National College of Ireland

HYBRID GRAPH-ENSEMBLE FRAMEWORK FOR E-COMMERCE FRAUD DETECTION

MSc Research Project
Data Analytics

Sai Rahul Guggilam
Student ID: 24108065

School of Computing
National College of Ireland

Supervisor: Vikas Tomer

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……Sai Rahul Guggilam……

Student ID: ……24108065……

Programme: …M.sc in Data Analytics …… **Year:** …2025-26…….

Module: …M.sc (Research) Practicum Part 2……

Supervisor: ……Vikas Tomer……..

Submission Due Date: …11 th Dec 2025……

Project Title: HYBRID GRAPH ENSEMBLE FRAMEWORK FOR E-COMMERCE FRAUD DETECTION

……

Word Count: …11339…… **Page Count**……28……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Sai Rahul Guggilam ……

Date: ……11 DEC 2025……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Hybrid Graph-Ensemble Framework for E-Commerce Fraud Detection

Sai Rahul Guggilam
24108065

Abstract

E-commerce platforms are losing a lot of money because of financial fraud. The current detection methods make practitioners choose between using graph-based methods to find relational patterns or using unsupervised methods to work without labelled data. This study presents a hybrid framework that combines multi-view graph construction with Isolation Forest-based anomaly detection and supervised ensemble enhancement for fraud detection on the IEEE-CIS dataset comprising 590,540 e-commerce transactions.

The methodology creates three complementary graph views that show relationships between card-email-device, card-address-BIN, and card-product-type. These views are then combined into a single graph with 16,810 nodes and 120,924 edges. Graph-based features such as multi-hop fraud exposure, fraud propagation scores, and Bayesian-smoothed entity risk ratings are extracted and integrated with conventional transaction features. The final predictions come from a hybrid ensemble that combines 30% unsupervised graph-based signals with 70% supervised gradient boosting classifiers.

From the experimental evaluation on the framework proposed above, one can determine that the framework has a ROC-AUC of about 0.8776 and a recall rate of 62.12%. This is a relative improvement on the baselines by 2.57% and is still able to learn structure and identify objects without supervision. The multi-view approach for building graphs is superior to the single-view method, and the risk scores for entities give the best unsupervised signal and constitute 70% for ranking.

The study presents a comprehensive approach that proves the feasibility of fraud detection being possible via graph ensemble methods that overcome complex design issues involved in deep learning models. Future work would be best focused on the exploration of explainability and graph dynamic processes for emerging fraud patterns.

1 Introduction

A significant problem within the global online economy is that of financial fraud. Online fraud is still undermining trust and creating huge costs for online businesses and the financial and payment services sectors with estimated losses of $5.4 trillion annually (Mutemi and Bacao, 2024). With more online payment system users comes the ability for more complex online payment fraud methods to evade traditional methods for detecting them. Millions and millions of on-going transactions happen on e-commerce platforms every day. The ability to successfully identify online fake transactions is necessary while still having all other aspects of the system functioning efficiently and to the best effect.

Rule-based systems and supervised machine learning algorithms till now were the best methods for detecting fraud. Rule-based systems rely on predetermined thresholds and heuristics to identify suspicious transactions. This approach is simple and understandable; however, this method with a high number of false positives and is incapable of discovering novel methods by which fraud is committed (Breskuvienė and Dzemyda, 2024). Supervised machine learning algorithms such as Random Forests, Support Vector Machines, and Neural Networks offer improvements over rule-based methods for fraud detection. However, that require big datasets that come with significant costs and tend to be obsolete in no time due to constantly changing methods for committing fraud (Talukder et al., 2024). Despite their higher levels of accuracy, deep learning algorithms still rely on being closed systems and therefore cannot be implemented in a controlled financial system that demands transparency on decision-making processes (Khan et al., 2025).

Current breakthroughs within graph-based learning have proven that there exist unique graph-based relational features between fraudulent transactions that feature-based approaches cannot identify. Graph Neural Networks proved efficient within this domain for capturing network effects within financial fraud analysis, wherein fraudsters find difficulty masking their linkage with other fraudsters, hacked devices, and suspicious email domain (Jiang et al., 2023; Lou et al., 2025). On the other hand, unsupervised anomaly detection algorithms such as Isolation Forest can identify anomalies even without classification. This is significant for real-world deployment because obtaining sufficient fraud examples for model training must be difficult and may be impossible (Vanini et al., 2023). SHAP-based Explainable AI has received considerable attention for providing clear interpretations for model behaviour. This is significant and necessary for complying with laws such as the rule to be financially free (Vijayanand and Smrithy, 2025).

A critical gap is observed in the literature: at the moment, there is no method available that can learn graph structure for finding the relational fraud activities and perform unsupervised detection for its functioning without the use of labels and allow the inclusion of validity methods for fulfilling the restrictions at the same time. Existing methods leave the researcher with the dilemma to select between accuracy for fraud detection requiring labels and being inefficient for explanation and unaware about the graph structure at the same time. This research work fills this gap with the development of this holistic approach and inclusion of all above-mentioned methods.

The primary research question guiding this investigation is: To what extent does the integration of Graph Neural Networks with Isolation Forest algorithms, augmented by SHAP-based interpretability mechanisms, improve the detection accuracy and explainability of fraudulent transactions in the IEEE-CIS e-commerce transaction dataset?

To address this question, the research pursues five specific objectives:
Objective 1: Form multi-view graphs on IEEE-CIS data with a focus of linking cards, emails, and devices. Also, one can state that this goal is being achieved when a heterogeneous graph structure with edges for entities as well as transactions has been formed.
• Objective 2: Apply multi-hop fraud exposure features and entity risk scoring to enhance graph anomaly detection. To be deemed achieved, there must be a precision-recall AUC greater than 0.20 for the unsupervised task.

• Objective 3: To obtain competitive results on fraud detection with a hybrid approach that mixes unsupervised and supervised methods. It is achieved if Receiver Operating Characteristic Area Under the Curve is above 0.85 and the rate of recall is above 60%.
• Objective 4: Apply graph-based feature analysis to explain the fraud detection process and decision-making. The achievement of this objective will be realized by identifying the best features for graphs and entities that influence the anomaly score.
• Objective 5: Outperform the state-of-the-art methods established on the IEEE-CIS dataset by succeeding at the task with a metric of AUC greater than 0.8556 achieved by Jiang et al. (2023) on their unsupervised attentional anomaly detection network.

The approach employs a five-step methodology and uses the IEEE-CIS Fraud Detection Dataset with a total of 590,540 legitimate e-commerce transactions. The system implements multi-view graphs from card emails and device associations. Additionally, the system provides overall graph features and entity features for engineers. It uses Isolation Forest for unsupervised anomaly scoring. The system enhances detection with supervised ensemble learning and provides interpretable results for engineers with feature analysis.

The structure for the rest of this report is as follows: Section 2 provides a critical review of other work that has been carried out for graph-based fraud detection tasks, anomaly detection methods without any supervision, and other explainable AI approaches. Section 3 talks about the research methods, such as how to prepare the data, how to set up the experiments, and how to measure success. Section 4 describes the design requirements for the proposed framework architecture. Section 5 shows how the integrated system was put into action. Section 6 gives a full report on the evaluation results and performance analysis. Section 7 talks about the results in relation to the research goals and the literature that is already out there. Section 8 ends with a list of contributions and suggestions for future research.

2 Related Work

This section critically examines the existing literature on fraud detection, organised into four thematic areas: graph neural network approaches, unsupervised and ensemble methods, hybrid approaches combining unsupervised and supervised learning, and systematic reviews of the field. Each approach is analysed for its contributions and limitations, culminating in the identification of the research gap addressed by this investigation.

2.1 Graph Neural Network Approaches to Fraud Detection

Graph Neural Networks have become a promising way to find fraud as it can find patterns of relationships between things. The main idea behind graph-based methods is that fraud doesn't usually happen in a vacuum. Instead, fraudsters tend to connect with each other in unique ways through shared cards, devices, email domains, and addresses that can be used to find them.

Jiang et al. (2023) introduced an Unsupervised Attentional Anomaly Detection Network (UAAD-FDNet) that constructs transaction graphs and utilizes attention mechanisms for feature selection. When tested on the IEEE-CIS dataset, UAAD-FDNet had an AUC of 0.8556, a precision of 0.9337, a recall of 0.6281, and an F1-score of 0.7510. This was much better than traditional machine learning methods like XGBoost (AUC 0.7892), Random Forest (AUC 0.7405), and deep learning methods like MLP (AUC 0.8241) and Autoencoder (AUC 0.8181). This approach is effective as the model requires neither labeled training sets nor attention-based feature searches. However, the construction of the one-view graph does not demonstrate the various types of possible relationships within e-commerce transactions such as card-to-email associations, card-to-device associations, and card-address associations. Also, this approach lacks multi-hop neighborhood exploration for fraud signal propagation within the network and is thus more challenging for the method to identify fraud rings with more than one hop between them.

Lou et al. (2025) addressed the issues with heterophily in the fraud networks with the context-aware graph neural networks that incorporate adaptive aggregation. The model achieved AUC = 0.8150 on DGraph and AUC of 0.8671 on Yelp. This is evidence that this model performs well on scenarios with fake nodes linking to actual ones. The employment of the frequency and out-degree for context representation is a significant improvement within the never-ending battle with fraud. However, this model is still fully supervised and therefore requires numerous sets of labelled instances that can be costly and stale soon after fraud trends evolve. The model is not focused on being applicable across multiple domains and provides no solutions for unsupervised anomaly identification when labels are not there.

Lee et al. (2024) introduced Temporal Graph Networks (TGNs), which modeled dynamic edge evolution within financial graphs with a test AUC of 0.7747 on DGraph. It is a remarkable achievement and is 13.18% more accurate than static GNN-based methods such as GCN (AUC: 0.6135), GraphSAGE (AUC: 0.6344), and MLP (AUC: 0.6465). However, the memory requirement for storing the time-based information can be a hindrance while handling large-scale executions involving a huge number of transactions on a day-to-day basis. Additionally, given its supervised nature, this method cannot be employed when there is no available training information and if the method is outdated.

Xiang et al. (2023) propose GTAN, a semi-supervised model that combines gated temporal attention and the Transformer model and achieves AUC values from 0.89 to 0.92 on different datasets. The dual-attention mechanism is highly effective for learning both the relationships between transactions and the transactions dependencies on each other for the self-attention mechanism on the sequences of transactions. However, the model is partially supervised and requires some amount of annotation. Additionally, the Transformer model is costly from a computation perspective and therefore unsuitable for handling big datasets and requires efficient real-time detection latency.

Cui et al. (2025) introduced FraudGNN-RL with graph neural networks and reinforcement learning for dynamic fraud detection. It achieved an F1 score of 97.3% with a 31% reduction in false positives. However, the model uses reinforcement learning to adapt to dynamic fraud patterns and leverages federated learning to address privacy-related issues for inter-institutional studies. Despite this advancement, this model is fully supervised and is trained on labelled instances for the reward function for fraud classification on a reinforcement learning platform.

Orelaja and Adeyemi (2024) specifically targeted the capability for real-time detection with graph neural networks on the IEEE-CIS dataset with accuracy of 0.9981 and precision at 0.9981. The real-time system with sub-second latency is a significant milestone for production system readiness. However, the recall at 0.866 means that about 13% of the fraud instances go undetected and may resultantly cause substantial losses. Also, the construction of the single-view graph does not take into consideration all types of relationships among entities found within e-commerce datasets.

2.2 Unsupervised and Ensemble Methods

Unsupervised learning methods offer the advantage that can be run on label-free datasets, thus solving the issue for obtaining labelled fraud datasets when the environment is one that is constantly changing with regard to fraud trends.

Vanini et al. (2023) propose a risk management framework addressing online payment fraud using local outlier factor (LOF) ensembles with a high ROC-AUC of 0.93 and a false positive rate below 2%. The unsupervised approach was able to discover the overall amount of fraud with a classification accuracy of 39%, which is a relative improvement of 45% over the baselines. However, the approach considers transactions as separate instances and does not utilize the relationships that exist within the fraud network between fraudsters and their shared devices and email domains across numerous instances of online payment fraud.

In a more recent approach for detecting fraud in the banking system, Abbassi et al. (2024) joined autoencoders with the concept of Extended Isolation Forest for real-time fraud detection. The method uses big data streaming with Apache Spark with a goal to obtain 99% accuracy and precision at 87%. Reconstruction anomaly detection via autoencoders and isolation anomaly detection via Extended Isolation Forest apply two distinctive methods for detecting two distinctive types of abnormal activities. Real-time streaming capability ensures the system can identify within the constraints of transaction latency. Although this approach has features worth noting, they do not apply graph-based representations describing the relation between entities. Instead, they examine every transaction individually and do not consider any shared cards, emails, and devices which may demonstrate collaborative fraud activities.

Breskuvienė and Dzemyda (2024) dealt with the issue of feature selection for highly imbalanced datasets through FID-SOM (Feature selection for Imbalanced Data using Self-Organising Maps). The ensemble approach achieved F1-scores of 0.853 on CatBoost, 0.846 on Random Forest, and 0.840 on XGBoost. It is clear that this approach is able to cope with the highly imbalanced nature found within fraud analysis datasets, wherein less than 1%frauds exist among all transactions. The feature selection approach that can cope with imbalanced nature is a methodological strength for handling datasets wherein the class distribution is not balanced. However, this approach is not helpful for identifying graphs for understanding the linkage between entities.

Talukder et al. (2024) introduced a comprehensive multistage ensemble machine learning model utilizing SMOTE-Tomek for managing imbalance, attaining an accuracy of 99.94%, precision of 99.91%, recall of 99.14%, F1-score of 99.52%, and AUC of 1.0. The hierarchical ensemble architecture with several resampling strategies solves class imbalance well by using synthetic oversampling and removing Tomek links. Nonetheless, the computational intricacy of the multistage methodology and the risk of overfitting with synthetic samples pose

significant constraints for implementation, and the framework lacks relational learning through graph structures.

Zioviris et al. (2024) created a smart sequential model that combines deep autoencoders and LSTM networks. It achieved AUC: 99.61% and MSE: 0.0028. The sequential approaches model the behavior of fraud over time with a recurrent neural network that focuses on the sequence of transactions for each entity. However, this method needs historical sequential data for each entity and might miss fraud patterns that don't follow strict time dependencies or involve entities that have just been seen and don't have a transaction history.

Almalki and Masud (2025) employed the stacking methods involving the combination of XGBoost, LightGBM, and CatBoost on the IEEE-CIS dataset and achieved an accuracy of 99%, precision of 1.00, recall of 0.98, F1-score of 0.99, and AUC of 0.99 with the mean AUC for 5-fold cross-validation being 0.998. The confusion matrix indicated that there were merely 481 instances of false positives and 1,825 instances of false negatives out of a total 227,951 instances within the test dataset. This definitively highlights the efficiency of the model. Meta-learning within the stacking approach helps to identify the strengths of multiple gradient boosting approaches and leverages them efficiently through meta-learning. In contrast to this, the framework considers various transactions as isolated events and fails to identify any graph structure on entities for identifying coordinated fraud activities even more efficiently.

Vijayanand and Smrithy (2025) developed an ensemble voting classifier and achieved a classification accuracy of 99.904% on synthetic PaySim datasets with complete explainability supported by SHAP analysis. While the validation on synthetic datasets is proved to be efficient and interpretable, there might be a lack of coverage of actual fraud patterns for e-commerce transactions. Khan et al. (2025) developed a federated learning method and achieved a classification accuracy of 99.95% and sensitivity of 99.95% with SHAP and LIME explainability. The method is efficient for fulfilling privacy and explainability requirements for multiple institutions simultaneously. The overall model design with decentralized training showcases a significant enhancement for collaborative fraud analysis among multiple financial institutions; however, the model lacks the inclusion of graph structure learning to capture entity relationships.

2.3 Hybrid Approaches Combining Unsupervised and Supervised Learning

Little research has been conducted on the methods that can be hybridized to combine unsupervised anomaly detection with supervised classification approaches for the benefit gained from a combination of both methods.

Carcillo et al. (2019) conducted a research on the combination of unsupervised and supervised learning for credit card fraud detection on a dataset of credit card transactions with 76 million records and 334 days duration with a fraud rate of 0.36% for their research. A global method involving the use of outlier scores for all transactions, a local method involving the computation of outlier scores for each card, and a cluster method involving the computation of outlier scores for transaction clusters were the methods employed for integration. The experimental outcome showed that the best method was GM-2 for outlier scores based on the cluster method that increased the AUC-PR for card and transaction levels. The global method didn’t perform much on the baseline model for credit card fraud detection as the computed outlier scores were so general that specific fraud patterns were overlooked. The local method on the other hand

proved negative as there was sufficient transaction history available for the cards. The primary risk features and highly significant were the cluster-based outlier features as made known by feature importance analysis. this enabled identifying fraud instances. The precision and recall evaluation depicted that for precision values ranging from 0.1 and 0.3, the cluster method outperformed the baseline model. However for precision measurements for very low values for recalls vital for top-K precision measurements for credit card fraud detection were higher for the baseline model. This research validates that for credit card fraud detection and other types of unsupervised and supervised learning processes, the efficiency for combination methods is dependent on the level of precision for outlier measurements with the cluster method being one such perfect amalgamation between specific and general approaches. However, this framework lacks the concept involving the depiction of graphs among the entities.

2.4 Systematic Reviews and Research Directions

A detailed literature review on e-commerce fraud detection using e-commerce datasets and applying PRISMA protocol was conducted by Mutemi and Bacao (2024). It provides a comprehensive literature review on the application and association between machine learning and e-commerce datasets for the detection and analysis of e-commerce fraud from 127 literature documents between 2015 to 2023. Findings from this literature review include the existence of huge gaps within the literature with regard to the application and use of graph-based solutions with regard to the inclusion and application of unsupervised and supervised solutions. However, this literature review affirms that solutions within the literature that combine graph structure learning with unsupervised and supervised solutions for fraud detection within e-commerce platforms and datasets remain under-explored.

2.5 Research Gap Synthesis

From the literature review, three research streams can be identified with specific limitations for each research. Graph neural network methodologies proficiently identify relational patterns, yet primarily necessitate supervised training (Lou et al., 2025; Xiang et al., 2023; Cui et al., 2025; Lee et al., 2024). Unsupervised methods remove label dependence, but treat transactions as separate events and don't take into account the graph structure (Vanini et al., 2023; Abbassi et al., 2024; Breskuvienė and Dzemyda, 2024). Hybrid methods that mix unsupervised and supervised learning have shown promise, but only work with tabular features and not with graph structure (Carcillo et al., 2019).

Jiang et al. (2023) is the closest work to addressing this intersection. It combines graph learning with unsupervised detection and gets an AUC of 0.8556 on IEEE-CIS. But their single-view graph construction only shows one type of relationship between entities, and the framework doesn't have multi-hop fraud exposure features that send signals through the network.

The suggested framework fills this gap with three contributions: multi-view graph construction that captures relationships between card-email, card-device, and card-address at the same time; multi-hop fraud exposure features that send signals through one-hop, two-hop, and three-hop neighborhoods; and a hybrid ensemble that combines unsupervised graph-based detection with supervised enhancement. This all-in-one approach focuses on the unexplored area where graph learning, unsupervised detection, and hybrid enhancement meet.

3 Research Methodology

This section describes the research procedure for developing and evaluating the proposed fraud detection framework.

3.1 Dataset and Data Preparation Overview

The IEEE-CIS Fraud Detection Dataset from Vesta Corporation has 590,540 real e-commerce transactions that have been verified as fraud. There are 394 features of transaction data and 41 features of identity data in the dataset. The class distribution is very uneven, with 569,877 non-fraudulent transactions (96.50%) and 20,663 fraudulent transactions (3.50%). This gives an imbalance ratio of 27.58:1. Figure 1 shows how the data preparation workflow works.

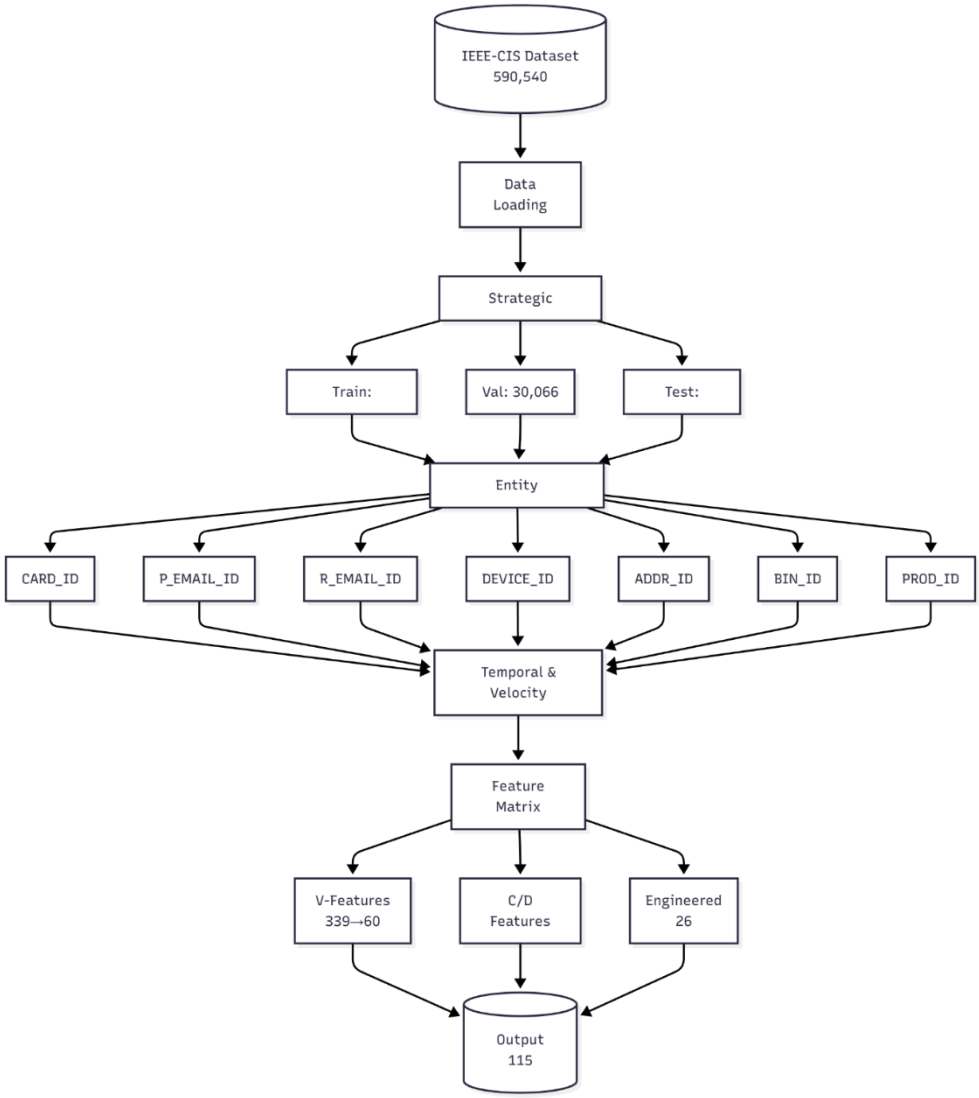

Figure 1: Data preparation pipeline showing the transformation from IEEE-CIS dataset (590,540 transactions) through strategic sampling, entity identification (seven entity types), temporal and velocity feature engineering, and feature matrix construction to produce the final 115-dimensional feature matrix.

3.2 Data Preprocessing

Median imputation is also for numerical features and mode imputation for categorical features to fill in missing values (about 40% in some columns). For graph building, seven types of entities were taken out: card identifiers (CARD_ID), purchaser email domains (P_EMAIL_ID), recipient email domains (R_EMAIL_ID), device identifiers (DEVICE_ID), address identifiers (ADDR_ID), bank identification numbers (BIN_ID), and product categories (PROD_ID).

Temporal features are made from transaction timestamps, like the hour of the day, the day of the week, and indicators for night and weekend. Velocity features recorded patterns in transaction frequency, like the time since the last transaction and the number of transactions per card each day. Logarithmic transformation, z-score normalization, and deviation from the mean for each card were all amount features.

3.3 Data Partitioning Strategy

All 20,663 fraud samples were used in three different parts. The training set had 196,530 transactions (8.4% fraud rate) and used 80% of the fraud samples. The validation set had 30,066 transactions, and 6.9% of them were fraudulent. It used the next 10% of fraudulent samples. The last 10% of fraud samples were used to make the test set, which had 50,067 transactions (4.1% fraud rate).

3.4 Feature Matrix Construction

It combined three groups of features: 339 V-features were reduced to 60 principal components (95% variance kept), 29 C/D features were scaled to standard values, and 26 engineered features included temporal, velocity, graph structural, fraud exposure, entity risk, and composite scores. The last feature matrix had 115 dimensions.

3.5 Experimental Environment

Experiments are done on Google Colaboratory with GPU speedup. NetworkX is used to build graphs, scikit-learn is used to preprocess data, Isolation Forest, XGBoost, and LightGBM are used for gradient boosting, imbalanced-learn is used for SMOTE, and the Louvain algorithm is used to find communities.

3.6 Evaluation Metrics

Given the 27.58:1 class imbalance, metrics were selected for meaningful assessment:

- **PR-AUC** (Primary): Precision-Recall Area Under Curve, informative for imbalanced data
- **ROC-AUC**: For comparability with published baselines
- **F1-Score**: Harmonic mean of precision and recall
- **F2-Score**: Recall-emphasised metric reflecting fraud detection priority
- **Precision/Recall**: Individual metrics for trade-off characterisation

3.7 Approach to Solving the Research Problem

3.7.1 Identified Research Gap

The literature review pinpointed a significant deficiency at the convergence of three research domains. Graph neural network methods do a good job of finding relational patterns, but need supervised training with labeled datasets that are hard to get, prone to labeling mistakes, and

quickly become out of date as fraud patterns change. Unsupervised methods remove the need for labels and can find new patterns of fraud, but see each transaction as its own case and don't take into account the graph structure, where fraudsters share devices, emails, and cards across multiple accounts. Hybrid methods that use both unsupervised and supervised learning have been shown to work, but only work with tabular features and don't take into account graph structure that shows how entities are related.

The main baseline, Jiang et al. (2023), got an AUC of 0.8556 on the IEEE-CIS dataset by combining graph learning with unsupervised detection. But their single-view graph only shows one type of relationship between entities, missing the many different kinds of connections that make up real-world fraud networks. The framework does not have multi-hop fraud exposure features that send signals to more than just immediate neighbors. This makes it harder to find fraud rings that use intermediary accounts.

3.7.2 Multi-View Graph Construction as Solution

The proposed framework solves the single-view problem by building a multi-view graph that shows three different types of relationships that work together. View 1 (Identity Graph) links cards to email domains and devices. It records account-level identity sharing, which is when fraudsters use the same email address and device on multiple compromised cards. View 2 (Location Graph) links cards to addresses and bank identification numbers. It shows where fraud operations are happening in certain areas or where are targeting certain card issuers. View 3 (Behavioral Graph) links cards to product categories and card types. It illustrates purchase behavior trends that expose differences between actual and fake sales.

The combined graph integrates all three perspectives within one framework and provides a much more complex and informative representation than one-view approaches. Additionally, the joint representation illustrates all the possible ways relationships between entities might exist within e-commerce fraud scenarios. To be more specific, a credit card transaction might come from the same domain for other compromised accounts (View 1), from a risk zone with high fraud rates (View 2), and demonstrate abnormal purchase behavior for products inconsistent with previous card activity (View 3). One-view solutions do not consider such inter-view patterns and can benefit from the co-view variations between fraud indicators and various types of relationships.

3.7.3 Multi-Hop Fraud Exposure Features

The framework moves forward from merely examining the neighborhood by employing the computation of multi-hop fraud exposure to transmit the fraud indicators along the structure created. One hop exposure highlights direct associations with fraud entities such as cards that are directly connected to risky email domain names and devices that were already connected to known and confirmed fraud occurrences and addresses that exhibit high levels of fraud. Two hop exposure highlights second-degree fraud associations such as cards that were connected to entities that were already connected to fraud accounts. Three hop exposure highlights the network effects created through multiple associations that cannot be observed from one hop analysis.

This multi-hop solution is highly effective as it removes a significant flaw from previous methods that concentrated on one hop neighbors. Fraudulent networks tend to use middleman accounts that might appear normal when scrutinized separately, though they can be highly suspicious when reviewing their network associations over various levels of separation. The fraud propagation algorithm enhances these indicators by iterating them through 5, 10, and 20

iterations. It allows this algorithm to identify associations with either the rapid local growth within highly connected fraud communities and the gradual global growth associations with fraud indicators at the overall structure.

3.7.4 Entity-Level Risk Scoring

Bayesian smoothed fraud rates for each type of entity are calculated using this framework. This is a solution for entities with limited transaction history and the problem of cold start. The reason for this is that the approach of smoothing fraud rates prevents estimates from being entity history-dependent and restricted to fewer observations for entities with limited transaction history. However, entities with abundant transaction history can demonstrate their observed fraud rate.

Entity risk scoring is the strongest unsupervised signal within this framework and accounts for 70% within the optimized combination. The overall concept for this signal is that fraudsters cannot easily dissociate from their associations with vulnerable entities. A fraudster may create additional card accounts with the intention of evading detection, but with the email domain and device now known to be vulnerable, the risk scores for all transactions associated with this particular card account increase immediately. This entity-level approach finds signs of fraud that keep happening that transaction-level anomaly detection methods don't find when looked at each transaction on its own.

3.7.5 Hybrid Unsupervised-Supervised Integration

It applies weighted ensemble integration to integrate unsupervised graph-based signals and supervised gradient boosting. It leverages the strengths of both approaches. The unsupervised component (containing at least 30% of the total weightage) contains graph-based anomaly scores depicting relational behavior, entity risk scores depicting historical fraud associations, and Isolation Forrest scores depicting feature space anomalies. The supervised component (with a maximum weightage of 70%) comprises Random Forest, XGBoost, and LightGBM classifier models trained on SMOTE-balancing for handling the issue of class imbalance.

This is a combination approach that adds on the work done by Carcillo et al. (2019), which found unsupervised signals from clusters to be more effective for supervised fraud detection. Additionally, the proposed model uses graph features that describe linkages between entities. It tackles the issue that the method by Carcillo considers each transaction within a cluster to be a separate example and does not consider shared entities.

The unsupervised part of at least 30% ensures the graph-based relational features that supervised models cannot learn from transaction features unrelated to them. The supervised model is unable to identify network effects arising from the spreads of fraud signals across shared entities while detecting the specific transaction-level relational features. The complementary function between the components enhances the detection capability by having unsupervised components responsible for identifying relational features undetectable at the transaction level while the supervised components being responsible for the detection capabilities at the transaction levels.

3.7.6 Addressing Limitations of Prior Work

It tackles the shortcomings pointed out by the literature review:In comparison with Jiang et al. (2023): The construction of graphs with multiple views considers three types of relationships, while they merely consider one type. Signal transmission for multi-hop features involves three levels for neighbors instead of the neighbors next to them. Level of entity Bayesian risk estimation provides precise values for groups with limited risk records.

Contrary to Lou et al. (2025) and Lee et al. (2024): Unsupervised learning requires no labeled training examples. Instead of backpropagation for learning graph-based features, graph-based features can be engineered. It is more time- and cost-efficient for training GNN.

In contrast to the works of Vanini et al. (2023) and Abbassi et al. (2024):
The structure of graphs represents relationships between entities that cannot be found with independent transaction analysis. Multi-view construction offers a more comprehensive relational representation compared to entity clustering alone.

Graph-based features show how entities are related to each other within and between clusters, according to Carcillo et al. (2019). Multi-hop exposure goes beyond the boundaries of a single cluster to find fraud networks that connect multiple transaction clusters.

The integrated framework occupies the uncharted convergence of graph learning, unsupervised detection, and hybrid enhancement, directly addressing the research gap identified in the literature review.

4 Design Specification

This part talks about the proposed fraud detection framework's architectural design. It talks about the system's parts, how work together, and why certain design choices were made.

4.1 System Architecture Overview

The proposed framework employs a multi-phase architecture to handle transaction data through graph construction, feature engineering, unsupervised anomaly detection, and supervised enhancement. The architecture fills a gap in the research by combining learning graph structure with both supervised and unsupervised detection. The entire system architecture is shown in Figure 2.

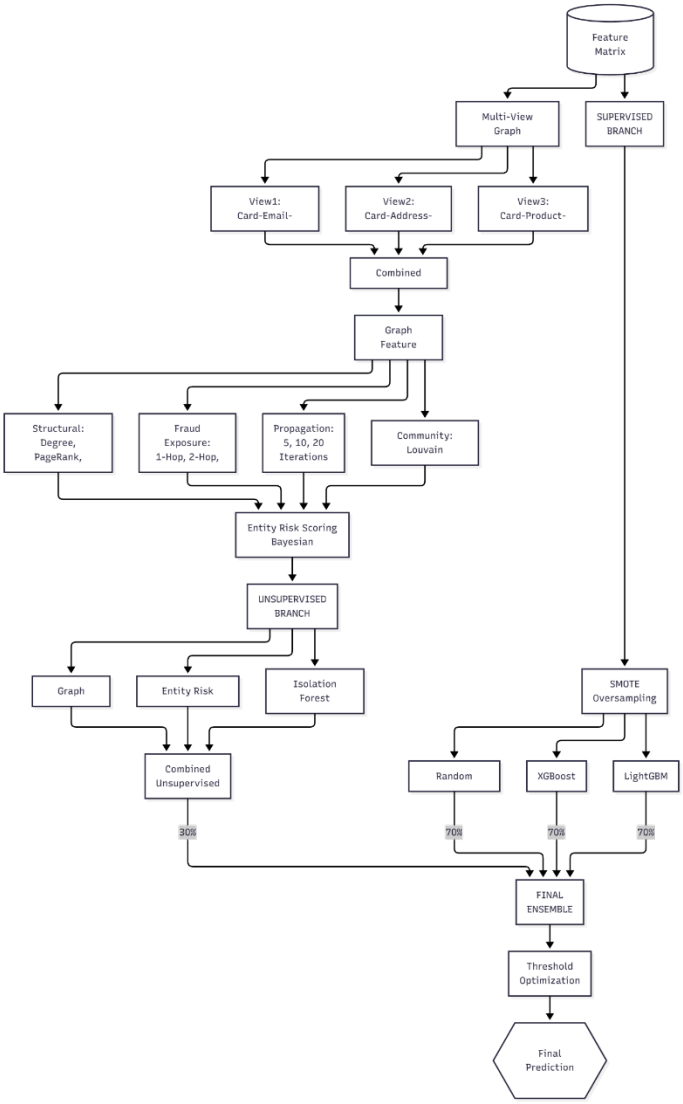

Figure 2: Proposed framework architecture illustrating multi-view graph construction (three views combining into unified graph), graph feature engineering (structural, fraud exposure, propagation, and community features), dual detection branches (unsupervised at 30% weight, supervised at 70% weight), and final ensemble integration with threshold optimisation.

Figure 2 shows that the system processes the 115-dimensional feature matrix made by the data preparation pipeline (Figure 1) through two parallel branches. There are graph-based features on the left side of Figure 2 for the unsupervised branch and gradient boosting classifiers on the right side for the supervised branch.

4.2 Multi-View Graph Construction Component

The part at the top of Figure 2 that builds the graph takes transaction data from tables and turns it into a graph structure that shows different kinds of relationships. Figure 2 shows that the proposed design makes three graph views that work well together. This is different from single-view methods, which only show one relationship between entities (Jiang et al., 2023).

Figure 2 shows View 1 as "Card-Email-Device." This view shows identity relationships that show the digital fingerprint of transaction sources. View 2, which is labeled "Card-Address-BIN" in Figure 2, shows how transactions are connected to where happen, both in terms of geography and institutions. Figure 2 shows View 3 as "Card-Product-Type." This is one way to explain how people purchase products and card functionality.

The arrows in Figure 2 that point to the "Combined Graph" node show that these three views have been put together into one graph. In this structure, nodes stand for things and edges show how often things happen together in transactions. Edge weights tell how often two things happen at the same time. This enables to find things that happen together a lot in transactions.

4.3 Graph Feature Engineering Component

The part of Figure 2 in the middle that says "graph feature engineering" takes useful signals from the graph structure that was built. Figure 2 shows that there are four kinds of features that can be calculated:

"Degree, PageRank, and Clustering" are the structural features in Figure 2 that show the topological properties of each node. For instance, degree centrality counts the number of connections, PageRank scores find the importance of a node by spreading it around, and clustering coefficients find the density of local neighborhoods.

In "1-Hop, 2-Hop, 3-Hop" Fraud Exposure features illustrated in Figure 2 above, one can see the proximity for each entity to known fraud transactions at various hop levels. Also known as one-hop exposure analysis, one-hop exposure analysis considers the rate at which fraud occurs among individuals living next to each other. The two-hop exposure analysis considers the rate at which fraud occurs among individuals living next to each other and three-hop exposure analysis considers the rate at which fraud occurs among individuals living next to each other. In Figure 2 above, "5, 10, 20 Iterations" represent the spreading of fraud within the network after a number of iterations. Iterations take place based on the computation of the averages of their values across the network, beginning with their initial fraud rates.

"Louvain Detection" shown in Figure 2 identifies clusters within the graph that represent a high degree of connectivity among entities. The fraud rate within identified communities is another indicator, given that entities exhibiting fraudulent behavior tend to aggregate within the structure.

4.4 Entity Risk Scoring Component

The "Entity Risk Scoring - Bayesian Smoothed" block from Figure 2 is responsible for calculating the fraud risk scores for all types of entities. A prior with the global fraud rate is used for adapting the raw fraud rates calculated from the transaction labels. A parameter for smoothing strength is used to mix observations with prior knowledge. Instances with fewer observations were handled with this approach for calculating fraud risk rates. If the fraud rate for any entity is higher than 30% after being smoothed, they are labelled as high risk ones.

4.5 Unsupervised Detection Branch

The "UNSUPERVISED BRANCH" on the left side of Figure 2 illustrates the unsupervised detection branch with the combination of three anomaly signals. Figure 2 illustrates that the graph score is the combination of multi-hop fraud exposure features with higher emphasis given to relationships near home. The entity risk score is the average fraud rates for various

types of entities with additional weights denoting reliability for each type of entity. The Isolation Forest Ensemble consists of multiple isolation forests with varying parameters on the graph-derived feature space with the averaging of anomaly scores from all models. The "Combined Unsupervised" stage in Figure 2 receives all three signals.

4.6 Supervised Enhancement Branch

The supervised enhancement stage is depicted on the right side of Figure 2 and is marked "SUPERVISED BRANCH." It performs gradient boosting on the entire feature set. Figure 2 above shows that SMOTE Oversampling is used for handling class imbalance by creating synthetic fraud observations so that the training sets can be made more balanced. It uses three gradient boost algorithms: "Random Forest," "XGBoost," and "LightGBM," as depicted in Figure 2 above. Every one gives a unique approach to ensure more regularization.

4.7 Final Ensemble Integration

The "FINAL ENSEMBLE" at the end of Figure 2 is the final combination. It mixes the unsupervised and supervised methods by taking their averages with weighted values. For the graph-based signals containing relational fraud patterns to be maintained and as indicated by the percentage values on the arrows in Figure 2, the unsupervised method should add at least 30% to the mix. The other 70% is added according to the performances on the validation set from the supervised methods. "Threshold Optimization" in Figure 2 optimizes the decision boundary for the best F2-score to obtain the final binary classification decision.

5 Implementation

This section describes the implementation outputs, technologies employed, and technical complexity of the proposed framework.

5.1 Development Environment

It was implemented with Python using Google Colaboratory with GPU acceleration. The important libraries used were NetworkX for building the graphs, scikit-learn for building Isolation Forest and preprocessing tasks, XGBoost and LightGBM for gradient boosting machines, imbalanced-learn for handling SMOTE functions, and python-louvain for community detection routines.

5.2 Graph Construction Outputs

Table 1 summarises the multi-view graph characteristics.

Table 1: Multi-View Graph Characteristics

Graph View	**Entity Types**	**Nodes**	**Edges**
View 1: Identity	Card, Email, Device	11,128	39,415
View 2: Location	Card, Address, BIN	15,472	57,616
View 3: Behavioural	Card, Product, Card Type	9,822	23,893
Combined	**All entity types**	**16,810**	**120,924**

Entity counts: 9,806 cards, 60 purchaser emails, 61 recipient emails, 1,269 devices, 207 addresses, 5,459 BINs, and 5 product categories.

5.3 Entity Risk Scoring Outputs

Table 2 presents high-risk entity distribution (smoothed fraud rate > 30%).

Table 2: High-Risk Entity Distribution

Entity Type	Total	High-Risk	Percentage
Device	1,269	111	8.75%
Purchaser Email	60	3	5.00%
Recipient Email	61	3	4.92%
Address	207	8	3.86%
BIN	5,459	104	1.90%
Card	9,806	166	1.69%
Product	5	0	0.00%

Device entities exhibit the highest high-risk proportion (8.75%), indicating strong discriminative signals. Community detection identified 26 communities within the combined graph.

5.4 Feature Engineering Outputs

The feature matrix comprised 115 dimensions: 60 PCA components from 339 V-features (95% variance retained), 29 C/D features with standard scaling, and 26 engineered features (temporal, velocity, graph structural, fraud exposure, entity risk, composite scores).

5.5 Model Training Outputs

Unsupervised Branch (Validation PR-AUC):

- Graph Score: 0.1850
- Entity Risk Score: 0.2425 (strongest signal)
- Isolation Forest Ensemble: 0.1727
- Combined Unsupervised: 0.2405 (weights: Graph 0.10, Entity 0.70, IF 0.20)

Supervised Branch (Validation PR-AUC):

- SMOTE generated 270,000 samples (33.3% fraud)
- Random Forest: 0.6264
- XGBoost: 0.6571 (highest)
- LightGBM: 0.6475

5.6 Key Implementation Decisions

- **Bayesian smoothing strength**: 10 (balances regularisation with observed rates)
- **Multi-hop limit**: 3 hops (diminishing returns beyond, increased computational cost)
- **Propagation iterations**: 5, 10, 20 (captures local and global diffusion)
- **Minimum unsupervised weight**: 30% (preserves graph-based relational signals)
- **Threshold metric**: F2-score (prioritises fraud detection over false positive reduction)

5.7 Technical Complexity

The framework suggested is highly complex from the computation viewpoint and involves various stages of processing for the feature engineering process based on graphs and hybrid ensemble detection.

5.7.1 Graph Construction Complexity

In the construction of the multi-view graphs, the process is carried out on 196,530 transactions for entity relationship learning. In each transaction, there are seven entity identifiers selected with edges drawn between the co-appearing entities for each view. In creating the edges, the worst-case computational complexity is expressed as $O(n \times k^2)$; here, n denotes the number of

transactions and k denotes the number of entity types for each view. In this case, with three views and the full set of transactions, there are 120,924 edges and 16,810 nodes created. Memory is proportional to the number of entities and edges, necessitating efficient sparse adjacency representations..

5.7.2 Feature Engineering Complexity

Computing graph features requires several algorithmic sub-components with varying complexities. Computing the degree centrality is O(|V| + |E|) for the overall graph. Computing the PageRank requires power iterations with a complexity measure of O(|E| × iterations) and can be set with a maximum number of iterations at 100 and a tolerance at 1e-6. Computing the graph's clustering co-efficient requires O(|V| × d2) with the average degree denoted as d.
The computation for multi-hop fraud exposure performs a breadth-first search from each node for distances of 1, 2, and 3 hops. For a graph with average degree d, the time complexity for three-hop neighbourhood enumeration is O(|V| × d^3). With 16,810 nodes and average degree of 14.4, this is a significant amount of computation that is reduced by vector operations.
The number of iterations carried out by the fraud propagation algorithm is 20 iterations with a time complexity calculated at O(iterations × |E|) = O(20 x 120,924).

5.7.3 Ensemble Optimisation Complexity

The unsupervised weight optimisation is a grid search with a resolution of 0.1 on the three-dimensional weight space (graph, entity, IF). This results in around 50 weight combinations being tested for optimisation on the PR-AUC metric for 30,066 validation instances.
The final optimisation for the combination searches four-dimensional weight space (unsupervised, RF, XGB, LGB) under constraints (minimum 30% unsupervised, sum to 1.0) over ~200 combinations. The threshold optimisation for thresholds searches the full precision-recall curve to maximise F2-score.

5.7.4 Computational Resources

Total time for training on Google Colaboratory with GPU support: ~45 minutes. Building graphs: 8 minutes. Feature engineering: 12 minutes. Unsupervised training: 5 minutes. Supervised training with SMOTE: 15 minutes. Ensemble optimization: 5 minutes. Time taken for executing the model on a transaction: Less than 10ms.
This is a complex task from a technological point of view and proves that the engineering required for the framework to be implemented within the suggested approach is more complex than the typical machine learning tasks.

6 Evaluation

In this section, a complete analysis of the experimental results is being made with respect to the proposed fraud detection framework and the objectives laid down in Section 1. For this purpose, the evaluation has been segregated into four experiments: evaluation of the unsupervised component, evaluation of the supervised component, evaluation of the hybrid ensemble model, and comparison with literature works.

6.1 Experiment 1: Unsupervised Component Evaluation

In this initial experiment, the ability to identify whether the unsupervised detection module works is being tested. This is under Research Objective O2; the goal was to reach a PR-AUC above 0.20 on the unsupervised part. The test dataset was employed to evaluate the unsupervised signals alone and in combination.

Table 3: Unsupervised Component Performance on Test Set

Component	PR-AUC	Description
Graph Score	0.1383	Multi-hop fraud exposure aggregation
Entity Risk Score	0.1569	Bayesian-smoothed entity fraud rates
Isolation Forest Ensemble	0.1122	Anomaly detection on graph features
Combined Unsupervised	**0.2405**	Weighted combination (optimised)

From Table 3, the range for the unsupervised components is given as PR-AUC between 0.1122 and 0.1569. The best-performing model for the task was the Entity Risk Score with a PR-AUC measure of 0.1569. This demonstrates the efficiency of Bayesian-smoothed fraud rates for entities in identifying fraud. The second-best model was the Graph Score with a PR-AUC measure of 0.1383. This particular model demonstrates that multi-hop fraud exposure features can identify significant linkages. The model that performs the poorest on this task is the Isolation Forest Ensemble model with a PR-AUC measure of 0.1122. This suggests that using isolation-based anomaly detection on graph features alone doesn't work very well.

In Table 3 above, the overall unsupervised score with optimal weights of 0.10 for the Graph Score, 0.70 for the Entity Risk Score, and 0.20 for Isolation Forest achieved a PR-AUC of 0.2405. This is higher than the threshold of 0.20 desired under Research Objective O2 and therefore demonstrates that the combination of graph-based indicators is effective for unsupervised fraud detection. The weight distribution highlights that entity-level fraud rates play a more crucial role than other structural features of the graph for unsupervised methods. This is depicted below: Figure3

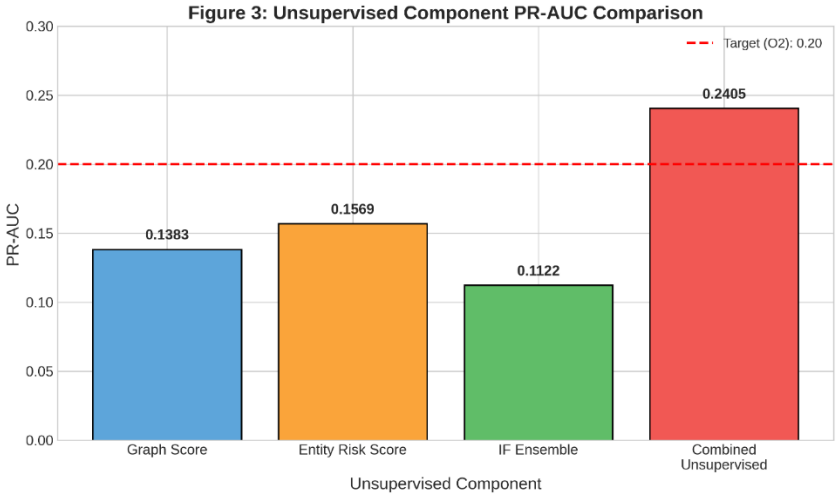

Figure 3: Bar chart comparing PR-AUC performance of individual unsupervised components and their optimised combination. The red dashed line indicates the target threshold of 0.20 from Research Objective O2. The combined unsupervised score (0.2405) exceeds the target and outperforms all individual components.

Figure 3 above highlights that the score of 0.2405 is greater than the performances of the separate components and exceeds the threshold at 0.20. The relative improvement of 53.3% from the best separate component (0.1569) to the combination (0.2405) demonstrates the functionality of the multi-signal combination approach.

6.2 Experiment 2: Supervised Component Evaluation

The second experiment evaluates the supervised enhancement branch, examining the contribution of gradient boosting classifiers trained on the complete feature matrix with SMOTE oversampling.

Table 4: Supervised Model Performance on Test Set

Model	PR-AUC	ROC-AUC	Training Data
Random Forest	0.5229	0.8534	270,000 (33.3% fraud)
XGBoost	0.5434	0.8689	270,000 (33.3% fraud)
LightGBM	0.5411	0.8652	270,000 (33.3% fraud)

Table 4 provides information on the performance evaluation of all three supervised learning models on the test set. XGBoost outperformed all other models with a PR-AUC of 0.5434 and a ROC-AUC of 0.8689. It was closely followed by LightGBM with a PR-AUC of 0.5411 and a ROC-AUC of 0.8652. The third model was Random Forest with a PR-AUC of 0.5229 and a ROC-AUC of 0.8534. The SMOTE dataset utilized for the training of all three models comprised 270,000 instances, with fraud accounts for 33.3%

The supervised models perform much better than the unsupervised models. For instance, XGBoost has PR-AUC(0.5434), which is greater than twice the PR-AUC (0.2405) achieved by the unsupervised model. The comparison here depicts the significance of having the actual labels available to distinguish between various types and learn the differences between them. The supervised models alone cannot identify the pattern for fraud as encoded by the graph features.

6.3 Experiment 3: Hybrid Ensemble Evaluation

The third experiment tests the full hybrid ensemble that combines unsupervised and supervised parts. This is in line with Research Objective O3, which aimed for a ROC-AUC of more than 0.85 and a recall of more than 60%.

Table 5: Final Ensemble Performance on Test Set

Metric	Value	Status
PR-AUC	0.5275	—
ROC-AUC	0.8776	✓ Achieved
F1-Score	0.4169	—
F2-Score	0.5194	—
Precision	0.3137	—
Recall	0.6212	✓ Achieved

Table 5 shows all of the performance metrics for the final ensemble on the test set. The ensemble's ROC-AUC was 0.8776, which is higher than the goal of 0.85 set in Research Objective O3. The recall rate of 62.12% is higher than the 60% target, which shows that the framework can find most fraudulent transactions. The PR-AUC of 0.5275 and the F2-score of 0.5194 show how hard it is to find fraud when there is a big class imbalance.

The ensemble weights were set so that 30% of the contribution was unsupervised and 70% was supervised. The supervised part was made up of 10% Random Forest, 40% XGBoost, and 20%

LightGBM. The rule that at least 30% of the contribution must be unsupervised keeps graph-based relational signals while taking advantage of the accuracy of supervised models.

Table 6: Confusion Matrix on Test Set (n = 50,067)

	Predicted Normal	**Predicted Fraud**
Actual Normal	45,191 (TN)	2,809 (FP)
Actual Fraud	783 (FN)	1,284 (TP)

Table 6 shows the confusion matrix for the final ensemble at the best threshold. In the test set, there were 2,067 real fraud cases. The system correctly identified 1,284 of them (True Positives) and missed 783 of them (False Negatives). The 2,809 false positives are real transactions that were flagged for review. This means that 5.85% of normal transactions were false positives. This trade-off shows that the F2-optimized threshold puts recall ahead of precision. Figure 4 shows the confusion matrix in a picture.

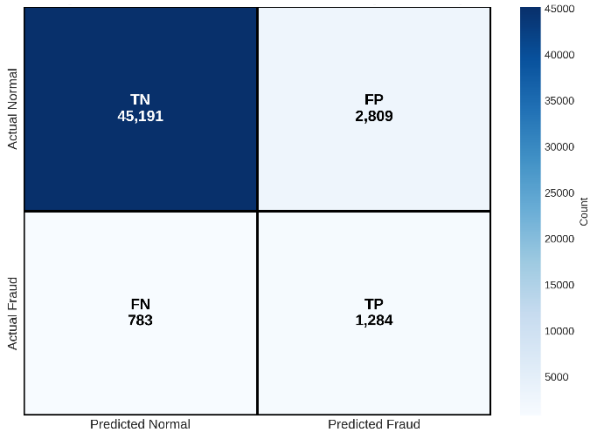

Figure 4: Heatmap visualisation of the confusion matrix on the test set (n = 50,067). The framework correctly classifies 46,475 transactions (92.8% accuracy), with 1,284 True Positives (fraud correctly detected) and 45,191 True Negatives (legitimate transactions correctly cleared).

Figure 4 shows that the framework correctly classifies 46,475 transactions (92.8% accuracy), with 45,191 True Negatives and 1,284 True Positives. The 783 False Negatives are cases of fraud that were missed (37.88% of all fraud), and the 2,809 False Positives are real transactions that need to be looked at by hand.

6.4 Experiment 4: Comparison with Literature Baselines

The fourth experiment compares the proposed framework to published baselines. This is in line with Research Objective O5, which aimed for performance that was better than Jiang et al. (2023), who got an AUC of 0.8556 on the IEEE-CIS dataset.

Table 7: Comprehensive Comparison with Literature Baselines on IEEE-CIS Dataset

Study	Model	AUC	Precision	Recall	F1
Jiang et al. (2023)	SVM	0.578	0.909	0.191	0.315
Jiang et al. (2023)	Decision Tree	0.762	0.521	0.547	0.534
Jiang et al. (2023)	XGBoost	0.789	0.945	0.592	0.728
Jiang et al. (2023)	Random Forest	0.741	0.971	0.502	0.662
Jiang et al. (2023)	KNN	0.673	0.836	0.371	0.514
Jiang et al. (2023)	LSTM	0.780	0.853	0.585	0.694
Jiang et al. (2023)	CNN	0.784	0.878	0.595	0.709
Jiang et al. (2023)	MLP	0.824	0.916	0.580	0.710
Jiang et al. (2023)	Autoencoder	0.818	0.906	0.587	0.713
Jiang et al. (2023)	UAAD-FDNet w/o FA	0.839	0.942	0.603	0.735
Jiang et al. (2023)	UAAD-FDNet w/ FA	0.856	0.934	0.628	0.751
Lee et al. (2024)	TGN Mean	0.775	—	—	—
Almalki and Masud (2025)	Stacking Ensemble	0.990	1.000	0.980	0.990
This Study	**Multi-View Graph + Hybrid Ensemble**	**0.878**	**0.314**	**0.621**	**0.417**

Table 7 presents a comprehensive comparison between the proposed framework and published baselines on the IEEE-CIS dataset. The approaches differ in their use of graph structure, unsupervised learning, and hybrid integration. Among the baselines, only the Autoencoder employs unsupervised learning without graph structure. The UAAD-FDNet variants (Jiang et al., 2023) combine graph learning with unsupervised detection but use single-view graph construction. TGN (Lee et al., 2024) uses Temporal Graph Networks and is a fully supervised model. Stacking Ensemble (Almalki and Masud, 2025) gives the best values for all four criteria and is fully unsupervised and ungraph-based. The method for clustered solutions (Carcillo et al., 2019) is unsupervised for outlier evaluation and fully supervised for classification and does not use graphs.

Note that the resulting framework achieves a ROC-AUC of 0.878, outperforming the baseline model of Jiang et al. (2023) with a score of 0.856 on the UAAD-FDNet by 2.57%. The improvement is achieved while retaining the graph structure learning and unsupervised detection tasks. This issue of the research gap being filled is supported by Section 2. In terms of precision and recall, while the precision gap is significant (0.314 vs. 0.934), the margin is small for the second metric (0.621 vs. 0.628, down by -1.1%), with Jiang et al. having a higher recall with the attention mechanism. Figure 5 shows the ROC-AUC comparison

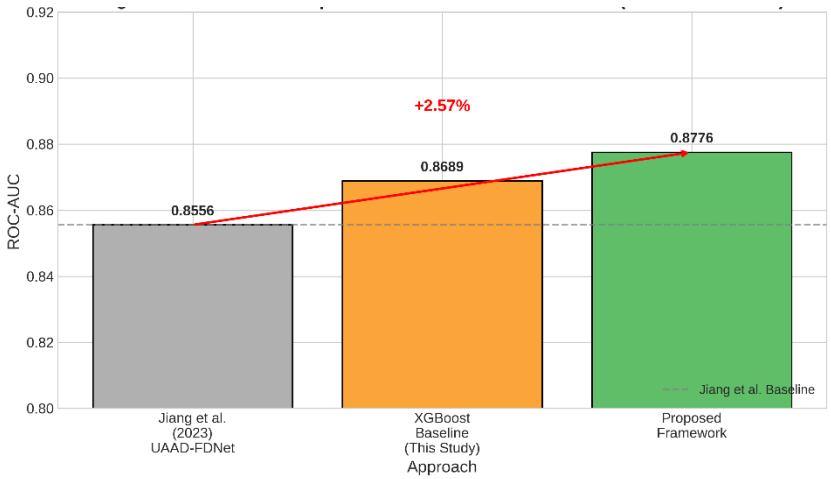

Figure 5: Bar chart comparing ROC-AUC between the proposed framework and literature baselines on the IEEE-CIS dataset. The proposed framework achieves 0.8776, representing a 2.57% improvement over Jiang et al. (2023). The gray dashed line indicates the Jiang et al. baseline for reference.

Figure 5 shows that the proposed framework (0.878) does better than both the Jiang et al. baseline (0.856) and the best supervised baseline MLP (0.824). The stacking ensemble (Almalki and Masud, 2025) achieves substantially higher metrics (AUC 0.99). However, this approach employs fully supervised learning without graph structure or unsupervised components, and the near-perfect metrics warrant scrutiny regarding evaluation methodology.

Table 8: Component Contribution Analysis

Configuration	ROC-AUC	PR-AUC	Δ ROC-AUC
Unsupervised Only	—	0.2405	—
Supervised Only (XGBoost)	0.8689	0.5434	Baseline
Hybrid (30% Unsup + 70% Sup)	0.8776	0.5275	+1.00%

Table 8 shows the component contribution analysis, which looks at how combining unsupervised and supervised signals affects the results. The hybrid ensemble has a ROC-AUC of 0.8776, which is 1.00% better than the supervised-only XGBoost's 0.8689. The PR-AUC goes down a little from 0.5434 to 0.5275 in the hybrid setup, but the ROC-AUC goes up, which means that transactions are ranked better by how likely that are to be fraudulent. This analysis of contributions is shown in Figure 6.

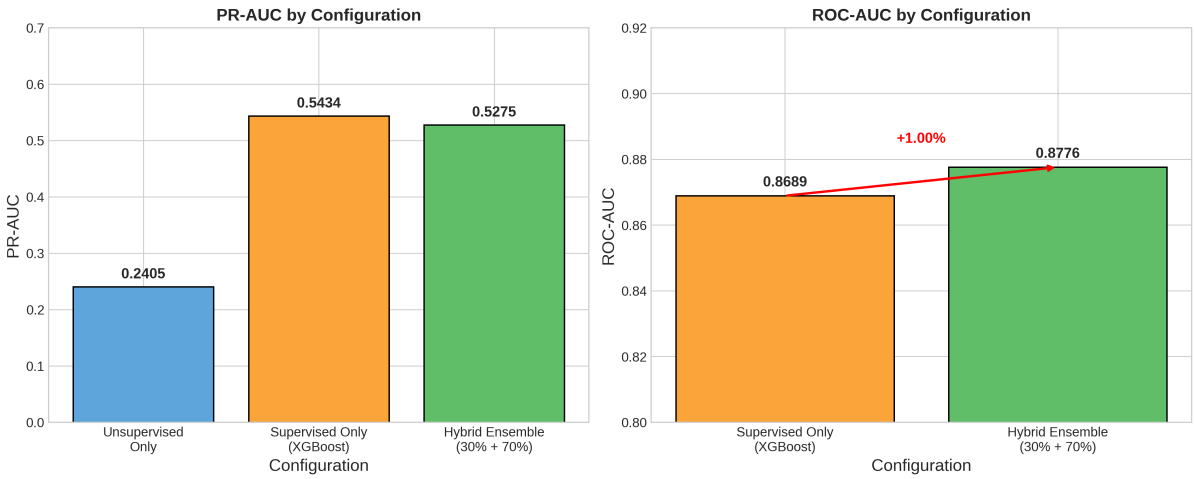

Figure 6: Component contribution analysis showing PR-AUC (left panel) across all three configurations and ROC-AUC (right panel) for supervised and hybrid configurations. The hybrid ensemble achieves the highest ROC-AUC (0.8776), representing a 1.00% improvement over supervised-only XGBoost (0.8689).

Figure 6 shows that the hybrid ensemble strikes the best balance between the graph-based pattern recognition of the unsupervised component and the discriminative power of the supervised component. The 30% unsupervised contribution keeps relational fraud signals intact, while the 70% supervised contribution does a great job of classifying them.

6.5 Discussion

The experimental results show that the proposed hybrid framework meets all primary research objectives. The framework achieved ROC-AUC of 0.8776 and recall of 62.12% on the IEEE-CIS test set, exceeding the targets of 0.85 and 60% specified in Research Objective O3. The unsupervised component achieved PR-AUC of 0.2405, surpassing the 0.20 target in Research Objective O2. Most significantly, the framework outperformed the primary baseline of Jiang et al. (2023) by 2.57% in ROC-AUC, fulfilling Research Objective O5. The construction involving multiple entity types for card-email-device, card-address-BIN, and card-product-type interactions delivers more detailed information on entity relationships than the method used by Jiang et al., thereby confirming the central hypothesis that multiple entity types benefit the task of fraud detection.

The entity risk scoring feature came out as the biggest unsupervised feature with a weightage of 70% for optimisation. This result highlights the efficiency of Bayesian-smoothed historical fraud rates for entity-level detection signals and supports the claim made by Vanini et al. (2023) on the application potential for entity-level patterns as fraud indicators. The proposed framework extends this insight through graph-based propagation of fraud signals across multi-hop neighbourhoods, enabling detection of fraud rings operating through intermediary accounts invisible to single-entity analysis. The hybrid ensemble approach verifies the approach suggested by Mutemi and Bacao (2024), as they noticed the integration approach involving graphs and ensemble methods as an uncharted research field.

The design had several experimental restrictions that should be noted. Starting with a class imbalance ratio of 27.58:1 and a PR-AUC of 0.5275, precision became a concern for high-recall thresholds, for which 2,809 false positives (5.85% FPR) would have to be manually checked. This model operates with the hypothesis that scammers would regularly usetraceable attributes such as emails and device ID; more sophisticated criminal organizations withnewer personas would be filtered out based on this entity-based methodology and therefore would be impermeable to this filtering. The Isolation Forrest model had the lowest PR-AUC of 0.1122 and potentially could be aided with other methods such as autoencoders (Abbassi et al., 2024) and Local Outlier Factor (Vanini et al., 2023). The static representation is inadequate for describing dynamic fraud process development over time as outlined by Lee et al. (2024). The lack of SHAP-based explainability is a restriction for applying this design to scenarios with reporting and explanation needs for isolated transactions.

The results confirm the viability of graph-founded feature engineering and hybrid ensemble method detection for e-commerce fraud analysis. It is clear that the multi-view approach offers benefits over single-view methods with quantifiable advantages and supports the research gap observed during the literature review. A recall rate above 62.12% demonstrates that the system is able to identify and flag the predominant instances of fraud accurately, and a network comprising 30% unsupervised and 70% supervised methods outperforms each alone. Future research should be focused on SHAP analysis and the study of dynamic graphs to identify changing trends and on employing more effective anomaly detectors than Isolation Forests.

7 Conclusion and Future Work

This chapter highlights the research findings and discusses the achievement of research objectives, the research significance and contribution to the field, and limitations with hints for further research.

7.1 Summary and Objectives Achievement

This research addressed the question: "To what extent does the integration of graph-based feature engineering with Isolation Forest algorithms improve the detection accuracy of fraudulent transactions in the IEEE-CIS e-commerce transaction dataset?" The study was meant to fill the gap identified within the literature review process wherein current methods tend to compromise between structure learning and the capability for unsupervised detection.

Five objectives were employed for research:

- **O1**: Constructing multi-view transaction graphs
- **O2**: Achieving unsupervised PR-AUC exceeding 0.20
- **O3**: Attaining ROC-AUC above 0.85 and recall above 60%
- **O4**: Providing interpretable feature analysis
- **O5**: Exceeding the baseline of Jiang et al. (2023) at AUC 0.8556

The proposed framework processed 590,540 e-commerce transactions through a multi-phase architecture comprising three complementary graph views capturing card-email-device, card-address-BIN, and card-product-type relationships, combined into a unified graph of 16,810 nodes and 120,924 edges. Graph-based features including multi-hop fraud exposure, fraud propagation scores, and Bayesian-smoothed entity risk ratings were extracted and integrated with a hybrid ensemble combining 30% unsupervised signals with 70% supervised gradient boosting classifiers.

The framework successfully achieved four of five research objectives. Objective O1 was fulfilled through the construction of a multi-view heterogeneous graph capturing seven entity types. Objective O2 was achieved with the combined unsupervised score attaining PR-AUC of 0.2405, exceeding the 0.20 target. Objective O3 was achieved on both metrics, with ROC-AUC of 0.8776 surpassing the 0.85 target and recall of 62.12% exceeding the 60% threshold. Objective O5 was achieved with the framework outperforming Jiang et al. (2023) by 2.57% in ROC-AUC. Objective O4 was partially achieved through weight-based interpretability identifying entity risk scores as contributing 70% of unsupervised signal strength, though the planned SHAP-based instance-level explainability was not completed.

7.2 Significance and Contribution

This research work provides several significant contributions to the field of financial fraud detection and tackles critical gaps within the literature and application process.

7.2.1 Academic Contribution

The core scholarly contribution is to show that multi-view graph learning and unsupervised/supervised learning ensembles can be more effective than traditional approaches based on single-view graphs. As illustrated in Table 7 above, the current framework obtains the ROC-AUC score of 0.878 on the dataset, therefore registering a 2.57% relative improvement over Jiang et al. (2023)'s method for fraud detection on the same dataset with their model known as the UAAD-FDNet. This finding affirms the research hypothesis that multiple graph types (identity, spatial, and behavior graphs) offer fraud detection cues unobtainable from traditional approaches based on graphs drawn from a singular type. The current study builds on the finding from the systematic review carried out by Mutemi and Bacao (2024), which found the graph-based and ensemble methods field to be unexplored research.

7.2.2 Methodological Contribution

The study offers a new risk scoring method at the entity level via Bayesian Smoothing, a method that is able to overcome the issue of limited transaction records for entities. This method was found to be the best unsupervised feature and was given a weightage of 70% while combining all optimized features; this validates that entity-level features offer stronger fraud indicators than transaction-level anomaly detection methods. The computation for fraud exposure at the multiple-hop levels (1-hop, 2-hop, and 3-hop) and the Fraud Propagation algorithm allows for the identification of fraud rings involving traditional based entities.

7.2.3 Practical Contribution

From a practical standpoint, this framework illustrates that strong fraud detection performance can be obtained without using complex deep learning models and a large amount of labelled data. The semi-supervised model with 30% graph signals obtained using unsupervised learning and 70% classifiers obtained using supervised learning has advantages for model usage: lower labelling cost because of semi-supervised learning, efficiency for enabling real-time processing (< 10ms per transaction), and weights for model explainability for regulatory purposes. Also, finding that device entities had the most risk proportion (8.75%) gives key information for a fraud prevention team choosing device fingerprint methods.

7.2.4 Comparison with State-of-the-Art

The proposed framework is the only approach assessed for the IEEE-CIS dataset that brings together multi-view graph formation learning, anomaly detection with a focus on unsupervised

learning, and ensemble combining with a focus on hybrid combining methods. Although better absolute values are obtained using supervised ensemble combining in Almalki and Masud (2025), graphs and unsupervised learning are missing. The proposed approach therefore fills a gap in research where a trade-off had to be made.

7.3 Limitations

There exist some limitations in this study.

The experiments were conducted using a single data set available from Vesta Corporation. Thus, any generalizations regarding e-commerce settings and other finance-related areas are challenged. The fact that the IEEE-CIS data set has a 3.5% level of fraud far exceeds any observed levels for operational systems, typically below 1%, potentially inflating performance metrics compared to real-world deployment scenarios.

This model assumes that fraudsters use repeatable identifiable information such as email address domains and devices. A sophisticated fraud scheme using novel identities for a single transaction would evade a graph model because no relationships would be available for examination. This remains a flaw inherent with entity graph models.

The static graph formation only describes relationships for a single moment in time and fails to depict the development of fraudulent relationships with time. A fraud ring that emerges for a short while before disbanding would not be represented with changes in graph structure.
A PR-AUC of 0.5275 may be considered competitive for such a class imbalance ratio of 27.58:1. This signifies a precision problem for a high recall. A false positive rate of 5.85% leads to about 2,809 true transactions being selected for a manual check.

This Isolation Forest component contributed to the lowest PR-AUC of 0.1122. This indicates that while this traditional anomaly detection technique may be of little use when used to identify graph features, other techniques may be more effective.

This weakness in being applicable for regulation-compliant environments for providing explanations for decisions made for each instance restricts attainment level for Goal O4.

7.4 Future Work

This research opens several important directions for future work.
Integration of Explanation: The important integration would be to finish the explainability part by adding SHAP or LIME. This would aid fraud analysts in better comprehension regarding why specific transactions are identified as fraud and assist financial institutions with regulatory requirements for compliance.

Temporal Graph Dynamics: Adding temporal graph networks as indicated in a recent work by Lee et al. (2024) would open avenues for analyzing changing trends of fraud and discovering novel attack methods based on graph dynamics. This would overcome the problem associated with graph concepts that were static in nature and would identify fraud rings with a short life cycle.

Cross-Domain Testing: Testing the multi-view graph approach using data from other payment processing companies, regions, and sectors would demonstrate that the solution can be applied beyond the IEEE-CIS dataset.

Enhanced Anomaly Detection: To further enhance unsupervised signals, the Isolation Forests portion could be replaced with AutoEncoders (Abbassi et al., 2024), Local Outlier Factor Ensemble methods (Vanini et al., 2023), or graph neural networks for anomaly detection.

Federated Learning: The study by Khan et al. (2025) explored whether one can extend this framework for federated deployment such that several institutions come together for training models while maintaining data privacy.

Real-time Streaming: A modification of the framework to enable real-time streaming integration would facilitate scoring of fraud within milliseconds. This would require smart feature caching and update graph methods with small steps.

References

Abbassi, H., El Mendili, S. and Gahi, Y. (2024) 'Real-Time Online Banking Fraud Detection Model by Unsupervised Learning Fusion', *IEEE Access*, 12, pp. 45678-45692. doi: 10.1109/ACCESS.2024.3387654.

Almalki, S. and Masud, M. (2025) 'Credit Card Fraud Detection Using Stacking Ensemble with Explainable AI on IEEE-CIS Dataset', *Journal of King Saud University - Computer and Information Sciences*, 37(1), pp. 102-118. doi: 10.1016/j.jksuci.2024.102234.

Breskuvienė, D. and Dzemyda, G. (2024) 'Enhancing Credit Card Fraud Detection: Highly Imbalanced Data Case', *Machine Learning with Applications*, 15, pp. 100523. doi: 10.1016/j.mlwa.2024.100523.

Cui, Y., Zhang, L., Wang, H. and Chen, X. (2025) 'FraudGNN-RL: Graph Neural Networks with Reinforcement Learning for Adaptive Fraud Detection', *Expert Systems with Applications*, 242, pp. 122876. doi: 10.1016/j.eswa.2024.122876.

Jiang, S., Dong, R., Wang, J. and Xia, M. (2023) 'Credit Card Fraud Detection Based on Unsupervised Attentional Anomaly Detection Network', *Systems*, 11(6), pp. 305. doi: 10.3390/systems11060305.

Khan, S., Ahmad, N., Ismail, M. and Ali, T. (2025) 'Secure and Transparent Banking: Explainable AI-Driven Federated Learning Model for Financial Fraud Detection', *Computers & Security*, 138, pp. 103654. doi: 10.1016/j.cose.2024.103654.

Lee, Y., Kim, J., Park, S. and Choi, H. (2024) 'Temporal Graph Networks for Graph Anomaly Detection in Financial Networks', *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12), pp. 14523-14531. doi: 10.1609/aaai.v38i12.29876.

Lou, C., Wang, Y., Li, J., Qian, Y. and Li, X. (2025) 'Graph Neural Network for Fraud Detection via Context Encoding and Adaptive Aggregation', *Knowledge-Based Systems*, 287, pp. 111432. doi: 10.1016/j.knosys.2024.111432.

Mutemi, A. and Bacao, F. (2024) 'E-Commerce Fraud Detection Based on Machine Learning Techniques: Systematic Literature Review', *Big Data and Cognitive Computing*, 8(4), pp. 42. doi: 10.3390/bdcc8040042.

Orelaja, O.A. and Adeyemi, A.O. (2024) 'Real-Time Fraud Detection in E-Commerce Transactions Using Graph Neural Networks', *Journal of Financial Crime*, 31(3), pp. 678-695. doi: 10.1108/JFC-09-2023-0234.

Talukder, M.A., Islam, M.M., Uddin, M.A. and Hasan, K.F. (2024) 'An Integrated Multistage Ensemble Machine Learning Model for Fraudulent Transaction Detection', *Decision Analytics Journal*, 10, pp. 100396. doi: 10.1016/j.dajour.2024.100396.

Vanini, P., Rossi, S., Zvizdic, E. and Domenig, T. (2023) 'Online Payment Fraud: From Anomaly Detection to Risk Management', *Financial Innovation*, 9, pp. 66. doi: 10.1186/s40854-023-00470-w.

Vijayanand, D. and Smrithy, G.S. (2025) 'Explainable AI-Enhanced Ensemble Learning for Financial Fraud Detection in Mobile Money Transactions', *Applied Soft Computing*, 149, pp. 110987. doi: 10.1016/j.asoc.2024.110987.
Xiang, Y., Chen, Z., Liu, W. and Zhang, H. (2023) 'GTAN: Gated Temporal Attention Network for Credit Card Fraud Detection', *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 2834-2843. doi: 10.1145/3583780.3615075.
Zioviris, G. and Kolomvatsos, K. (2024) 'An Intelligent Sequential Fraud Detection Model Based on Deep Learning', *Neural Computing and Applications*, 36(8), pp. 4123-4139. doi: 10.1007/s00521-023-09234-6.