

Explainable AI in Financial Fraud Detection: Evidence from the IEEE-CIS Fraud Dataset

Delia DIACONU

*Bucharest University of Economic Studies, Bucharest, Romania
delia.diaconu@fin.ase.ro*

Daniel Traian PELE*

*Bucharest University of Economic Studies, IDA Institute Digital Asset, Bucharest, Romania
Institute for Economic Forecasting, Romanian Academy, Bucharest, Romania*

**Corresponding author, danpele@ase.ro*

Abstract. *Financial fraud detection systems increasingly rely on advanced machine learning models, whose decisions should remain easily interpretable, in order to make transparent decisions and to align with regulatory requirements. In this study, we explore explainable AI methods for fraud detection, using the IEEE-CIS Fraud detection database. Thus, we compare the results obtained using tabular models (XGBoost, CatBoost) with those generated by graph-based models (GraphSAGE) and observe that although gradient boosting models have a better performance - measured in terms of ROC-AUC -, graph-based models capture relational structures, which cannot be described by tabular models. We propose a multi-level explainability framework, which combines methods of explaining model variables (SHAP values) with graph analysis methods, such as edge and node ablation. The results highlight a trade-off between discriminatory power and the interpretation of interactions between network nodes. Thus, although tabular models are better at identifying risk at feature level, graph-based models allow for a better interpretation of transactional fraud patterns. These results demonstrate that explainable AI in fraud detection should not be based on a single model, but it should combine several types of models that capture both the relational structure and the predictive power at feature level.*

Keywords: Explainable Artificial Intelligence (XAI), Financial Fraud Detection, IEEE-CIS Fraud Dataset, Graph Neural Networks (GNNs), GraphSAGE, Heterogeneous Graphs.

Introduction

Financial fraud has become a major challenge in the modern digital economy, driven by the rapid growth of electronic payments, e-commerce platforms, and digital identification systems. The increase in the volume and velocity of financial transactions, combined with increasingly sophisticated fraud schemes, have supported the widespread adoption of machine learning models in risk analytics and fraud prevention systems. Traditional rule-based and expert scenario systems are no longer sufficient for detecting complex fraud patterns, leading financial institutions to implement more and more predictive models based on data.

In recent years, gradient boosting models, such as XGBoost and CatBoost, have gained popularity among transaction fraud detection methods due to their very high discriminatory power, robustness to noise, and ability to model non-linear relationships between transaction characteristics. These models have produced state-of-the-art results when used on multiple public and proprietary databases, including the IEEE-CIS Fraud Detection database. However, these models rely on a flattened representation of complex fraud processes, which limits their ability to capture fraud paths across interconnected entities. In the real world, fraud in financial systems

rarely occurs in isolation, and rather it develops in a network of shared identities, devices, addresses, and reuse of the same infrastructure, coordinated behaviors, and common patterns of interactions between multiple entities.

The relational nature of fraud has supported the rapid adoption of graph-based approaches, where transactions and associated entities are represented as network nodes, and interactions and common attributes are represented as network edges. Graph Neural Networks propose a framework for learning from these relational structures by passing messages between entities and aggregating information from neighboring nodes. By encoding relational inductive biases, GNNs allow models to learn interactions that are not accessible to classical tabular models. At the same time, the increasing complexity of machine learning systems has raised fundamental concerns about explainability, transparency, and trustiness, especially in regulated areas such as finance. Regulatory frameworks, model risk management standards, and governance requirements demand not only high accuracy, but also interpretable and auditable results and accountable decision-making systems. Thus, explainable AI has become a key requirement for implementing fraud detection models in operational environments. Most explainability methods in fraud detection focus on feature- and attribute-level analyses in tabular models, using techniques such as SHAP and LIME. While these methods provide valuable insights about the importance of variables, they do not capture relational structures or dependencies. At the same time, explainability methods in graph models focus on local, instance-level explanations, facing stability and scalability challenges in heterogeneous graphs in financial systems.

The paper proposes a comparative study, from a multiple perspective, for fraud detection explainability. Using the IEEE-CIS Fraud detection database, we test tabular models (XGBoost, CatBoost) and graph-based models (GraphSAGE) in a unified framework. The paper does not aim to describe the problem as a competition to choose the best model in terms of performance, but rather investigates how different models produce different and complementary results in terms of explainability of artificial intelligence. Thus, explainability methods are analyzed by integrating attribute-level interpretation with global interpretation, within the graph, based on edge ablation and hop ablation techniques. By combining the evaluation of model performance with analysis of explainable AI, the paper shows that explaining fraud detection should not be approached as a simple optimization problem, but rather as a multidimensional framework that integrates complementary perspectives: statistical learning and relational structure modeling.

Literature review

Machine learning models have become a standard approach in detecting financial fraud across payment systems, e-commerce platforms, and digital financial services, where predictive modelling faces challenges such as large class imbalances, non-stationarity of fraud patterns, and adversarial behaviour. Classical methods include logistic regression, decision trees, random forests, and neural networks, while gradient boosting algorithms such as XGBoost, LightGBM, and CatBoost have consistently demonstrated top performance when used on tabular databases (*Bahnsen et al., 2015*). In an operational environment, these are accompanied by post-hoc explainability methods, notably SHAP (*Lundberg & Lee, 2017*) and LIME (*Ribeiro et al., 2016*), which provide information about the contribution of features at both the local and global levels. These methods have begun to be widely adopted by financial institutions for regulatory compliance, model governance, and audit.

The IEEE-CIS Fraud Detection database has become a benchmark for such approaches, where gradient boosting models combined with extensive feature engineering have achieved top

performance. However, several studies highlight a fundamental limitation of tabular representations: transactions are modelled as independent observations, while in the real world, fraud occurs based on coordinated behaviors, across multiple entities, shared devices, ITP addresses, emails, identities. Thus, tabular models can achieve very good performance, but have limited capabilities in capturing relational structures, infrastructure reuse, and similar fraud trajectories.

Graph-based models address this limitation by modelling financial systems as networks of transactions and entities. Early research in graph analysis demonstrated the effectiveness of relationship modelling to detect fraud rings, fraudulent groups, and anomalous structures (*Pandit et al., 2007; Akoglu et al., 2015*). With the emergence of Graph Neural Networks (GNNs,) relational learning evolved to an end-to-end representation of graph learning. *Hamilton et al. (2017)* introduced GraphSAGE, an inductive framework for aggregating information from neighbouring nodes, allowing for generalization of behaviour to unseen nodes and large dynamic graphs. This property is particularly relevant for financial systems, where new accounts, devices, and identities are constantly entering the graph.

However, the use of graphs in fraud detection poses unique challenges, including intentional camouflage strategies and heterogeneous relational structures. *Dou et al. (2020)* proposed CARE-GNN, with the aim of improving the robustness of the model against camouflaged fraudsters by filtering and reweighting information from the neighborhood. Subsequent work has explored graph learning under low-homophily conditions, including label-enhanced propagation mechanisms, which stabilize learning in large and noisy graphs (*Li et al., 2023*). More recent studies have also investigated self-explainable graphs for fraud detection, which integrate predictive and explainability objectives, with the aim of overcoming the limitations of post-hoc interpretation in high-stakes domains (*Li et al., 2024*).

Several works have specifically constructed heterogeneous transaction–identity–device graphs from real-world financial datasets, including IEEE-CIS, demonstrating that relational representations reveal fraud clusters, shared infrastructure, and coordinated behaviours that are not visible in tabular representations alone. These studies consistently show that graph modelling provides structural insights even in cases where predictive performance gains over tabular models are modest, supporting the view that relational learning offers complementary value rather than purely accuracy-driven improvements.

Explainability for graph-based models has become an active research area. *Ying et al. (2019)* introduced GNNExplainer, which generates local explanations by learning edge and feature masks that identify compact explanatory subgraphs for individual predictions. *Luo et al. (2020)* proposed PGExplainer, a parametric explainer that learns a global explanation model capable of producing explanations across multiple instances, improving scalability and amortization compared to per-instance optimization. Subsequent methods such as SubgraphX (*Yuan et al., 2021*) and GraphMask (*Schlichtkrull et al., 2021*) further advanced subgraph-based and masking-based explanation paradigms. Recent surveys (*Yuan et al., 2020*) emphasize that most GNN explainability methods remain predominantly local, often unstable, computationally expensive, and difficult to operationalize in large-scale heterogeneous graphs. This has motivated growing interest in global and structural explainability, including relation-type importance analysis, structural ablation, motif discovery, community-based interpretation, and topological sensitivity analysis. These approaches aim to explain model behavior in terms of network mechanisms rather than individual predictions, aligning more closely with the needs of financial risk governance,

where understanding systemic patterns, infrastructure reuse, and relational dependencies is often more valuable than isolated instance-level explanations.

In this context, comparative studies between tabular and graph-based models increasingly emphasize that model evaluation should not be limited to predictive metrics alone. While gradient boosting models often achieve superior ROC-AUC on tabular benchmarks, graph-based models provide complementary explanatory value by revealing relational structures and coordinated fraud mechanisms that cannot be captured through feature attribution alone. This perspective motivates a hybrid evaluation framework in which tabular models contribute to predictive strength and stability, while graph-based models contribute to relational understanding, structural interpretability, and systemic risk insight, forming a more complete foundation for explainable AI in financial fraud detection.

Methodology

The study presents a comparative analysis of the models used and an explainability framework designed to describe how different perspectives of data representation – tabular and graph-based models – support fraud detection and data interpretation from financial transactions. More than proposing a single solution, the methodology integrates several types of models and complementary explainability measures into a unified framework, and proposes a comparison between statistical learning in feature space and the relational learning in network space.

Two families of models are considered. In the tabular representation, transactions are represented as independent observations, described by a fixed-length feature vectors. In the graph representation, transactions are presented through relational structures that aggregate dependencies between different entities. This dual representation allows a comparison between classical feature learning and relational modelling, reflecting two fundamentally different perspectives of fraud detection.

For the tabular representation, two gradient boosting benchmark models were tested - XGBoost and CatBoost. Both models are trained and vectorized on transactional data and optimized with a binary target. XGBoost is trained using one-hot encoding for categorical variables, while CatBoost allows for the native use of categorical variables, without additional transformations.

Both XGBoost and Catboost models are based on gradient boosting decision trees algorithm, where the final prediction is built based on additive ensemble of weak learners:

$$\hat{y}_i = \sum_{m=1}^M f_m(x_i), \quad f_m \in \mathcal{F} \quad (1)$$

where f_m denotes an individual decision tree and $\mathcal{F}$ represents the space of regression trees.

The training objective minimizes a regularized empirical loss:

$$\mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{m=1}^M \Omega(f_m) \quad (2)$$

where $l(\cdot)$ is the binary classification loss function and $\Omega(f_m)$ penalizes model complexity to prevent overfitting. At each boosting iteration t , a new tree f_t is added to minimize the following objective:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t). \quad (3)$$

The dataset is split into training and validation subsets using a standard 70/30 random split, with the training set used for model fitting and the validation set for performance evaluation and explainability analysis. This setup reflects standard supervised learning practice in fraud detection and ensures a clear separation between training and evaluation data.

In parallel, a graph-based representation is constructed in which transactions are modelled as nodes connected to multiple types of entity nodes through heterogeneous relations. On this relational structure, a heterogeneous GraphSAGE model is trained for node-level fraud classification. GraphSAGE performs inductive representation learning through neighbourhood aggregation, allowing each transaction node to learn an embedding that integrates information from its connected entities and relational neighborhood. Message passing occurs across multiple node and edge types, enabling the model to capture relational dependencies between transactions, identities, devices, addresses, and other associated entities.

The transaction network is modelled as a heterogeneous graph $G = (V, E, \mathcal{R})$, where V denotes the set of nodes, $E \subseteq V \times \mathcal{R} \times V$ is the set of edges, and $\mathcal{R}$ is the set of relation types. Nodes correspond to transactions and associated entities (e.g. identities, devices, addresses), and each node $v \in V$ is described by a feature vector $x_v \in \mathbb{R}^d$. GraphSAGE learns node embeddings through iterative neighbourhood aggregation. Let $h_v^{(0)} = x_v$ denote the initial feature representation of node v . At layer k , the representation is updated as:

$$h_v^{(k)} = \sigma \left(W^{(k)} \cdot AGG^{(k)} \left(\left\{ h_u^{(k-1)} : u \in \mathcal{N}(v) \right\} \right) \parallel h_v^{(k-1)} \right) \quad (4)$$

where

- $\mathcal{N}(v)$ denotes the neighbourhood of node v ,
- $AGG^{(k)}(\cdot)$ is a permutation-invariant aggregation function (in this case, the mean),
- $W^{(k)}$ is a trainable weight matrix,
- $\parallel$ denotes concatenation,
- $\sigma(\cdot)$ is a non-linear activation function.

After k layers, the final embedding $z_v = h_v^{(k)}$ is used for binary fraud classification via

$$\hat{y}_v = \sigma(w^T z_v) \quad (5)$$

where $\hat{y}_v$ represents the predicted fraud probability for node v .

In contrast to the tabular setting, for the graph-based model, training and evaluation are performed using a node-level data split rather than a dataset-level partition. Transactions are represented as nodes in a heterogeneous graph, and the full graph structure – all nodes and edges – is preserved during both training and validation. A subset of transaction nodes is assigned to the training set and another disjoint subset to the validation set using boolean masks, while supervision and loss computation are restricted only to the labeled nodes in the corresponding split. This strategy allows the model to exploit the complete relational structure of the network during learning, while preventing label leakage between training and validation through node-level separation.

All models are trained using supervised learning with binary classification objectives. XGBoost and CatBoost are optimized using standard classification losses and evaluated using ROC-AUC as the primary performance metric. The GraphSAGE model is trained using node-level binary cross-entropy loss on transaction nodes, with embeddings for all node types learned jointly through message passing. Evaluation is conducted on validation transaction nodes using the same ROC-AUC metric, enabling consistent performance comparison across model classes. In addition to ROC-AUC, model performance is evaluated using precision, recall and F1-score, reflecting practical operational requirements in real-world fraud detection systems. Given the strong class imbalance and asymmetric cost structure of fraud detection, where false negatives correspond to direct financial losses and false positives generate operational costs and customer friction, threshold dependent metrics provide more actionable insight than threshold-independent measures alone. This evaluation framework enables a more realistic assessment of model utility in deployment settings, where optimizing detection effectiveness must be balanced against alert volume and investigation capacity. Performance curves across decision thresholds are analyzed to identify operating points that maximize practical effectiveness rather than purely statistical discrimination.

Explainability is treated as a central component of the methodology rather than as a post-hoc add-on. For tabular models, explainability is performed using SHAP (Shapley Additive Explanations), which provides feature-level attributions based on cooperative game-theoretic principles. SHAP values are computed to obtain global feature importance rankings and to characterize the contribution of individual variables to fraud risk predictions, enabling the identification of dominant risk drivers in feature space.

For a given prediction $f(x)$, SHAP expresses the model output as an additive decomposition:

$$f(x) = \phi_0 + \sum_{j=1}^p \phi_j \quad (6)$$

where:

- ϕ_0 is the expected model output over the background dataset,
- ϕ_j represents the contribution of feature j to the prediction,
- p is the number of input features.

Each SHAP value ϕ_j is computed as:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|! (p - |S| - 1)!}{p!} [f_{S \cup \{j\}}(x_{S \cup \{j\}}) - f_S(x_S)] \quad (7)$$

where:

- F is the full feature set,
- S represents a subset of features not containing j ,
- $f_S(\cdot)$ denotes the model evaluated using only features in subset S .

For the graph-based model, explainability focuses on structural and relational mechanisms rather than feature attribution. Two complementary global structural explainability strategies are employed. First, an edge-type ablation procedure is applied, in which each relation type in the heterogeneous graph is removed independently and the model is re-evaluated on the modified graph. The resulting performance degradation quantifies the global importance of each relation

type, producing a relational importance ranking that identifies which entity interactions contribute most to fraud detection. This provides a form of global relational explainability, revealing which network connections carry predictive signal at the system level.

Let:

- $G = (V, E, \mathcal{R})$ be the heterogeneous graph,
- $\mathcal{R} = \{r_1, \dots, r_m\}$ the set of relation types,
- $f(G)$ the trained GraphSAGE model.

For each relation type $r \in \mathcal{R}$, we define the modified graph: $G^{-r} = (V, E \setminus E_r)$, where $E_r \subset E$ are the edges of type r . The importance of relation r is quantified as:

$$\Delta_r = AUC(G) - AUC(G^{-r}) \quad (8)$$

where

- $AUC(G)$ is the baseline performance measured as the validation ROC-AUC of the trained model evaluated on graph G ,
- $AUC(G^{-r})$ is the performance after removing relation type r .

A larger Δ_r indicates greater global relational importance.

Second, a hop-based ablation strategy is used to analyze the role of neighborhood depth in relational learning. The graph is restricted to k -hop neighbourhoods around transaction nodes, with increasing values of k controlling the depth of message passing. Model performance is evaluated at each level, ranging from isolated nodes without relational information to multi-hop relational propagation. This analysis quantifies how predictive performance depends on relational distance and reveals the structural depth at which fraud-related information is concentrated in the network.

Let $\mathcal{N}_k(v)$ denote the k -hop neighbourhood of a transaction node v . We define the restricted graph to k -hop neighbourhood around transaction nodes as:

$$G^{(k)} = \text{Subgraph induced by } \bigcup_{v \in V_{txn}} \mathcal{N}_k(v).$$

Model performance at depth k is $AUC_k = AUC(G^{(k)})$, and the structural contribution of depth k is:

$$\Delta_k = AUC(G) - AUC(G^{(k)}). \quad (9)$$

Together, these explainability methods enable a multi-level interpretation of fraud detection models. Tabular models provide feature-based explanations that identify variable-level risk factors, while graph-based models provide structural explanations that reveal relational and network-level mechanisms. The methodology therefore supports a comparative explainability framework in which different model classes are treated not as competitors, but as complementary representations of fraud: one capturing statistical dependencies in feature space, and the other capturing relational dependencies in network space.

Data and Graph Construction

This study is based on the IEEE-CIS Fraud Detection dataset, a large-scale real-world transaction dataset released for the IEEE-CIS fraud detection challenge. The dataset consists of anonymized e-commerce transactions labeled as fraudulent or legitimate and reflects realistic operational conditions, including high dimensionality, heterogeneous data sources, and strong class imbalance. Fraudulent transactions represent a small minority of the total sample, reflecting real-world fraud detection environments where fraud prevalence is low but economically significant.

Each transaction is described by a high-dimensional feature space composed of multiple categories of variables. The C variables capture counting and behavioral consistency patterns, the D variables encode temporal delays and transaction timing structures, the V variables consist of anonymized numerical attributes derived from transactional patterns, and the ID variables represent identity and infrastructure-related information such as device identifiers, address proxies, and email domain attributes. Additional categorical variables describe transaction context, product categories, and payment-related attributes. This heterogeneous structure reflects the multi-source nature of fraud data, where risk emerges from the interaction of behavioral, infrastructural, and transactional signals.

Table 1. Dataset Description

Category	Description
Dataset name	IEEE-CIS Fraud Detection Dataset
Total observations	590,540
Fraudulent transactions	20,663
Fraud rate	3.50%
Total number of features	432

Source: Authors' own computation results.

The dataset includes a temporal variable, *TransactionDT*, which encodes elapsed time since a reference point and enables reconstruction of transaction ordering. Transformation of this variable into calendar-aligned temporal features allows the analysis of fraud evolution over time. Exploratory analysis indicates that fraud prevalence remains relatively stable over the observation period, without strong seasonal spikes or abrupt regime shifts, suggesting a structurally persistent fraud process rather than isolated temporal anomalies.

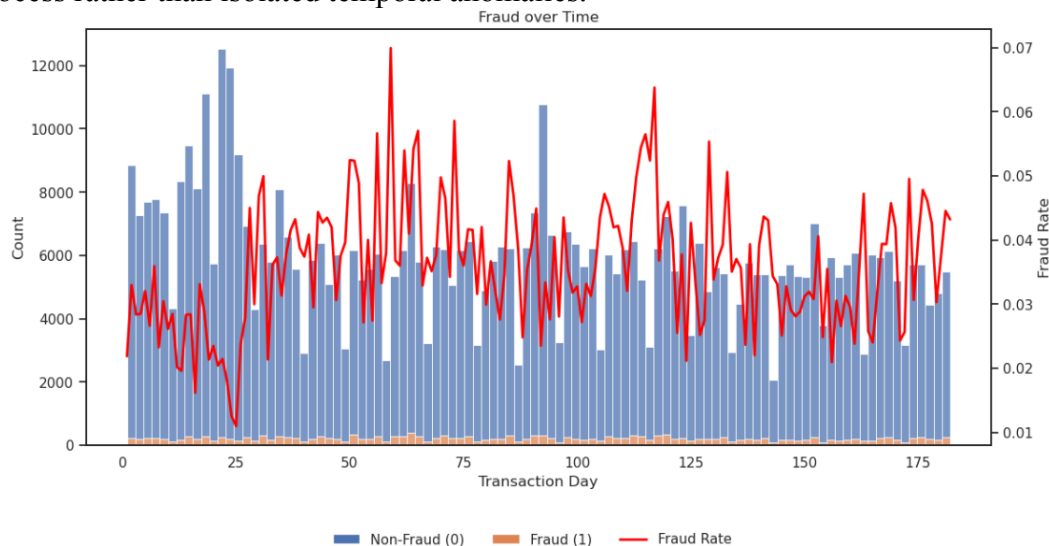


Fig. 1. Fraud rate evolution over time

Source: Authors' own computation results.

To capture relational dependencies, the dataset is transformed into a heterogeneous graph representation. Transactions are modeled as nodes connected to multiple types of entity nodes representing semantically meaningful abstractions of shared attributes. Identity-related variables

define identity entities, infrastructure-related variables define device and address entities, and communication-related variables define email domain entities. This approach transforms high-dimensional attribute spaces into structured relational entities, enabling explicit modelling of shared identity, shared infrastructure, and shared behavioral channels across transactions.

Edges represent relational associations between transactions and entities, such as a transaction being linked to a specific identity, device, address, or email domain. For each relation, a corresponding reverse relation is included to enable bidirectional information propagation. This construction yields a multi-relational, heterogeneous transaction network in which transactions are embedded within a structured entity ecosystem rather than treated as isolated events.

The resulting graph encodes relational dependencies such as identity reuse, infrastructure sharing, and behavioral overlap, enabling the emergence of relational clusters and coordinated patterns that are not observable in tabular representations. This graph representation allows neighbourhood aggregation and message passing to operate across heterogeneous entity types and relational structures.

By combining a high-dimensional tabular representation with a structured heterogeneous graph representation, the dataset supports a multi-representational analysis of fraud detection, enabling both statistical learning in feature space and relational learning in network space. This dual representation forms the basis for the comparative explainability framework described in this study.

Results and discussion

This section describes the empirical results obtained based on both types of models, integrating the assessment of discriminative power and the explainability analysis. The aim is to reflect the dual objective for fraud detection: effective risk mitigation and transparency in decision-making.

First, the results of the tabular models are described, which serve as good benchmarks for the performance analysis. The XGBoost model, optimized with cross validation, achieves the strongest results, where the best configuration is represented by 1000 trees, with depth 10. The categorical variables were encoded using one-hot encoding, thus building a multidimensional spatial of the vector of features, which allows the model to learn non-linear trajectories for class separation. The CatBoost model, trained on the same database, which uses the categorical variables in their raw, native form, achieves similar values in performance, but slightly below those generated by XGBoost. However, the results generated by CatBoost remain competitive and robust, in the context of a more compact representation of the features vector. The difference in performance between the two types of models is attributed more to the optimization and learning mechanism of the representations, than to the proportion of categorical variables in the dataset.

In addition to the classic discrimination measure - the ROC-AUC curve- , models performance is evaluated using precision and recall metrics, and threshold-dependent metrics, which are much more relevant in real-world fraud detection systems. In an environment where the fraud rate is very low, where we are talking about strongly imbalanced classes, the ROC-AUC measure alone does not provide a complete framework for evaluating operational performance. Practical performance is obtained by selecting an operational cut-off that balances the number of correctly detected frauds with the number of false positive alerts. This balance is essential in the operational environment, where missing fraud detection and burdening the workforce with many false alarms generate high costs. Thus, we observe that both models have high fraud detection power, with high recall values supported by acceptable precision rates.

Table 2. Model performance indicators

Model	ROC-AUC	Precision	Recall	F1-Score	Optimal Threshold
XGBoost	0.9763	0.8955	0.7509	0.8169	0.2055
CatBoost	0.9636	0.8915	0.7091	0.7899	0.2386

Source: Authors' own computation results.

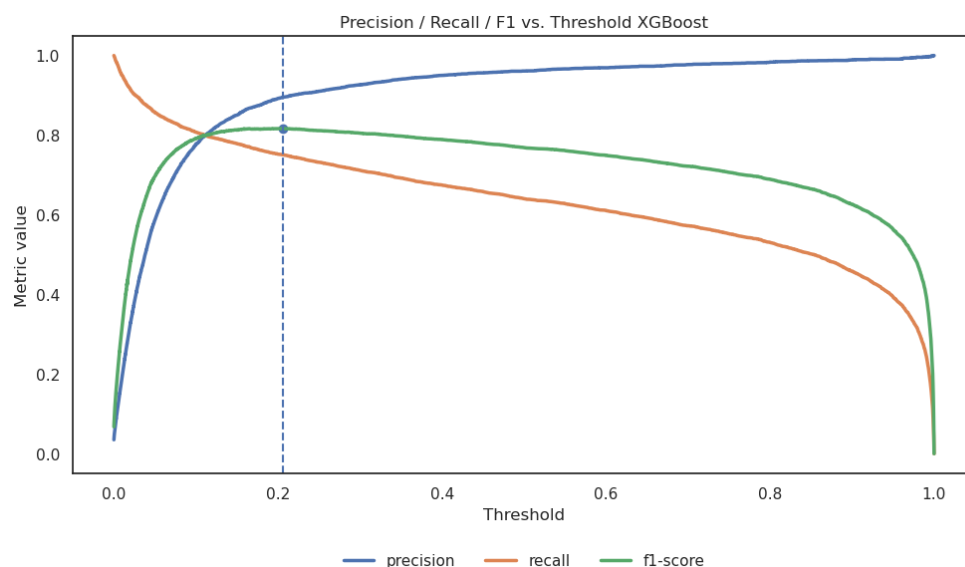


Fig. 2. Precision, Recall, F1-Score for XGBoost model

Source: Authors' own computation results.

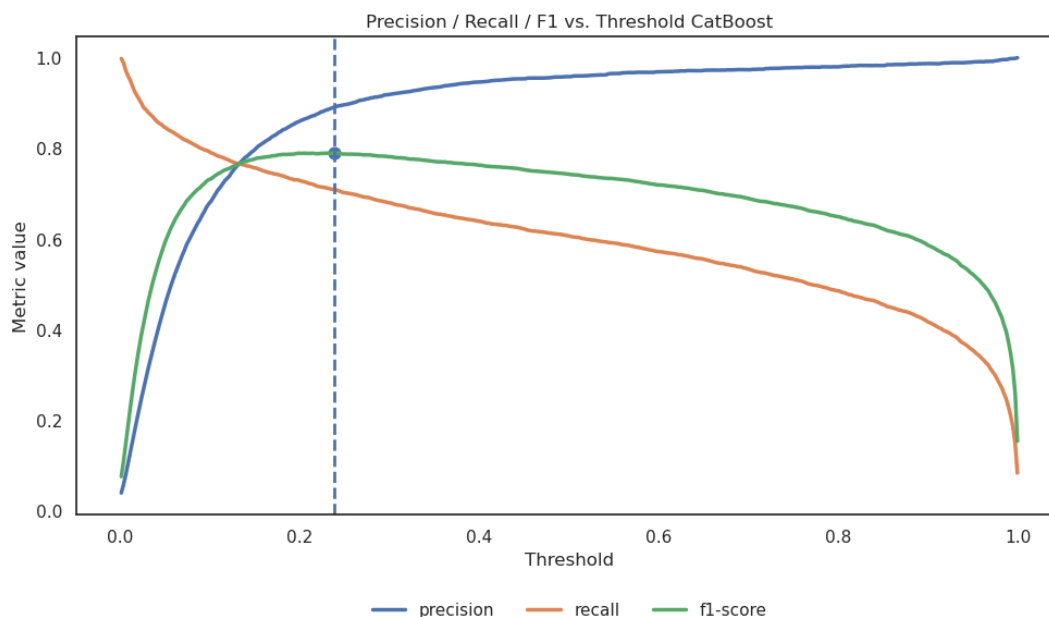


Fig. 3. Precision, Recall, F1-Score for CatBoost model

Source: Authors' own computation results.

Explainability at the feature level is addressed using SHAP (SHapley Additive exPlanations) for the tabular models. Global SHAP analysis reveals that fraud predictions are mainly driven by behavioural consistency features, identity-related attributes, and transaction pattern variables. These results illustrate the importance of behavioural regularity, infrastructure reuse, and anomalous dynamics in fraud detection. SHAP explanations are presented below for both the XGBoost and CatBoost models. The ranking of the most influential variables is largely consistent across the two models, indicating stable and robust explanatory patterns despite architectural differences. This consistency suggests that both models rely on similar behavioural and transactional signals when identifying fraudulent activity. For CatBoost, SHAP values for categorical variables are not computed in the same way as for numerical features. These variables appear in grey in the SHAP summary plot from Fig. 5 and require separate category-level interpretation. This way, an individual category-level analysis is needed to understand their contribution to fraud risk. Examples of this type of analysis are provided for two categorical variables in Fig. 6 and 7.

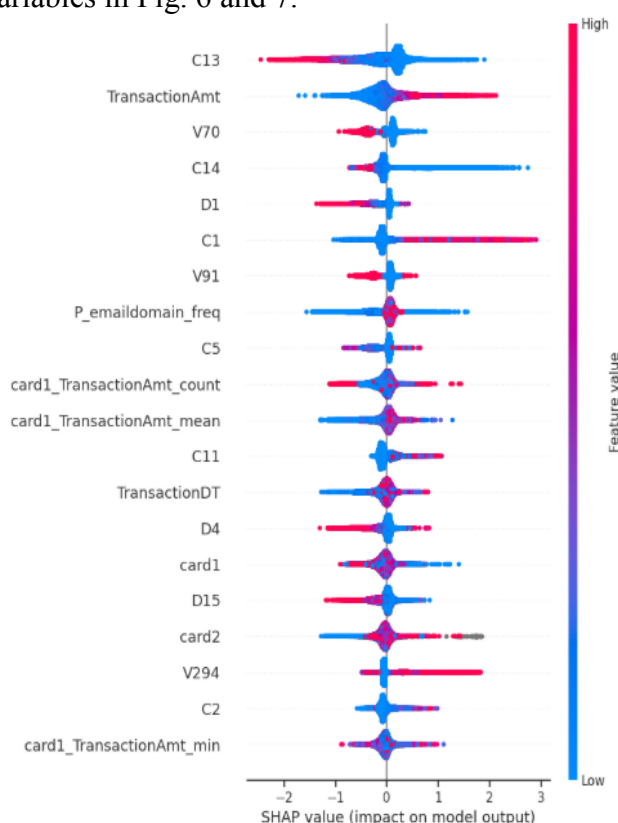


Fig. 4. SHAP Values - XGBoost model

Source: Authors' own computation results.

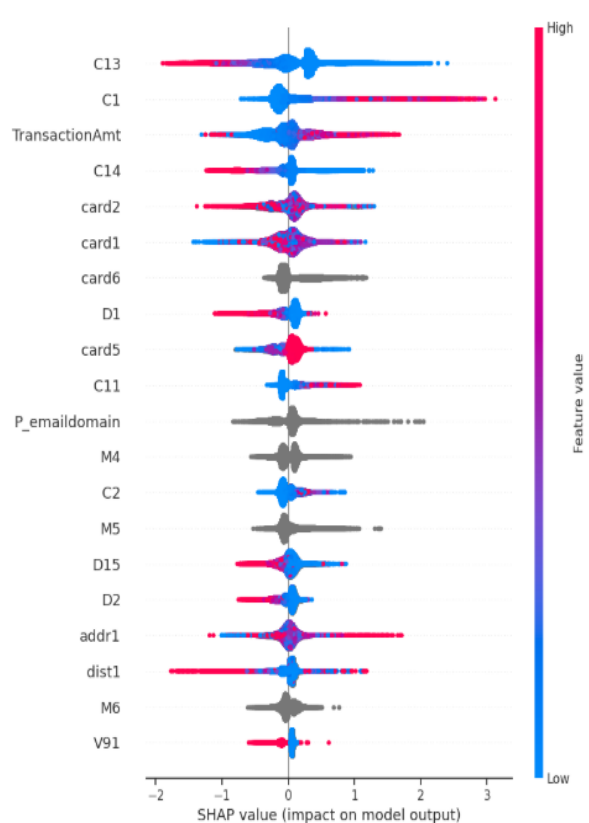


Fig. 5. SHAP Values - CatBoost model

Source: Authors' own computation results.

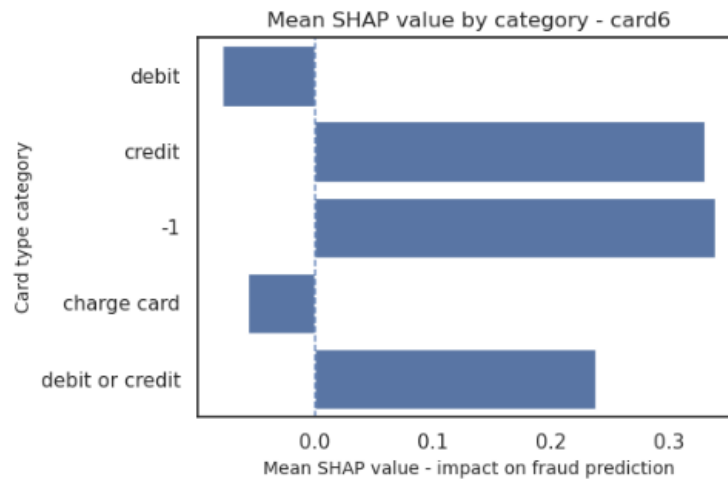


Fig. 6. SHAP Values – categorical var. *card6*

Source: Authors' own computation results.

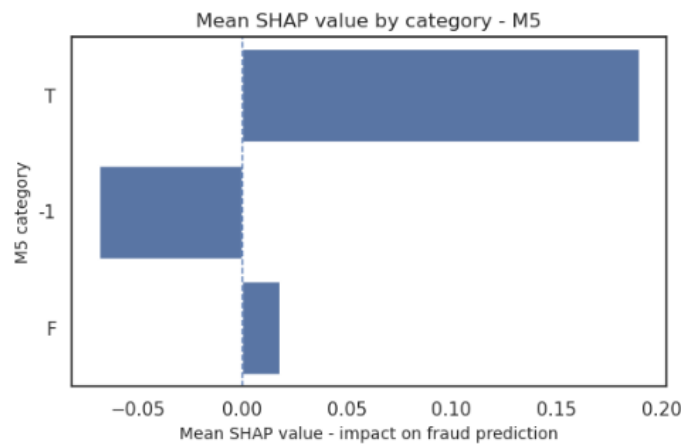


Fig. 7. SHAP Values - categorical var. *M5*

Source: Authors' own computation results.

Relational learning is evaluated using a heterogeneous GraphSAGE model trained on the graph representation of the dataset. The model shows lower predictive performance than the tabular gradient boosting models with ROC-AUC score equal to 0.8729, which reflects the natural trade-off between representation power and predictive efficiency. While graph representations capture relational structure and network effects, they also introduce higher modelling complexity and noise propagation, which can reduce top classification performance to highly optimized tabular models. The main value of the graph-based approach in this study is not peak classification performance, but its ability to model relational structure and provide network-level explainability. The graph representation encodes direct links between entities such as shared identities, shared infrastructure, and shared communication channels. This allows the model to capture fraud patterns that emerge from relationships, not only from individual transaction attributes.

Structural explainability is evaluate through performance-based ablation analysis. This is done by systemically removing edges and nodes and measuring the resulting change in model performance. The results described in Table 3 show that removing specific relation types leads to clear performance degradation, ensuring the construction of a global relational importance ranking. This ranking

describes which types of entity interactions contribute the most to fraud detection and provide system level insight into how fraud risk propagates through relational structures. Hop-based ablation shows that model performance drops significantly when no relational information is available ($k = 0$) – Table 4. When only first-hop neighbours are included ($k = 1$), performance fully recovers to the baseline level, while adding higher order neighbours does not improve performance. This indicated that the predictive signal in the graph is given by direct relational dependencies, while multi-hop relational propagation does not contribute with additional useful information.

Table 3. Edge-type ablation results

Source Node	Relation	Target Node	Base AUC	Ablated AUC	Delta AUC
id_31	rev_uses_id_31	transaction	0.8729	0.8426	0.0303
addr1	rev_uses_addr1	transaction	0.8729	0.8474	0.0255
P_emaildomain	rev_uses_P_emaildomain	transaction	0.8729	0.856	0.0169
transaction	uses_addr1	addr1	0.8729	0.8675	0.0054
DeviceType	rev_uses_DeviceType	transaction	0.8729	0.8702	0.0027
transaction	uses_DeviceType	DeviceType	0.8729	0.8713	0.0016
transaction	uses_P_emaildomain	P_emaildomain	0.8729	0.8717	0.0012
transaction	uses_id_31	id_31	0.8729	0.8718	0.0012

Source: Authors' own computation results.

Table 4. Node-type ablation results

Hop (k)	Base AUC	AUC (k-hop graph)	Delta AUC
k=0	0.8729	0.754	0.1189
k=1	0.8729	0.8729	0
k=2	0.8729	0.8729	0

Source: Authors' own computation results.

Together, these results show that graph-based models offer a complementary perspective to tabular models. Tabular approaches provide strong predictive performance and clear feature-level interpretability. Graph models, in contrast, provide structural and relational interpretability, revealing network-level mechanisms such as infrastructure sharing and coordinated behavioural patterns. Rather than competing, the two approaches offer different but complementary views of fraud risk: one based statistical feature learning. This supports a multi-level interpretation framework for fraud detection, where performance, structure, and explainability are jointly considered.

Limitations

The analysis has several limitations. First, the results are based on a single public database (IEEE-CIS). Its variables are largely anonymized, and cannot provide a business-level interpretation. Therefore, the explanations are based more on statistical significance, to the detriment of understanding the variables in the business domain.

Second, the comparison between tabular models and graph-based model is based on a fixed split train-validation. A different approach in choosing the validation sample, possibly split according to the time horizon, may lead to different results.

Finally, the computational cost and application of the models on a larger scale, in a real-world financial transaction network, has not been tested in detail and may generate difficulties and a greater complexity in implementation.

Conclusion

This study investigated explainable AI in financial fraud detection using both tabular and heterogeneous graph neural networks. The results showed that tabular models (XGBoost and CatBoost) generate high predictive performance and stable feature-level explanations, making them the benchmark for fraud detection in operational systems. SHAP values analysis shows that both types of models are based on similar groups of variables, mainly behavioural variables, identity-related attributes, and transactional patterns. This consistency of the results across the two models strengthens the confidence in the feature-level explanations in the tabular modelling framework.

The GraphSAGE model achieves a lower performance compared to gradient boosting models, but it brings added value by modelling the relational links between entities/transactions. Edge ablation shows that certain specific links, such as shared identities and the infrastructure used, play an important role in fraud detection. Hop ablation further indicates that the most important relationship is contained in first order neighbours, while higher order neighbours do not bring any additional information.

Overall, the results support the use of a multi-level modelling framework to explain fraud detection, where gradient boosting models have very high discriminatory power, and graph models bring relational context information at system level. This mixed approach provides a more informative, reliable and easily interpretable framework in fraud risk management.

Future research will focus on developing integrated models that combine tabular and graph models representations. This approach could lead to top predictive performance, complemented by information on the relational structure of the entities. Increased attention should be paid to the robustness of the results and explainability throughout time, in the context of data drift for transparency in financial systems and alignment with regulatory requirements.

Acknowledgements

This paper is supported through the project “IDA Institute of Digital Assets”, CF166/15.11.2022, CN760046/ 23.05.2023; the project “AI for Energy Finance (AI4EFin)”, CF162/15.11.2022, CN760048/23.05.2023, financed under the Romania’s National Recovery and Resilience Plan, Apel nr. PNRR-III-C9-2022-I8; and the Marie Skłodowska-Curie Actions under the European Union's Horizon Europe research and innovation program for the Industrial Doctoral Network on Digital Finance, acronym DIGITAL, Project No. 101119635.

This paper was co-financed by The Bucharest University of Economic Studies during the PhD program.

During the preparation of this work, the authors utilized ChatGPT 5 to assist with language refinement, and rephrasing of certain sections of the text. The generated or revised content was carefully reviewed, edited, and validated by the authors, who take full responsibility for the final content of the publication.

References

- Akoglu, L., Tong, H., & Koutra, D. (2015). Graph-based anomaly detection and description: A survey. *Data Mining and Knowledge Discovery, Springer*, V.29, pp. 626-688.
- Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications*, Vol. 42, Issue 19, pp. 6609-6619.
- Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., & Yu, P. S. (2020). Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. *ACM CIKM Conference on Information and Knowledge Management*, pp. 315-324.
- Hamilton, W. L., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Neural Information Processing Systems*.
- Li, Y., Chen, X., Zhang, J., et al. (2025). Graph neural networks for financial fraud detection: a review. *Frontiers of Computer Science*, Vol. 19(9).
- Li, Z., Wang, Y., Zhang, Q., et al. (2023). Label Information Enhanced Fraud detection against low homophily in graphs. *Proceedings of the ACM Web Conference*, pp. 406-416.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4766-4777.
- Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., & Zhang, X. (2020). PGExplainer: Parameterized Explainer for Graph Neural Network. *Neural Information Processing Systems*, pp. 19620-19631.
- Pandit, S., Chau, D. H., Wang, S., & Faloutsos, C. (2007). NetProbe: A fast and scalable system for fraud detection in online auction networks. *Proceedings of the 16th international conference on WWW*, pp. 201-210.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144.
- Schlichtkrull, M., De Cao, N., & Titov, I. (2021). Interpreting Graph Neural Networks for NLP With Differentiable Edge Masking. *ICLR Conference*.
- Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating Explanations for Graph Neural Networks. *Neural Information Processing Systems*.
- Yuan, H., Tang, J., Hu, X., & Ji, S. (2021). XGNN: Towards Model-Level Explanations of Graph Neural Networks. *Proceedings of the 40th International Conference on Machine Learning*, pp. 33674-33719.
- Yuan, H., Yu, H., Gui, S., Ji, S. (2020). Explainability in graph neural networks: a taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol, 45(5), pp. 5782-5799.