Journal of Intelligent Decision and Computational Modelling

www.jidcm.reapress.com

J. Intell. Decis. Comput. Model. Vol. 2, No. 1 (2026) 50–57.

Paper Type: Original Article

Detection of Fraudulent Transactions in the Cryptocurrency Market Using an Explainable Graph Isomorphism Network

Fatemeh Jalalzaei[1], Akbar Esfahanipour[2,*], Ali Reza Keivanimehr[1]

[1] Department of Management, Science, and Technology, Amirkabir University of Technology, Tehran, Iran; fatemeh.jalalzaei@aut.ac.ir; alireza.keivanimehr@aut.ac.ir.

[2] Department of Industrial Engineering & Management Systems, Amirkabir University of Technology, Tehran, Iran; esfahaa@aut.ac.ir.

Citation:

Received: 22 August 2025 *Revised: 19 October 2025* *Accepted: 08 January 2026*	Jalalzaei, F., Esfahanipour, A., & Keivanimehr, A. R. (2026). Detection of fraudulent transactions in the cryptocurrency market using an explainable graph isomorphism network. *Journal of intelligent decision and computational modelling, 2*(1), 50-57.

Abstract

Cryptocurrency markets are highly vulnerable to fraud due to the lack of comprehensive regulatory frameworks and the pseudonymous nature of transactions. This paper proposes a graph-based learning approach for detecting suspicious transactions in decentralized networks, with a focus on Bitcoin. We evaluate three Graph Neural Network (GNN) models: Graph Convolutional Networks (GCN), Graph Attention Networks (GAN), and Graph Isomorphism Networks (GIN). Experiments on the Elliptic dataset, which contains highly imbalanced data between legitimate and illicit transactions, show that GIN outperforms other models in detecting fraudulent patterns. To enhance interpretability and provide insights into the decision-making process of the models, we integrate explainable Artificial Intelligence (AI) techniques, specifically using GNNExplainer, which identifies the most influential nodes and features contributing to each fraud prediction. This approach not only enables analysts and regulators to understand why a transaction is classified as suspicious, improving trust and accountability in automated cryptocurrency fraud detection systems, but also achieves a high detection F1-score of 96.22% for illicit transactions using GIN.

Keywords: Fraud detection, Graph isomorphism network, Graph neural network, Explainable artificial intelligence, Deep learning, Bitcoin.

1 | Introduction

Fraud detection involves identifying patterns in data that deviate from expected behaviors, commonly referred to as fraud, outliers, or anomalies. This task is particularly critical in cryptocurrency markets, where fraudulent activities can lead to substantial financial losses and undermine trust in decentralized systems. Detecting fraud in cryptocurrencies requires understanding blockchain, a tamper-proof distributed ledger that records transactions using cryptography and consensus algorithms [1]. While blockchain ensures transaction integrity, its pseudonymous and highly interconnected nature creates unique challenges for identifying suspicious transactions, making traditional detection methods less effective.

Corresponding Author: esfahaa@aut.ac.ir

https://doi.org/10.48314/jidcm.v2i1.89

To address these challenges, Graph Neural Networks (GNNs) have emerged as powerful tools, capable of modeling complex relationships in graph-structured data. Among GNN variants, Graph Isomorphism Networks (GINs) are particularly effective due to their high expressive power, enabling them to capture subtle structural patterns in cryptocurrency transaction networks that are indicative of illicit activity [2], [3]. GINs aggregate information from neighboring transactions while preserving node identity, making them suitable for fraud detection in highly connected and imbalanced graphs. By leveraging both node features and structural information, GINs can detect subtle anomalies that might be missed by traditional machine learning approaches.

A key challenge in using deep GNN models is understanding why a model flags a transaction as suspicious. To tackle this, our approach integrates Explainable Artificial Intelligence (XAI), specifically GNNExplainer, which identifies the most influential nodes and features that contribute to each prediction [4]. This capability allows analysts and regulators to trace back suspicious patterns in the network and understand the reasoning behind the model's decisions. By combining GINs with GNNExplainer, our method not only achieves superior detection performance but also provides actionable explanations for flagged transactions, enhancing transparency and trust in automated cryptocurrency fraud detection. This capability is particularly important for regulatory compliance, as it allows investigators to justify alerts and focus their efforts on the most critical transaction patterns.

2 | Background

Fraud detection in cryptocurrency networks is challenging due to the pseudonymous nature of transactions, high transaction volumes, and complex interactions among entities. Traditional anomaly detection and classical machine learning methods often struggle to capture the intricate relational patterns in blockchain data [5], [6]. Consequently, recent studies have increasingly employed graph-based approaches, particularly GNNs, to model structural and transactional dependencies more effectively.

In a study by Ashfaq et al. [7], machine learning models such as XGBoost and Random Forest (RF) were used to classify Bitcoin transactions as legitimate or illicit. Blockchain and smart contracts were also employed to analyze and categorize transactions. One of the main challenges in [7] was data imbalance, which was addressed using the SMOTE technique. Additionally, cybersecurity threats such as double-spending and Sybil attacks were highlighted as potential risks.

Asiri et al. [8] proposed a Graph Convolutional Network (GCN) to detect illicit Bitcoin transactions, combined with traditional machine learning models including Logistic Regression, RF, SVM, and LSTM. The main challenge in [8] was the severe data imbalance and the difficulty in accurately identifying illicit transactions, as they were significantly underrepresented compared to legitimate ones. To address this, the authors employed semi-supervised learning, leveraging both labeled and unlabeled data to improve detection performance. They also utilized graph structural features, such as degree and closeness centrality, to identify important nodes in the transaction network, enhancing model accuracy.

In a study by Simone Marasi and Ferretti [9], multiple GNN models, including GCN, GAT, ChebNet, and GraphSAGE, were applied to detect illicit Bitcoin transactions. The study addressed challenges such as data imbalance and scalability by employing techniques such as pseudo-labeling and data augmentation, along with graph centrality measures to identify key nodes. ChebNet and GraphSAGE outperformed traditional machine learning models in terms of accuracy and recall, demonstrating the effectiveness of GNNs for cryptocurrency fraud detection.

Building upon these studies, the present work aims to further improve fraud detection by focusing on data imbalance and model explainability. The focal loss method [10] is applied to handle class imbalance, giving higher weight to rare illicit transactions during the training phase of the network. GIN is applied to capture complex relational patterns, and GNNExplainer is integrated to enhance interpretability by identifying the

most influential nodes and features for each prediction. Combining the GIN with GNNExplainer improves both the accuracy and trustworthiness of our automated cryptocurrency fraud detection system.

3 | Materials and Method

This section presents the methodology used to detect illicit Bitcoin transactions through a combination of machine learning, GNNs, and interpretability tools. This section is structured into three main parts: Data preparation, graph modeling of GIN, and model interpretability.

3.1 | Data Preparation

This study utilizes the Elliptic dataset [9], a publicly available labeled graph dataset released by the cryptocurrency intelligence firm Elliptic, aimed at advancing Anti-Money Laundering (AML) research. The dataset consisted of 203,769 Bitcoin transactions (nodes) and 234,355 directed payment flows (edges), mapping the flow of Bitcoin between transactions. Each transaction is labeled as licit, illicit, or unknown depending on the entity that initiated it. Notably, only 2% of the transactions are labeled illicit, while 21% are labeled licit, leading to a significant class imbalance. Each transaction node contains 166 handcrafted features derived from both raw transaction metadata and one-hop aggregated graph statistics. The first 94 features represent local attributes (e.g., number of inputs/outputs, transaction fee, BTC volume), whereas the remaining 72 are Aggregated Features (AF) from neighboring nodes, including statistical measures such as mean, standard deviation, min/max, and correlation.

To prepare the data, missing values were removed, and features were normalized using z-score [11] to ensure consistency and improve model learning. Transactions are then represented as nodes in a directed graph, where the edges denote the BTC flow between transactions over time. Owing to the extreme class imbalance, standard loss functions tend to be biased toward the majority (licit) class. To address this, the focal loss method [10] was used during training. Unlike the cross-entropy loss, the focal loss dynamically scales the contribution of each sample to the loss based on how well it is classified, placing more emphasis on hard-to-classify (typically illicit) examples using *Eq. (1)*.

$$\mathrm{FL}_{(p_T)} = -\alpha_t(1-p_t)^{\gamma}\mathrm{Log}_{(p_T)}, \quad (1)$$

where p_t represents the predicted probability of the correct class label. The term α_t is a weighting factor introduced to balance the importance of different classes, which is particularly useful for imbalanced datasets. The parameter γ, known as the focusing parameter, adjusts the extent to which the loss function down-weights well-classified examples. The component $(1-p_t)$ acts as a modulating factor that reduces the loss contribution from examples that the model classifies with high confidence while increasing the loss from those it finds harder to classify.

In essence, focal loss helps the model focus more on the misclassified or uncertain examples rather than the easy ones. This characteristic is especially important in the context of the Elliptic dataset, where illicit transactions are rare but vital to detect accurately.

3.2 | Graph Isomorphism Network

To analyze both the behavior of transactions and the structural patterns of the transaction network, we employed three graph-based neural architectures: GCN, Graph Attention Network (GAT), and GIN.

The GCN [12] serves as a foundational baseline. In each convolutional layer, node features are updated by aggregating features from neighboring nodes using a normalized adjacency matrix, as shown in *Eq. (2)*.

$$H^{(l+1)} = \sigma\left(D^{-\frac{1}{2}}(A+I)D^{-\frac{1}{2}}H^{(l)}W^{(l)}\right), \quad (2)$$

where $A + I$ adds self-loops, D is the degree matrix of A, $H^{(l)}$ is the node feature matrix at layer l, $W^{(l)}$ is the learnable weight matrix, and σ is the ReLU activation. This operation enforces a smoothing effect by combining local neighborhood features into a more informative representation. In the context of illicit transaction detection, this allows the model to capture the correlations between suspicious transactions and their immediate neighbors.

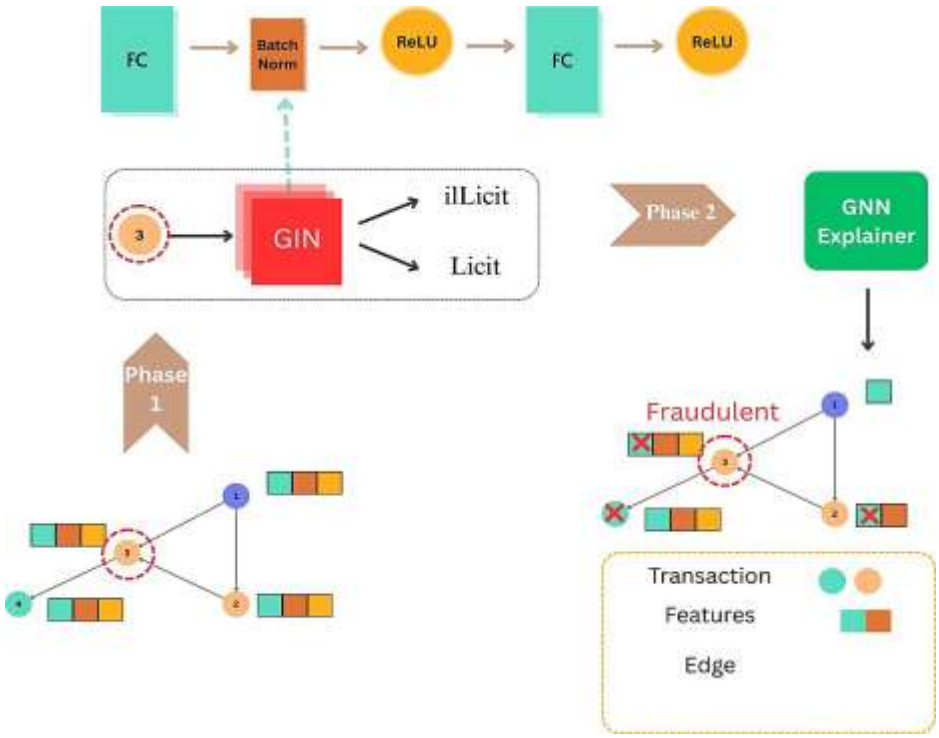

Fig. 1. Interpretable GIN framework for fraud detection.

In Phase 1, graph-oriented data are given to a GIN (FC stands for fully connected) for classification (Node 3 is chosen for the classification task). Phase 2, GNNExplainer is applied to retain the most influential features and neighbors that contribute to node 3's class prediction. GNNExplainer allows us to extract Interpretable subgraphs that can expose fraudulent patterns in cryptocurrency transactions.

The GAT [13] introduces an attention mechanism that assigns learnable, dynamic weights to neighbors during feature aggregation. For a given node i and its neighbor j, the attention coefficient is computed as shown in *Eq. (3)*.

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a^T[Wh_i||Wh_j]))}{\sum_{K\in N_i}\exp(\text{LeakyReLU}(a^T[Wh_i||Wh_j]))}, \quad (3)$$

where W is a shared linear transformation, a is the attention vector, and $||$ denotes concatenation. Multiple attention heads are used to stabilize learning and capture different aspects of neighborhood influence. In transaction graphs, this enables the model to assign higher weights to connections that are more indicative of fraudulent activity, even if the structural position is similar.

The GIN [2] is designed to match the discriminative power of the Weisfeiler–Lehman graph isomorphism test, making it particularly adept at distinguishing fine-grained structural differences. The update rule for node v at layer (k-1) is calculated by *Eq. (4)*.

$$h_v^{(k)} = \text{MLP}^{(k)}\left(\left(1+\epsilon^{(k)}\right)\cdot h_v^{(k-1)} + \sum_{u\in N} h_u^{(k-1)}\right). \quad (4)$$

Given GIN's theoretical strength, we selected it as our primary model and incorporated GNNExplainer [4] for interpretability. GNNExplainer generates local explanations by identifying the minimal subgraph and feature subset most influential for a particular prediction, allowing us to visualize why certain transactions are classified as illicit and which connections or attributes contribute most to the decision.

In GIN's equation, $\epsilon^{(1)}$ is either a learnable parameter or a fixed scalar, and MLP denotes a multi-layer perceptron. This formulation allows the model to learn how much to emphasize a node's own features relative to its neighbors, enabling robust detection of both local patterns (e.g., a single anomalous transaction) and global structures like complex laundering chains. The Elliptic dataset presents two key challenges: sparse label distribution, where illicit transactions form a small minority, and Heterogeneous structural patterns, where fraudulent activities may be hidden in both tightly connected clusters like scams and long transaction chains including money laundering.

GCN, with its normalized aggregation, captures broad local correlations but may oversmooth features in high-degree nodes, causing subtle anomalies to be lost. GAT mitigates this by weighting neighbors differently, improving sensitivity to influential edges, but its effectiveness can decline in highly noisy neighborhoods where attention scores are difficult to calibrate.

GIN, by contrast, is structurally more expressive and can distinguish between graphs that GCN and GAT would treat as equivalent. This expressiveness proved crucial for detecting illicit transactions that mimic legitimate patterns but differ in subtle neighborhood configurations. The combination of GIN with GNNExplainer not only provided higher predictive accuracy but also yielded interpretable subgraph-based justifications, making it a strong candidate for real-world deployment in financial forensics.

As shown in *Fig. 1*, in Phase 1, the graph data is given to the GIN model, which performs node classification, exemplified by the selection of node 3 for this task. In Phase 2, the GNNExplainer is used to identify and preserve the most important features and neighboring nodes that influence the classification outcome of node 3. The framework highlights a subgraph containing nodes and edges that are most relevant to the prediction, marking suspicious activity as fraudulent. This approach not only improves classification accuracy but also enhances interpretability by revealing specific patterns and features associated with fraudulent behavior, providing valuable insights for analysts.

3.3 | Interpretability

The detection of illicit transactions in cryptocurrency networks is not efficient without interpretability [14], especially in financial forensics, where investigative decisions require transparent justifications. In this work, we employ GNNExplainer as an interpretation tool to complement the predictive capabilities of the GIN. GNNExplainer [4] provides instance-level explanations for predictions made by a GNN. For any given node in the transaction graph, it identifies both the key node features from the 166 transaction attributes in the Elliptic dataset and the critical subgraph structure that most strongly influenced the classification decision. Examples of such influential features include transaction amount, time step, and fee-to-output ratio, while the critical subgraph corresponds to the minimal neighborhood whose presence is most responsible for the predicted class label.

The GIN method achieves this by learning two masks: a feature mask applied over the input features to select the most relevant dimensions, and an edge mask applied over the local subgraph to retain only the most informative connections. These masks are optimized to maximize the mutual information between the GNN's prediction on the full graph and its prediction on the masked explanatory subgraph as expressed by *Eq. (5)*.

$$\max_{G'.X'} I(Y; G'.X'), \quad (5)$$

where G' is the explanatory subgraph and X' is the subset of relevant features.

In our implementation, GNNExplainer operates on the node embeddings produced by GIN and outputs both ranked feature importance scores for each transaction and interpretable subgraphs that highlight influential neighbors. This combination not only reveals which specific attributes contributed to a classification, but also visualizes the structural context, for instance, uncovering that a transaction flagged as illicit may be directly connected to a high-risk cluster hub, embedded in a layering scheme, or part of a peel chain. Such explanations bridge the gap between accurate predictions and human interpretability, allowing illicit patterns in the Elliptic dataset to be traced and understood in a transparent manner, as shown in *Eq. (6)*.

$$[-\log P(GIN(G)|GIN(G') + \lambda \cdot |E_{G'}|]G' = \arg\min_{G' \subseteq G}. \quad (6)$$

Since G'⊆G is a descriptive subgraph we discussed earlier in this section, where |EG'| denotes the number of edges in G', and λ is a regularization parameter that balances the trade-off between simplicity of description and accuracy. Such subgraphs highlight suspicious clusters and distinct transactional behaviors (e.g., heavy transactions occurring in a short time frame), assisting forensic analysts in tracing fraudulent patterns.

This method not only renders the model's output interpretable, but also provides a foundation for engineering new features and developing hybrid strategies for fraud detection. Overall, the proposed combination of GIN and graph interpretability techniques offers a novel, accurate, and transparent framework for Bitcoin transaction fraud analysis. *Eq. (6)*, introduced here for the first time, defines the smallest, most informative subgraph that preserves the explanatory power of the original network while discarding extraneous connections, forming the mathematical backbone of our proposed interpretability-driven fraud detection framework.

4 | Results and Discussion

In this section, we present and discuss our experimental results on detecting illegal Bitcoin transactions, using GNNs applied to the Elliptic dataset. All experiments were conducted on the Elliptic dataset, which contains over 200k labeled and unlabeled Bitcoin transactions with 166 features per node, as described in Section 3.

Table 1. GIN's performance comparison with baselines.

Models	Illicit Transactions		
	Precision	**Recall**	**F1**
RF^{AF+NE}	95.60%	94.31%	94.72%
GCN	93.94%	94.34%	93.53%
GAT	93.38%	94.32%	94.11%
GIN_{XAI}	96.23%	96.39%	96.22%

As shown in *Table 1*, the GCN baseline achieved strong performance on licit transactions but struggled with illicit ones, obtaining 93.94% precision, 94.34% recall, and 93.53% F1-score. The GAT model, which incorporates attention mechanisms, slightly improved minority-class detection, reaching an F1-score of 94.11%. The GIN model delivered the best results, achieving 96.23% precision, 96.39% recall, and a 96.22% F1-score. This improvement is likely due to the GIN's stronger ability to capture subtle structural differences in transaction patterns. *Table 1* demonstrates the superiority of the GIN over other baseline models.

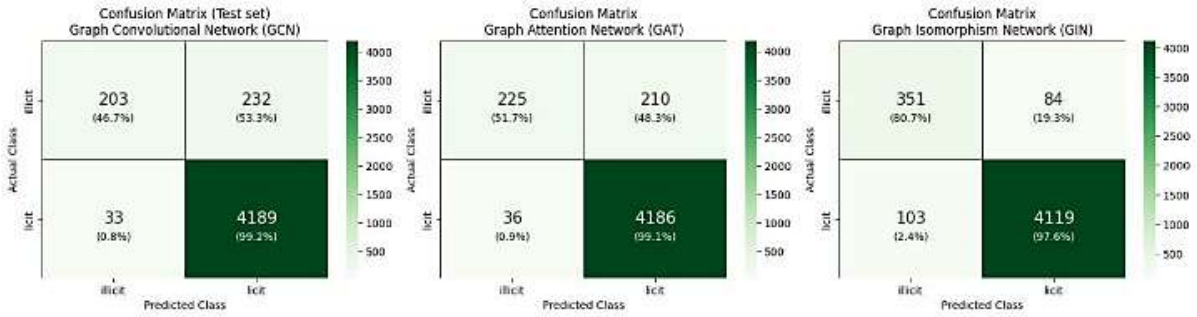

Fig. 2. Confusion matrices for GCN, GAT, and GIN on the test set.

Our t-test results show that GIN performs significantly better than all other models (GCN, GAT, and RF) for detecting fraudulent Bitcoin transactions. The p-values (all below 0.05) prove these differences are statistically important, not just random. The most notable improvement is over GCN with p=0.009, demonstrating that GIN's Weisfeiler-Lehman-inspired aggregation better captures structural fraud patterns than standard graph convolutions. While GAT's attention mechanism narrows the performance gap, the GIN maintains superiority with p=0.038, likely due to its stronger discriminative power on complex transaction graphs. RF shows competitive mean F1-scores but exhibits higher variance and statistically worse performance (p=0.025), highlighting the limitations of non-graph approaches in this domain.

Table 2. Statistical comparison (F1-score) of GIN vs. baselines using t-tests (α=0.05).

Model	Mean F1 ± Std	p-value (vs. GIN)	Significant? (α=0.05)
GIN (Proposed)	0.96 ± 0.22	—	—
GCN	0.93 ± 0.13	0.009	Yes (✓)
GAT	0.94 ± 0.28	0.038	Yes (✓)
RF	0.94 ± 0.04	0.025	Yes (✓)

The confusion matrices in *Fig. 2* show that the GIN significantly outperforms the GCN and GAT in detecting illicit transactions. While the GCN correctly identifies only 46.7% of illicit samples and GAT improves this to 51.7%, the GIN achieves 80.7%, marking a substantial gain in recall for the minority class. Although the GIN has a slightly higher false positive rate for licit transactions compared to the GCN and GAT, this trade-off leads to much better coverage of illicit activity, which is crucial for fraud detection.

5 | Conclusion

This study explored how GNNs can help to detect fraud in cryptocurrency transactions. The experiments showed that the GIN works better than other models because it can capture small but important patterns in the data. In our work, we trained the model on the Elliptic dataset, and the results were very promising. The GIN with GNNExplainer (GINXAI) reached 96.23% precision, 96.39% recall, and 96.22% F1-score, which shows that the model is strong in detecting suspicious activities. By using GNNExplainer, the system can also give reasons for its predictions, which makes it more transparent and useful for investigators and regulators. For future work, we will test the method on larger and more diverse transaction networks to see if it can scale well. We also plan to study the time dimension of fraud, because criminals often change their behavior over time. In addition, we want to include expert knowledge and rules from the financial sector, so the system can be more practical and easier to apply in real cases.

Authors' Contributions

All aspects of the research and manuscript preparation were carried out by the author. The author has read and approved the final version of the manuscript.

Funding

Not applicable.

Data Availability

All data supporting the reported findings in this research paper are provided within the manuscript.

Conflict of Interest

The author declares that they have no conflicts of interest.

Consent for Publication

The author confirms consent for the publication of this work.

Ethics Approval and Consent to Participate

This article does not contain any studies with human participants performed by the author.

References

[1] Li, X., Jiang, P., Chen, T., Luo, X., & Wen, Q. (2020). A survey on the security of blockchain systems. *Future generation computer systems*, *107*, 841–853. https://doi.org/10.1016/j.future.2017.08.020

[2] Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). *How powerful are graph neural networks?* https://doi.org/10.48550/arXiv.1810.00826

[3] Nasir, W. (2026). Cryptocurrency fraud and financial crime: Analyzing convicted fraudsters in digital asset scams. *Authorea*, *3*(12). https://doi.org/10.22541/au.174180456.65670202/v1

[4] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Žitnik, & Jure Leskovec. (2019). GNNExplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems* (PP. 9240–9251). Curran Associates, Inc. https://pmc.ncbi.nlm.nih.gov/articles/PMC7138248/

[5] Liu, S., Cui, B., & Hou, W. (2024). A survey on blockchain abnormal transaction detection. *Blockchain and trustworthy systems* (pp. 211–225). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-8101-4_15

[6] Bandarupalli, G. (2025). The evolution of blockchain security and examining machine learning's impact on ethereum fraud detection. *2025 17th international conference on electronics, computers and artificial intelligence (ECAI)* (pp. 1–6). IEEE. https://doi.org/10.1109/ECAI65401.2025.11095594

[7] Ashfaq, T., Khalid, R., Yahaya, A. S., Aslam, S., Azar, A. T., Alsafari, S., & Hameed, I. A. (2022). A machine learning and blockchain based efficient fraud detection mechanism. *Sensors*, *22*(19), 7162. https://doi.org/10.3390/s22197162

[8] Asiri, A., & Somasundaram, K. (2025). Graph convolution network for fraud detection in bitcoin transactions. *Scientific reports*, *15*(1), 11076. https://doi.org/10.1038/s41598-025-95672-w

[9] Marasi, S., & Ferretti, S. (2024). Anti-money laundering in cryptocurrencies through graph neural networks: A comparative study. *2024 IEEE 21st consumer communications & networking conference (CCNC)* (pp. 272–277). IEEE. https://doi.org/10.1109/CCNC51664.2024.10454631

[10] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988). IEEE. https://doi.org/10.1109/ICCV.2017.324

[11] Chen, Y., Tang, X., Qi, X., Li, C. G., & Xiao, R. (2022). Learning graph normalization for graph neural networks. *Neurocomputing*, *493*, 613–625. https://doi.org/10.1016/j.neucom.2022.01.003

[12] Jiang, B., Zhang, Z., Lin, D., Tang, J., & Luo, B. (2019). Semi-supervised learning with graph learning-convolutional networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11313-11320). IEEE. https://doi.org/10.1109/CVPR.2019.01157

[13] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). *Graph attention networks*. https://info.arxiv.org/help/api/index

[14] Li, K., Yang, T., Zhou, M., Meng, J., Wang, S., Wu, Y., …, & Tong, Y. (2024). SEFraud: Graph-based self-explainable fraud detection via interpretative mask learning. *Proceedings of the 30th ACM sigkdd conference on knowledge discovery and data mining* (pp. 5329–5338). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3637528.3671534