electronics

Article

An Abnormal Account Identification Method by Topology Feature Analysis for Blockchain-Based Transaction Network

Yuyu Yue [1], Jixin Zhang [1,*], Mingwu Zhang [1,2] and Jia Yang [1]

[1] School of Computer Science, Hubei University of Technology, Wuhan 430068, China; 102101029@hbut.edu.cn (Y.Y.); mzhang@mail.hbut.edu.cn (M.Z.); jia_yang@hbut.edu.cn (J.Y.)
[2] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China
* Correspondence: zhangjx@hbut.edu.cn

Abstract: Cryptocurrency, as one of the most successful applications of blockchain technology, has played a vital role in promoting the development of the digital economy. However, its anonymity, large scale of cryptographic transactions, and decentralization have also brought new challenges in identifying abnormal accounts and preventing abnormal transaction behaviors, such as money laundering, extortion, and market manipulation. Recently, some researchers have proposed efficient and accurate abnormal transaction detection based on machine learning. However, in reality, abnormal accounts and transactions are far less common than normal accounts and transactions, so it is difficult for the previous methods to detect abnormal accounts by training with such an imbalance in abnormal/normal accounts. To address the issues, in this paper, we propose a method for identifying abnormal accounts using topology analysis of cryptographic transactions. We consider the accounts and transactions in the blockchain as graph nodes and edges. Since the abnormal accounts may have special topology features, we extract topology features from the transaction graph. By analyzing the topology features of transactions, we discover that the high-dimensional sparse topology features can be compressed by using the singular value decomposition method for feature dimension reduction. Subsequently, we use the generative adversarial network to generate samples like abnormal accounts, which will be sent to the training dataset to produce an equilibrium of abnormal/normal accounts. Finally, we utilize several machine learning techniques to detect abnormal accounts in the blockchain. Our experimental results demonstrate that our method significantly improves the accuracy and recall rate for detecting abnormal accounts in blockchain compared with the state-of-the-art methods.

Keywords: cryptocurrency; transaction graph; singular value decomposition; generative adversarial networks

Citation: Yue, Y.; Zhang, J.; Zhang, M.; Yang, J. An Abnormal Account Identification Method by Topology Feature Analysis for Blockchain-Based Transaction Network. *Electronics* **2024**, *13*, 1416. https://doi.org/10.3390/electronics13081416

Academic Editor: Myung-Sup Kim

Received: 8 March 2024
Revised: 1 April 2024
Accepted: 4 April 2024
Published: 9 April 2024

1. Introduction

Cryptocurrency [1] is a type of distributed ledger application based on blockchain technology [2], characterized by decentralization, anonymity, and tamper-resistance. Since Bitcoin was first introduced in 2009, the cryptocurrency market has rapidly developed. As a new type of digital asset, cryptocurrency has improved transaction efficiency, reduced transaction costs, and attracted the attention and participation of industry and academia, promoting the development of financial inclusion.

Meanwhile, due to the decentralized and anonymous nature of blockchain, cryptocurrency has also led to illegal activities such as money laundering [3], extortion [4], gambling [5], drug trafficking [6], illegal transactions [7], and cybercrime [8]. There exist significant security problems and challenges in applying blockchain-based cryptocurrency. For example, in 2021, the British police arrested multiple suspects of cryptocurrency money laundering and seized over USD 250 million. In 2022, two people were arrested for alleged conspiracy to launder USD 4.5 billion in stolen cryptocurrency. Over the past decade, global cryptocurrency investors have suffered losses exceeding USD 30 billion. In 2022

alone, a total of 436 individual cryptocurrency-related crimes were reported, surpassing the previous year's count by over four times. According to Chainalysis, the estimated lower limit of sum of illegal cryptocurrency transactions has reached an unprecedented high of USD 20.6 billion [9].

Recent works propose machine-learning-based approaches for detecting abnormal transactions in the blockchain-based transaction network to identify potential fraud and cybercrime activities in cryptocurrency. P. Monamo et al. [10] investigated trimmed k-means clustering for fraud detection. Gu et al. [11] detected abnormal transaction amounts in cryptocurrency exchanges, which can be indicative of various abnormal or abnormal behaviors. Their approach identifies the most important factors and uses a prediction model based on deep learning to detect abnormal transaction amounts. Wu et al. [12] proposed a method for analyzing the behavior features of blockchain transactions and identifying suspected addresses through 19 features. Furthermore, Lorenz et al. [13] addressed the challenge of detecting money laundering by using machine learning. They proposed an active learning solution that achieves comparable performance. Mohy et al. [14] built the NIDS using the KNN algorithm to improve the IDS accuracy and detection rate. Furthermore, principal component analysis, the univariate statistical test, and a genetic algorithm were used for feature selection separately to improve the data quality and select the ten best performing features.

However, some blockchains encrypt the data (such as "money") in transactions to protect privacy, so some features extracted from the data may not be useful in abnormal behavior detection. Moreover, the above approaches use machine learning to train the abnormal behavior detection model with some datasets, but they do not consider that the normal and abnormal transactions are imbalanced in reality. As a result, traditional machine learning methods may not be effective as they are not equipped to handle such imbalanced datasets. In such situations, training a model on such an imbalanced dataset where only a small proportion of accounts and transactions are identified as abnormal may result in poor performance.

To address these challenges, we propose an abnormal account identification method for blockchain-based transaction networks. Since the key data may be encrypted for privacy protection, we first transform the accounts and transactions in the blockchain into graph nodes and edges and analyze the topology feature to represent the abnormal behaviors in transactions. We then use an adjacency matrix to represent the transaction topology. However, the adjacency matrix is sparse and not adapted to machine learning training, and its dimension may be very large, which leads to significant computation cost. So, we next use the singular value decomposition (SVD) method to extract topology features from the adjacent matrix to reduce the feature dimension and solve the sparse feature problem.

As the abnormal activities are much less common than the normal activities, the abnormal activities in transactions can be easily hidden among a large number of legitimate transactions, and such imbalanced training data make it difficult to identify abnormal transactions and distinguish them from normal ones, which in turn affects the accuracy and reliability of the abnormal behavior detection model. Distinct from previous works, our approach uses a generative adversarial network (GAN) to generate similar samples with the same distribution of the topology features for training data augmentation. The GAN-based training data augmentation method makes the training balanced. The main contributions of this paper are summarized as follows:

- We propose an abnormal account identification method for blockchain-based transaction networks by using topology feature analysis. Our approach can be used for some blockchain-based cryptocurrency that encrypts the data for privacy protection.
- We propose a SVD-GAN-based feature reduction and augmentation method to solve the sparse problem and imbalanced training data problem in topology features of transactions.

- We use real-life datasets in our experiments to demonstrate that our approach significantly improves the performance compared with some machine (deep) learning methods and state-of-the-art methods.

The remainder of this paper is organized as follows. Section 2 presents the related works. Section 3 introduces the methodology of this paper. Section 4 shows experimental results, and Section 5 shows the conclusion.

2. Related Works

In this section, we review the existing literature related to the generation of user topology features and their application in anomaly detection in blockchain networks. We begin by discussing the studies that have examined similar research questions or hypotheses. Next, we describe the studies that have used similar methods or techniques for feature generation and anomaly detection in blockchain networks. Finally, we highlight the key findings from the literature and identify any gaps or limitations that our study aims to address. In recent years, there has been an increasing amount of research focused on detecting abnormal blockchain accounts, and this area of research is considered to be of paramount importance in both industry and academia as it can help identify unusual activities on the blockchain.

2.1. Literature Review

To detect abnormal transactions in the blockchain, some machine learning methods have been employed, like active machine learning solutions (Lorenz et al. [13]), a collaborative clustering-characteristic-based data fusion approach (Liang et al. [15]), a secure fraud detection model based on XGboost and the random forest method (Ashfaq et al. [16]), and using a semi-supervised generative adversarial network, which efficiently detects abnormal attacks within the Ethereum network (Sanjalawe et al. [17]). In another method, the system integrates blockchain at base stations and cluster heads in a wireless sensor network to enhance security using a machine learning classifier called Histogram Gradient Boost to identify and revoke malicious nodes (Nouman et al. [18]).

Scattered graph neural-network-based detection is a novel method to capture abnormal behavior in the Bitcoin blockchain. Examples include the graph convolutional network and linear layers method (Alarab et al. [19]), the temporal graph convolutional network solution (Sharma et al. [20]), the graph contrastive learning method (Chen et al. [21]), and the synthesis between the existing literature on the application of graph-based anomaly approaches in fraud detection published between 2007 and 2018 (Pourhabibi et al. [22]). In a abnormal transaction cryptocurrency detection method with spatiotemporal and global representation, the CTDM combines EvolveGCN with MGU and global representations to achieve better performance (Xiao et al. [23]). An extensive survey of the blockchain anomaly transaction detection was conducted, and graph convolutional networks were applied to the domain of blockchain anomaly detection (Liu et al. [24]).

In addition, a few supervised methods, similar to comparative analysis of the performance of classical supervised learning methods (Alarab et al. [25]), aimed to detect illicit transactions in the Bitcoin network using the K-nearest neighbors (KNN) algorithm (Elbaghdadi et al. [26]) and explore the local topology and geometry of the Bitcoin network method (Nerurkar et al. [27]). The system architecture for detecting fraudulent transactions and attacks in the BC network is based on machine learning. Machine learning was used to check medical data from sensors and block abnormal data from entering the blockchain network (Mohammed et al. [28]. Data structures known as sketches, specifically bloom filters and hyperLogLog, have been utilized to identify suspicious accounts without requiring the examination of the entire blockchain data, and methods have been developed to identify accounts with high transaction volume, frequency, and node degree (Voronov et al. [29]). Another study highlights how an interplay between blockchain and ML would allow both technologies to assist 130 cybersecurity-related use cases (Venkatesan et al. [30]).

Furthermore, certain rules-based detection, in the style of constructing a temporal Bitcoin network combining time constraints, attribute constraints, and structure constraints, along with the multi-constrained meta-path, can be considered as a means of detecting abnormal user behavior. All state-of-the-art studies provide insights into the effectiveness of different methods and techniques for topology feature generation and anomaly detection in blockchain networks.

2.2. Challenges and Proposed Method

Table 1 summarizes and compares the main issues addressed in the existing literature. and we can find that one of the main limitations of the existing literature is that most of the studies focus only on specific cryptocurrencies or blockchain networks, which largely limits the broad applicability and generalization of the findings. In addition, there is a relative lack of research on the impact of different data preprocessing and feature selection techniques on the performance of anomaly detection based on user graph features. Meanwhile, the existing literature is also relatively scarce in terms of comparative studies of different feature dimension reduction and data augmentation methods and their technical effectiveness in anomaly detection based on high-dimensional and sparse graph features in the blockchain.

Table 1. The comparison of the literature review. (✓indicates the problem addressed by the literature).

Literature	Machine Learning	Graph Neural Network	Data Imbalance	High-Dimensional Features	Blockchain Abnormal Behavior Detection
Lorenz et al. [13]	✓				✓
Mohy et al. [14]	✓			✓	✓
Liang et al. [15]	✓				✓
Ashfaq et al. [16]	✓		✓		✓
Sanjalawe et al. [17]	✓		✓		✓
Muhammad et al. [18]	✓				✓
Alarab et al. [19]		✓		✓	✓
Sharma et al. [20]	✓	✓			✓
Chen et al. [21]		✓		✓	✓
Pourhabibi et al. [22]					✓
Xiao et al. [23]		✓		✓	✓
Liu et al. [24]		✓		✓	✓
Alarab et al. [25]	✓				✓
Elbaghdadi et al. [26]	✓				✓
Nerurkar et al. [27]		✓			✓
Mohammed et al. [28]	✓				✓
Voronov et al. [29]				✓	✓
Venkatesan et al. [30]	✓				✓

Given this, this study aims to evaluate the efficacy of these methods by constructing a comprehensive framework dedicated to generating, downgrading, and data-enhancing user graph features in blockchain networks evaluating the efficacy of these methods on multiple types of cryptocurrencies and blockchain networks. The core goal of this paper is to fill the gaps in the existing literature on abnormal behavior detection in the blockchain and to promote further development in the field. With this comprehensive approach, this paper expects to improve the generalizability and generalization of the anomaly detection models and to provide more effective solutions for blockchain network security. This research is dedicated to more effective identification methods of abnormal transaction behavior in the blockchain, and through summary analysis, there are two urgent problems in this regard in small sample scenarios: the first aspect is that the abnormal transaction mapping features in the blockchain are high-dimensional and sparse; the second aspect is that it is difficult to accurately detect the abnormal transaction behavior in the blockchain, especially in the case where the user transaction data are imbalanced.

Our study focuses on the challenges of small samples and high-dimensional sparse features in detecting abnormal transactions in the blockchain. To address these issues, we propose a novel method for generating informative topology features that can be used as input to machine learning models.

Generating an informative transaction graph feature: Our approach involves constructing user transaction graphs using historical transaction data, reducing the feature dimensionality to alleviate sparsity, and employing data augmentation techniques to tackle the problem of a limited number of abnormal topology features in blockchain transactions. After reducing the dimensionality of user topology features, the abnormal topology features are subjected to data augmentation methods to generate the final user topology features. The application of generative adversarial networks in generating minority abnormal topology features exhibits significant advantages in capturing the underlying features of genuine topology features and synthesizing new abnormal topology features that conform with the real transaction data.

Application of generative adversarial networks (GANs) for data augmentation: Abnormal transaction detection tasks are faced with the challenge that abnormal transaction samples usually constitute a negligible fraction of the entire transaction dataset, thereby constraining traditional machine learning algorithms in their capability to learn features from limited abnormal samples and fabricate new abnormal transaction data. A GAN, on the other hand, can generate fresh abnormal transaction data from a small set of abnormal transaction samples, augmenting the diversity and quantity of abnormal transaction data and consequently improving the accuracy and dependability of abnormal account detection. Furthermore, a GAN can bolster the robustness and generalization capacity of the model via adversarial training, augmenting the identification ability of unexplored abnormal transaction samples and enhancing the efficacy of abnormal account detection. Subsequently, we utilize machine learning models to predict and identify abnormal accounts in the blockchain.

Utilizing machine learning models for abnormal behavior detection: Our proposed approach presents an efficient solution to augment the detection of abnormal accounts in blockchain transactions through the synthesis of informative topology features for machine learning models. The potential impact of our study is substantial, as it could facilitate the advancement of more secure and reliable blockchain systems, leading to greater adoption and integration of blockchain technology across diverse industries.

It would be beneficial to conduct further research to optimize the performance of our proposed method and explore its applicability in other domains. In addition, our findings suggest that using advanced machine learning techniques in combination with blockchain technology can enhance the security and trustworthiness of transaction systems, thereby creating new opportunities for innovation and growth in the digital economy [31].

2.3. Potential Impact and Future Research

One limitation of the existing literature is that most studies focus on specific types of cryptocurrencies [32] or blockchain networks, which limits the generalizability of their findings. Additionally, there is a lack of research on the impact of different data preprocessing and feature selection techniques on the performance of user topology feature-based abnormal behavior detection [33]. Furthermore, the literature lacks studies that compare the effectiveness of different methods and techniques for topology feature reduction and data augmentation [34] in user topology feature-based abnormal behavior detection. Our study aims to address these gaps by developing a comprehensive framework for user topology feature generation, dimensionality reduction, and data augmentation for abnormal behavior detection in blockchain networks and evaluating its effectiveness on multiple types of cryptocurrencies and blockchain networks.

3. Methodology

3.1. Overview of Our Method

Our approach proposes to identify abnormal accounts by the topology features of the cryptocurrency transaction graph. Figure 1 shows the architecture of our method, which consists of four stages:

- Representation of the topology features, which uses the adjacent matrix to represent the topology feature of the cryptocurrency transaction graph.
- Feature dimension reduction of the topology features. Since the topology feature of the transaction graph is high-dimensional and sparse, it is not suitable for machines, so we use SVD to reduce the feature dimension while encoding the sparse feature vectors.
- Data augmentation for the topology features of abnormal accounts. Since in reality the abnormal accounts and transactions are far less common than normal accounts and transactions, we use the generative adversarial network to generate samples, which has the same feature distribution as the abnormal accounts. These samples are sent to the training dataset to produce an equilibrium of abnormal/normal accounts.
- Machine-learning-based abnormal account identification. After feature reduction and data augmentation, we use machine learning methods for training the abnormal account identification model.

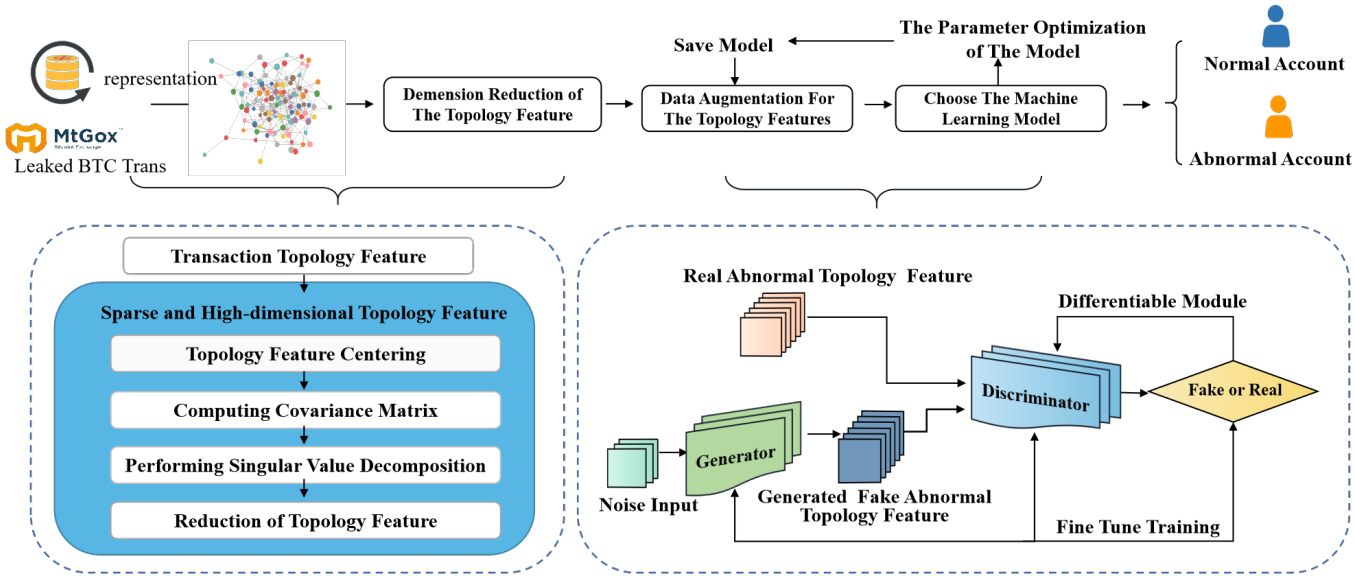

Figure 1. The overall architecture of the abnormal account identification method.

3.2. Representation of Topology Features

We first represent the behaviors of the accounts in the blockchain transactions by building an undirected transaction hyper-graph, in which the graph nodes represent the accounts and the edges represent the transactions among them (there may be more than one transaction between two nodes), as shown in Figure 2, where Figure 2a shows the behaviors of normal transactions and Figure 2b shows the behaviors of abnormal transactions. Intuitively, the abnormal transaction behaviors seem more aggregated, which is a realistic basis for the use of our transaction graph to separate normal and abnormal accounts. Let $G = (V, E)$ be the transaction graph, transforming the transaction between different accounts into a connection relationship between nodes. $V = \{v_i\}$ is the set of nodes in the transaction graph G, and $E = \{e_i\}$ is the set of edges in the transaction graph G. Let $A_{n\times n} = [v_i, v_j]$ be the adjacency matrix of the transaction graph G, in which n is the number of the account nodes, v_i is the row index, v_j is the column index, and the element $a_{i,j}$ represents the number of connections between two accounts v_i and v_j, which shows the frequency of the interactions between them. We represent the topology feature for each node in the graph by using the row vectors in the adjacency matrix.

(**a**) The graph of normal behavior. (**b**) The graph of abnormal behavior.

Figure 2. Comparison of normal and abnormal transaction graphs.

3.3. Feature Dimension Reduction of the Topology Features

Since the topology feature is sparse and its dimension is very large, which leads to significant computation cost and convergence time, and the length of the topology feature is unfixed, we use the singular value decomposition (SVD) method to reduce the dimension of the topology feature to a small fixed value without losing much topology feature information. The SVD method can preserve the original information and remove redundant information about topology features after dimension reduction. The dimension reduced topology feature is nonsparse. Moreover, it can also reduce computational complexity and convergence speed by projecting topology features into a low-dimensional subspace.

We compute the SVD of the adjacency matrix $A_{n\times n}$ of the transaction graph $G = (V, E)$ according to Equation (1), where U is the left singular vector matrix of dimension $n \times n$, and Σ is the diagonal matrix of singular values, with dimensions $n \times n$. V^T is the right singular vector matrix of dimension $n \times n$. For the adjacency matrix $A_{n\times n}$, where the row number n represents the number of nodes, the column number n represents the number of features.

$$A = U\Sigma V^T \tag{1}$$

Next, we select the top k singular values and their corresponding columns U_k $(n \times k)$ and V_k^T $(k \times n)$. We compute the dimension reduced topology feature matrix Z_k according to Equation (2).

$$Z_k = U_k \Sigma_k V_k^T \tag{2}$$

3.4. Data Augmentation for the Topology Features of Abnormal Account

We used the dimension-reduced topology feature of normal/abnormal accounts as the training data to train an abnormal behavior detection model by machine learning. However, since in reality the abnormal accounts and activities are much fewer than the

normal accounts and activities, such imbalanced training data make it difficult to train an abnormal behavior detection model to efficiently identify abnormal transactions and distinguish them from normal ones, so in this method, we use generative adversarial networks (GANs) to generate samples with the similar distributions of the topology features of abnormal accounts. Both the generated data and the original data are sent to the machine learning model as the balance training data for training the abnormal account detection model. Figure 3 shows the architecture of our GAN network. Our GAN network consists of a generator network $G()$ and a discriminator network $D()$. The two networks are trained alternately in an architecture and learn together to progress, finally reaching a Nash equilibrium state.

The generator network G comprises four fully connected layers followed by LeakyReLU activation functions and a final sigmoid activation function. The initial input is a vector of k-dimensional random values, which is then transformed through the sequential layers to produce generated topology features. The output is a vector representing the generated topology features.

The discriminator network D consists of four linear layers with LeakyReLU activation functions and dropout layers to enhance the network's generalization ability. The dimension-reduced topology features and the generated topology features as inputs are jointly fed into the discriminator network for training to distinguish between real topology features and the generated ones. The output of the discriminator is a scalar value representing the probability of the input data being real (having a value close to 1) or generated (having a value close to 0).

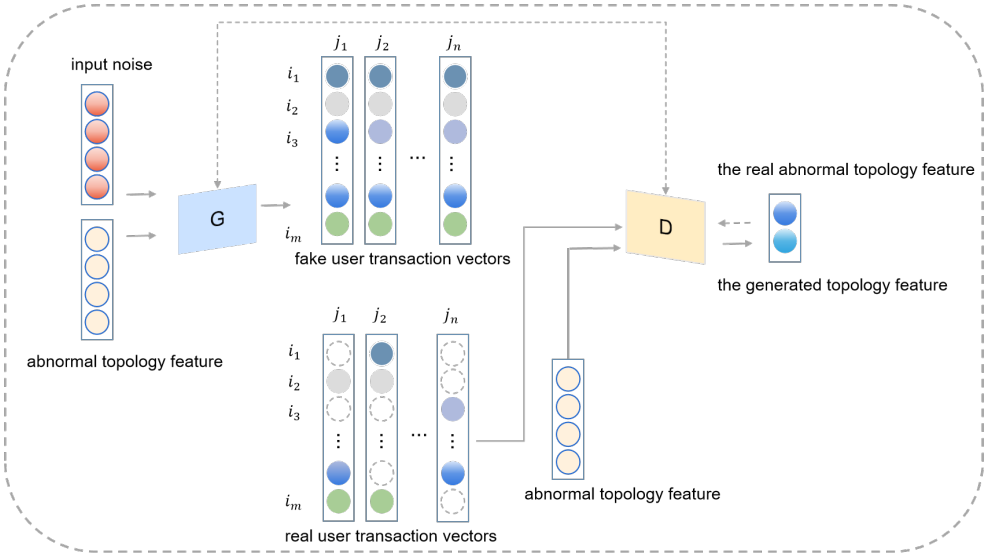

Figure 3. The transaction graph of abnormal behavior.

In the training process, we utilize the binary cross entropy (BCE) loss function for both the discriminator and generator networks, according to Equation (3), where Z_k is the k-dimensional reduced topology features, z indicates a k-dimensional random vector as the input of generator network, $G(z)$ is the output of the generator network when given the random vector z as input, $V(D, G)$ is the gap function between the real topology features and generated topology features, $D(Z_k)$ is the output of the discriminator network when given the real topology features z as input, $D(G(z))$ is the output of the discriminator when given the generated topology features as input, $P_{data(Z_k)}$ is the distribution of the k-dimensional reduced topology features, and $P_{z(Z)}$ is the distribution of the random vectors.

We separate Adam optimizers to update the parameters of the generator and discriminator independently. To train the generator network, we generate a random vector (drawn from a Gaussian distribution) as input to the generator network. The random noise vector is transformed through the generator network to produce the generated topology features.

Then, the discriminator takes both real topology features and generated topology features as input and performs classification with the following objective function. After repeatedly alternating between training the generator and the discriminator, the generator's adversarial loss gradually decreases, and the discriminator's adversarial loss stabilizes. This indicates that the GAN network has reached a certain level of convergence. The adversarial training process allows the generator to enhance its ability to generate more realistic topology features, while the discriminator becomes more adept at distinguishing between real and generated topology features. If the discriminator network cannot identify the generated topology features from the real topology features, that means the generated topology features have a similar distribution and can be used for data augmentation of the topology features of abnormal accounts.

$$\min_G \max_D V(D,G) = \mathbb{E}_{Z_k \sim p_{data}(Z_k)}[\log D(Z_k)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \tag{3}$$

To validate the similarity of the topology generated by the GAN network to abnormal accounts, the GAN reaches a level of convergence where the topology feature generator's adversarial loss is decreasing and the topology feature discriminator's adversarial loss is stabilizing. This suggests that the generator is producing increasingly realistic topology features, and the discriminator is improving at distinguishing between real and generated features. Then, the discriminator's performance is evaluated by feeding it a mix of real abnormal account topology features and the generated ones. If the discriminator struggles to differentiate between the two, it indicates that the generated features closely resemble the real ones.

3.5. Machine-Learning-Based Abnormal Account Identification

Once we reduce the topology feature dimension and prepare the balance training dataset, we then use several (deep) machine learning methods such as support vector machine (SVM), logistic regression (LR), decision trees (DTs), random forest (RF), and XGboost to train an abnormal account detection model.

Given the labeled normal/abnormal transaction accounts with topology features as the training dataset, we obtain an abnormal account detection model through training with the topology features $\{z_i | i = 1, 2, \ldots, N\}$ and the labels y_i, where $y_i = -1$ or 1 in the SVM algorithm and $y_i = 0$ or 1 in the other machine learning algorithms. In the process of the model training, n-fold cross-validation is carried out to improve the generalization ability of the detection model. Here, we take the SVM and XGboost methods as examples to illustrate our approach.

3.5.1. Abnormal Behavior Detection Model Training by SVM Algorithm

The SVM model aims to find the hyperplane that optimally separates the normal/abnormal topology feature points by using the Radial Basis Function (RBF) kernel, according to Equation (4), where σ is the scale parameter of the RBF kernel, and $\|z_i - z_j\|$ is the Euclidean distance between the two topology feature points z_i and z_j.

$$K(z_i, z_j) = \exp\left(-\frac{\|z_i - z_j\|^2}{2\sigma^2}\right) \tag{4}$$

The objective of the SVM method is according to Equation (5), where w is the weight vector that defines the orientation of the decision boundary (hyperplane), b is the bias term that shifts the decision boundary, C is a hyperparameter that controls the trade-off between maximizing the margin and minimizing the account classification error, $\frac{1}{2}\|w\|^2$ encourages the maximization of the margin between the normal/abnormal classes, and $C\sum_{i=1}^{N} \xi_i$ represents the classification error penalty to minimize the sum of slack variables ξ_i.

Furthermore, next, we apply the optimization process to determine the optimal hyperplane to effectively identify the abnormal account.

$$\begin{aligned} \text{Maximize:} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{Subject to:} \quad & y_i(w^T z_i + b) \geq 1 - \xi_i \quad \text{for all } i = 1, 2, \ldots, N \\ & \xi_i \geq 0 \quad \text{for all } i = 1, 2, \ldots, N \end{aligned} \tag{5}$$

3.5.2. Abnormal Behavior Detection Model Training by XGboost Algorithm

XGboost leverages gradient-boosted trees within a framework of supervised learning. We utilize the labeled topology features z_i encompassing multiple attributes to forecast account labels y_i as either normal or abnormal, with y_i belonging to the set $\{0, 1\}$ for all i in the range from 1 to N.

The objective function [35] is initially defined to encapsulate the performance of the abnormal transaction account detection model, and it bears emphasizing that the choice of parameters can lead to diverse outcomes of our abnormal behavior detection model. The objective function is composed of two primary components, as shown in Equation (6): The first component is the training loss, which measures the predictive performance of our abnormal behavior detection model. The second component is the regularization term, which penalizes the size of the model parameters, thus helping to prevent over-fitting. By adjusting the L2 regularization coefficient value of λ, we can balance the complexity and generalization ability of the detection model. XGboost iteratively constructs a series of trees from the training topology features and sums their predictions, refining the final abnormal transaction accounts through an optimization process.

$$\text{obj}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\log(1 + e^{-y_i\theta^T z_i}) + \frac{\lambda}{2}\sum_{j=1}^{J}\theta_j^2 \tag{6}$$

4. Experiments

4.1. Setup

In this work, we conducted experiments to compare the performance of our approach and the state-of-the-art methods. All of the methods were implemented with the same environment, platform, and configurations. The CPU we used in the experiments is an Intel i5, the capacity of RAM was 8G, the capacity of the SSD was 500G, the operating system was Windows 10, and the development platform was PyCharm.

4.2. Datasets

We used the Mt.Gox [36] dataset to test all of the comparison methods, since Mt.Gox's online trading history was leaked, with Mt.Gox data dumps providing approximately 18 million matching buy and sell transactions from April 2011 to November 2013. The data include several items: TradeId represents a complete transaction from the buyer to the seller. The source TradeId represents the buyer, the target TradeId represents the seller, and the binary label indicates whether the user of the transaction is an abnormal transaction.

4.3. Data Pre-Processing

Due to a large number of duplicate entries in the transaction data, we adopted a data-cleaning method similar to most studies [36–39]. Specifically, we used a primary index combined with the SourceId and TargetId to delete duplicate entries. By doing so, the data were reduced from 1.05 million rows to 460,000 rows. In the training dataset, there are 16,728 normal accounts with normal trading behavior and 1357 negative accounts with abnormal trading behavior. Next, we used SourceId and TargetId as the row index and column index in the adjacent matrix to extract the topology features.

4.4. Performance Comparison of State-of-the-Art Imbalanced and Sparse Data Processing Methods

To show the performance of our approach, we compare the performance of our approach with other imbalanced data processing methods. Table 2 shows the performance of the six machine learning classifiers (SVM, LR, DT, XGboost, KNN, Ada) with the imbalanced data processed by various imbalanced data processing methods which include our approach (SVD-based dimension reduction and GAN-based data augmentation) against the other two approaches, specifically undersampling as proposed in [40] and oversampling as proposed in [41]. Analyzing the data from the table, it becomes evident that:

- Our method achieves a significant improvement in precision and recall compared with the other imbalanced data processing methods across various machine learning methods.
- If we do not employ any imbalanced and sparse data processing methods, the performance will be worst of all of the machine learning methods.
- Since the original topology feature vectors are very sparse, the compared two imbalanced data processing methods use SVD to reduce the feature dimension. The experimental results show that SVD-based feature dimension reduction is necessary.
- Since the numbers of abnormal accounts and legitimate ones are imbalanced, the GAN-based data augmentation methods improve the precision and recall, as well as the two imbalanced data processing methods [40,41].
- When SVD-based feature dimension reduction is applied to the imbalanced data processing methods [40,41], the overall performance increases compared to the method without data augmentation.
- In our approach, we compare the SVD_GAN and GAN_SVD methods; the SVD_GAN method first takes the SVD-based feature dimension reduction to process the sparse feature and then takes the GAN-based imbalanced data augmentation, while the GAN_SVD method first takes GAN-based imbalanced data augmentation and then takes the SVD-based feature dimension reduction to process the sparse feature. The performance results show that the SVD_GAN method achieves better performance than the GAN_SVD method.

Table 2. Experimental results of the performance comparison of different imbalanced data processing methods.

Methods	SVM		LR		DT	
	Precision	Recall	Precision	Recall	Precision	Recall
None	69.6%	73.3%	67.9%	39.9%	51.4%	46.1%
SVD	70.2%	75.4%	73.6%	62.0%	64.6%	65.5%
GAN	79.9%	72.8%	83.8%	71.6%	86.3%	68.7%
US [40]	65.1%	78.7%	33.3%	75.5%	26.2%	76.9%
OS [41]	68.9%	72.1%	45.1%	56.3%	53.2%	47.8%
SVD_US	71.4%	69.3%	66.5%	81.6%	54.8%	64.9%
SVD_OS	68.7%	76.7%	70.5%	74.6%	58.2%	64.8%
Our (GAN_SVD)	80.8%	73.4%	84.0%	71.7%	64.0%	63.9%
Our (SVD_GAN)	91.0%	57.7%	83.7%	71.1%	68.9%	66.4%
Methods	**KNN**		**Ada**		**XGboost**	
	Precision	Recall	Precision	Recall	Precision	Recall
None	80.4%	38.2%	71.5%	39.7%	77.9%	41.6%
SVD	67.1%	57.8%	80.9%	67.7%	89.6%	66.4%
GAN	93.2%	56.0%	89.5%	70.2%	85.5%	70.8%
US [40]	59.6%	36.6%	29.8%	71.0%	31.4%	74.9%
OS [41]	20.1%	73.4%	63.1%	45.4%	79.3%	45.1%
SVD_US	53.4%	61.4%	63.3%	82.9%	62.5%	81.9%
SVD_OS	59.3%	74.5%	63.8%	69.9%	77.6%	62.4%
Our(GAN_SVD)	91.7%	57.7%	75.8%	69.6%	89.1%	67.8%
Our(SVD_GAN)	93.4%	57.1%	81.9%	70.2%	89.3%	69.0%

4.5. Performance Comparison of Ate-of-the-Art Abnormal Behavior Detection Methods

We compare the performance of our approach with the state-of-the-art methods for detecting abnormal blockchain-based transaction accounts [16,42,43]. Refs. [16,42] present two different machine-learning-based abnormal behavior detection methods that take XGboost and random forest algorithms to address security challenges in IoT and Bitcoin

networks. Ref. [16] employs the synthetic minority oversampling technique (SMOTE) to address the imbalance of data. Then, XGboost classifiers are used to classify legitimate and abnormal transactions. Ref. [42] proposes to detect DDoS attacks in blockchain-enabled IoT Networks by using Standard-Scaler data normalization for scaling the feature values and using the random forest method for intrusion detection. Ref. [43] transforms the vulnerability data to a graph and proposes a graph convolutional neural network-based vulnerability detection method in the blockchain. All of the methods use our transaction topology features for abnormal account detection. The performance comparison is shown in Table 3:

- By using SVD-based feature dimension reduction and GAN-based imbalanced data augmentation, our approach achieves the best performance compared with three different abnormal behavior detection methods. The precision of our approach is 89.3%, and the recall of our approach is 69.9%.
- The two machine-learning-based abnormal behavior detection methods [16,42] achieve good precision and recall, since Ref. [16] uses the SMOTE to handle the imbalanced data, and the sparse feature is not sensitive to the random forest method used in [42]. However, the performance of the above two methods is still worse than our proposed method.
- The graph convolutional neural network-based abnormal behavior detection method used in [43] achieves poor performance, which does not mean the method performance is poor in classification tasks, only that the sparse feature is not suitable for deep neural networks, and the imbalanced training data make the results even worse.

Table 3. Experimental results of the performance comparison of different state-of-the-art methods for detecting abnormal blockchain-based transaction accounts.

Methods	Precision	Recall
XGboost [16]	88.9%	66.3%
Random forest [42]	85.8%	66.5%
Graph convolutional neural network [43]	17.8%	10.7%
Our approach	89.3%	69.0%

4.6. Parameter Settings of Machine Learning Methods

Table 4 shows the parameter settings for various machine learning algorithms employed in the abnormal account detection methods. The parameters of each algorithm have been carefully fine-tuned to maximize performance for the abnormal account detection task.

Table 4. Abnormal account detection for different machine learning parameter setting.

Methods	Parameters
SVM	kernel = rbf c = 1 gamma = 0.001
XGboost	classifier = binary:logistic max_depth = 20 learning_rate = 0.01 n_estimators = 100
DT	criterion = entropy max_depth = 100
LR	c = 1 penalty = l2 tol = 0.0001 max_iter = 100
Ada	n_estimators = 100 learning_rate = 0.001
KNN	n_neighbors = 6

- For the support vector machine (SVM) classifier, the RBF kernel is chosen for the SVM method since the RBF kernel surpasses the performance of the linear kernel. The value of c provides a balanced trade-off between achieving a low training error and avoiding overfitting. The gamma parameter determines the influence of a single training topology feature on the decision function, allowing for a smooth transition between topology features.
- XGboost is a gradient boosting framework that employs the binary:logistic objective function, indicating a binary classification problem. The learning_rate is a common starting point that strikes a good trade-off between the speed of convergence and the quality of the final abnormal behavior detection model. The max_depth of the

trees is set to prevent overfitting, and the n_estimators parameter provides a sufficient number of iterations for the XGboost algorithm to refine the abnormal behavior detection model.

- For the KNN algorithm, the n_neighbors parameter is a common choice to prevent overfitting while still capturing the local structure of the topology features.
- For the Ada classifier, the learning_rate is a conservative choice to ensure slow and steady learning, while the n_estimators are sufficient to ensure robust performance without incurring significant overfitting.
- The decision tree classifier employs entropy as the criterion for node splitting, which is a measure of impurity and results in a more balanced tree. The tree is capped at the maximum depth max_depth to manage the complexity of the tree and prevent overfitting.
- Logistic regression is configured with a regularization strength of c, which indicates a moderate level of regularization. This is used to prevent overfitting by penalizing the model for having too many large coefficients. The L2 penalty further aids in regularizing our abnormal behavior detection model, helping to prevent overfitting. Additionally, the tolerance level, tol, and the maximum number of iterations, max_iter, govern the convergence criteria for the optimization algorithm.

4.7. Parameter Settings of the Data Augmentation Method

Table 5 shows the parameter settings of the data augmentation method. The data augmentation model is governed by several parameters.

Table 5. Parameters optimized by data augmentation model.

Parameter	Values
The optimizer	The Adam Optimizer
The first moment estimate	0.99
The second moment estimate	0.999
batch_size	20
feature dimension	8000
latent_dim	1
learning_rate	0.0001
n_cpu	8

The Adam optimizer is utilized in this method. In a GAN, the generator and discriminator often have different training dynamics and convergence speeds, and the adaptive learning rate helps both to update at an appropriate pace. Within the Adam optimizer, we have two hyperparameters: the first moment estimate parameter is an empirical value, and we use the default values, which means that the Adam algorithm gives more weight to recent gradients but still considers the history of gradients to smooth out the updates. The second moment estimate parameter also is an empirical value, and we use the default values, which provide a longer-term memory of past squared gradients, helping to stabilize the parameter updates. The batch_size defines the number of topology data per training iteration. The feature dimension parameter specifies the dimensions of each topology feature. The latent_dim encapsulates the dimensionality of the latent space. The learning rate controls the pace at which the model refines its parameters during training. The n_cpu specifies the number of CPU threads to use during batch generation.

5. Conclusions

In this paper, we propose an abnormal account identification method for blockchain-based transaction networks by using topology feature analysis. We propose an SVD-GAN-based feature reduction and augmentation method to solve the sparse problem and imbalanced training data problem. We used real-life datasets in our experiments to demonstrate that our approach significantly improves performance compared with some machine learning methods and the state-of-the-art methods. In the future, we will apply

our approach to more fields, not only abnormal behavior detection in the blockchain but also common network intrusion detection.

However, the development of blockchain technology is rapidly changing, and there is still vast room for research on abnormal account identification in the blockchain. The limitations of this study are discussed below:

- Long-term effectiveness and iterative model updating. Although the method proposed in this paper has achieved some results in dealing with small samples and high-dimensional sparse features, there are still some limitations. In real-world environments, the model may face different challenges from those of experimental environments, including new attack methods, more complex transaction patterns, and changing blockchain network structures. Therefore, it is crucial to track and evaluate the performance of the model in the long term, and long-term performance monitoring and data collection can be performed in future research by deploying the model in a real-running blockchain system to verify the lasting effectiveness of the model.
- Introducing more types of transaction data. This paper mainly focuses on user transaction mapping features, but in practical applications, there may be other types of transaction data, such as smart contract execution data and transaction amounts. Future research can try to incorporate more types of transaction data into the analysis framework to enrich the feature dimensions and improve the performance of abnormal account identification.

Author Contributions: Methodology,writing—original draft preparation, Y.Y.; method,writing—review and editing, J.Z.; project administration, M.Z.; Supervision, J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by the Major Research Plan of Hubei Province under Grant No. 2023BAA027, the National Natural Science foundation of China under Grant No. 62002106 and 62202146, the Natural Science Foundation of Hubei Province under Grant No. 2022CFB914, and the Research Foundation of Hubei University of Technology under Grant No. BSQD2020066.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, Y.; Tsyvinski, A. Risks and returns of cryptocurrency. *Rev. Financ. Stud.* **2021**, *34*, 2689–2727. [CrossRef]
2. Wang, H.; Ma, S.; Dai, H.-N.; Imran, M.; Wang, T. Blockchain-based data privacy management with nudge theory in open banking. *Future Gener. Comput. Syst.* **2020**, *110*, 812–823. [CrossRef]
3. Campbell-Verduyn, M. Bitcoin, crypto-coins, and global anti-money laundering governance. *Crime Law Soc. Chang.* **2018**, *69*, 283–305. [CrossRef]
4. Keshavarzi, M.; Ghaffary, H.R. I2ce3: A dedicated and separated attack chain for ransomware offenses as the most infamous cyber extortion. *Comput. Sci. Rev.* **2020**, *36*, 100233. [CrossRef]
5. Gainsbury, S.M.; Blaszczynski, A. How blockchain and cryptocurrency technology could revolutionize online gambling. *Gaming Law Rev.* **2017**, *21*, 482–492. [CrossRef]
6. Jamil, F.; Hang, L.; Kim, K.; Kim, D. A novel medical blockchain model for drug supply chain integrity management in a smart hospital. *Electronics* **2019**, *8*, 505. [CrossRef]
7. Liu, C.; Xiao, Y.; Javangula, V.; Hu, Q.; Wang, S.; Cheng, X. Normachain: A blockchain-based normalized autonomous transaction settlement system for iot-based e-commerce. *IEEE Internet Things J.* **2018**, *6*, 4680–4693. [CrossRef]
8. Ahram, T.; Sargolzaei, A.; Sargolzaei, S.; Daniels, J.; Amaba, B. Blockchain technology innovations. In Proceedings of the 2017 IEEE Technology & Engineering Management Conference (TEMSCON), Santa Clara, CA, USA, 8–10 June 2017 ; IEEE: New York, NY, USA, 2017; pp. 137–141.
9. The 2023 Crypto Crime Report: Everything You Need to Know about Cryptocurrency-Based Crime. Chainalysis. Available online https://go.chainalysis.com/2023-crypto-crime-report.html (accessed on 20 February 2023) .
10. Monamo, P.; Marivate, V.; Twala, B. Unsupervised learning for robust bitcoin fraud detection. In Proceedings of the 2016 Information Security for South Africa (ISSA), Johannesburg, South Africa, 17–18 August 2016; pp. 129–134.

11. Gu, Z.; Lin, D.; Wu, J. On-chain analysis-based detection of abnormal transaction amount on cryptocurrency exchanges. *Phys. A Stat. Mech. Appl.* **2022**, *604*, 127799. [CrossRef]
12. Wu, Y.; Tao, F.; Liu, L.; Gu, J.; Panneerselvam, J.; Zhu, R.; Shahzad, M.N. A bitcoin transaction network analytic method for future blockchain forensic investigation. *IEEE Trans. Netw. Sci. Eng.* **2021**, *8*, 1230–1241. [CrossRef]
13. Lorenz, J.; Silva, M.I.; Aparício, D.; Ascensão, J.T.; Bizarro, P. Machine learning methods to detect money laundering in the bitcoin blockchain in the presence of label scarcity. In Proceedings of the First ACM International Conference on AI in Finance, ACM, New York, NY, USA, 15–16 October 2020.
14. Mohy-Eddine, M.; Guezzaz, A.; Benkirane, S.; Azrour, M. An efficient network intrusion detection model for iot security using k-nn classifier and feature selection. *Multimed. Tools Appl.* **2023**, *82*, 23615–23633. [CrossRef]
15. Liang, W.; Xiao, L.; Zhang, K.; Tang, M.; He, D.; Li, K.-C. Data fusion approach for collaborative anomaly intrusion detection in blockchain-based systems. *IEEE Internet Things J.* **2021**, *9*, 14741–14751. [CrossRef]
16. Ashfaq, T.; Khalid, R.; Yahaya, A.S.; Aslam, S.; Azar, A.T.; Alsafari, S.; Hameed, I.A. A machine learning and blockchain based efficient fraud detection mechanism. *Sensors* **2022**, *22*, 7162. [CrossRef] [PubMed]
17. Sanjalawe, Y.K.; Al-E'mari, S.R. Abnormal transactions detection in the ethereum network using semi-supervised generative adversarial networks. *IEEE Access* **2023**, *11* , 98516–98531. [CrossRef]
18. Nouman, M.; Qasim, U.; Nasir, H.; Almasoud, A.; Imran, M.; Javaid, N. Malicious node detection using machine learning and distributed data storage using blockchain in wsns. *IEEE Access* **2023**, *11*, 6106–6121. [CrossRef]
19. Alarab, I.; Prakoonwit, S.; Nacer, M.I. Competence of graph convolutional networks for anti-money laundering in bitcoin blockchain. In Proceedings of the 2020 5th International Conference on Machine Learning Technologies, ACM, Beijing, China, 19–21 June 2020.
20. Sharma, S.; Sharma, R. Forecasting transactional amount in bitcoin network using temporal gnn approach. In Proceedings of the 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), The Hague, The Netherlands, 7–10 December 2020; pp. 478–485.
21. Chen, B.; Zhang, J.; Zhang, X.; Dong, Y.; Song, J.; Zhang, P.; Xu, K.; Kharlamov, E.; Tang, J. Gccad: Graph contrastive learning for anomaly detection. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 8037–8051. [CrossRef]
22. Pourhabibi, T.; Ong, K.-L.; Kam, B.H.; Boo, Y.L. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decis. Support Syst.* **2020**, *133*, 113303. [CrossRef]
23. Xiao, L.; Han, D.; Li, D.; Liang, W.; Yang, C.; Li, K.-C.; Castiglione, A. Ctdm: Cryptocurrency abnormal transaction detection method with spatio-temporal and global representation. *Soft Comput.* **2023**, *27*, 11647–11660. [CrossRef]
24. Liu, S.; Cui, B.; Hou, W. A survey on blockchain abnormal transaction detection. In Proceedings of the International Conference on Blockchain and Trustworthy Systems, Haikou, China, 8–10 August 2023; pp. 211–225.
25. Alarab, I.; Prakoonwit, S.; Nacer, M.I. Comparative analysis using supervised learning methods for anti-money laundering in bitcoin. In Proceedings of the 2020 5th International Conference on Machine Learning Technologies, Jaipur, India, 13–15 February 2020; pp. 11–17.
26. Elbaghdadi, A.; Mezroui, S.; El Oualkadi, A. K-nearest neighbors algorithm (knn): An approach to detect illicit transaction in the bitcoin network. In *Integration Challenges for Analytics, Business Intelligence, and Data Mining*; IGI Global: Hershey, PA, USA, 2021; pp. 161–178.
27. Nerurkar, P.; Patel, D.; Busnel, Y.; Ludinard, R.; Kumari, S.; Khan, M.K. Dissecting bitcoin blockchain: Empirical analysis of bitcoin network (2009–2020). *J. Netw. Comput. Appl.* **2021**, *177*, 102940. [CrossRef]
28. Mohammed, M.A.; Boujelben, M.; Abid, M. A novel approach for fraud detection in blockchain-based healthcare networks using machine learning. *Future Internet* **2023**, *15*, 250. [CrossRef]
29. Voronov, T.; Raz, D.; Rottenstreich, O. A framework for anomaly detection in blockchain networks with sketches. *IEEE/ACM Trans. Netw.* **2023**, *32*, 686–698.
30. Venkatesan, K.; Rahayu, S.B. Blockchain security enhancement: An approach towards hybrid consensus algorithms and machine learning techniques. *Sci. Rep.* **2024**, *14*, 1149. [CrossRef] [PubMed]
31. Zhang, Z. The impact of the artificial intelligence industry on the number and structure of employments in the digital economy environment. *Technol. Forecast. Soc. Change* **2023**, *197*, 122881. [CrossRef]
32. Kufo, A.; Gjeci, A.; Pilkati, A. Unveiling the influencing factors of cryptocurrency return volatility. *J. Risk Financ. Manag.* **2023**, *17*, 12. [CrossRef]
33. Li, H.; Chen, M.; Sun, X.; Chen, J. Mtm-net: A multidimensional two-stage memory-guided network for vedio abnormal behavior detection. *Multimed. Tools Appl.* **2023**, 1–25. [CrossRef]
34. Garcea, F.; Serra, A.; Lamberti, F.; Morra, L. Data augmentation for medical imaging: A systematic literature review. *Comput. Biol. Med.* **2023**, *152*, 106391. [CrossRef] [PubMed]
35. Dhaliwal, S.S.; Nahid, A.-A.; Abbas, R. Effective intrusion detection system using xgboost. *Information* **2018**, *9*, 149. [CrossRef]
36. Chen, W.; Wu, J.; Zheng, Z.; Chen, C.; Zhou, Y. Market manipulation of bitcoin: Evidence from mining the mt. gox transaction network. In Proceedings of the IEEE Conference on Computer Communications, (IEEE INFOCOM), Paris, France, 29 April–2 May 2019.
37. Feder, A.; Gandal, N.; Hamrick, J.T.; Moore, T. The impact of DDoS and other security shocks on bitcoin currency exchanges: Evidence from mt. gox. *J. Cybersecur.* **2017**, *3*, 137–144. [CrossRef]

38. Gandal, N.; Hamrick, J.; Moore, T.; Oberman, T. Price manipulation in the bitcoin ecosystem. *J. Monet. Econ.* **2018**, *95*, 86–96. [CrossRef]
39. Wei, J.; Chen, J.; Zhou, Z. Strategies to "bitcoin-gold" trading decoupling the qualitative decision and the quantitative investment. In Proceedings of the International Conference on Information Economy, Data Modeling and Cloud Computing, ICIDC 2022, Qingdao, China, 17–19 June 2022.
40. Alarab, I.; Prakoonwit, S. Effect of data resampling on feature importance in imbalanced blockchain data: Comparison studies of resampling techniques. *Data Sci. Manag.* **2022**, *5*, 66–76. [CrossRef]
41. Vassallo, D.; Vella, V.; Ellul, J. Application of gradient boosting algorithms for anti-money laundering in cryptocurrencies. *SN Comput. Sci.* **2021**, *2*, 143. [CrossRef]
42. Kumar, R.; Kumar, P.; Tripathi, R.; Gupta, G.P.; Garg, S.; Hassan, M.M. A distributed intrusion detection system to detect ddos attacks in blockchain-enabled iot network. *J. Parallel Distrib. Comput.* **2022**, *164*, 55–68. [CrossRef]
43. Zhuang, Y.; Liu, Z.; Qian, P.; Liu, Q.; Wang, X.; He, Q. Smart contract vulnerability detection using graph neural network.In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI 2020), Yokohama, Japan, 7–15 January 2021; pp. 3283–3290.