electronics

MDPI

Article

Detection of Illegal Transactions of Cryptocurrency Based on Mutual Information

Kewei Zhao [1], Guixin Dong [2,3] and Dong Bian [1,*]

[1] School of Microelectronics, Shandong University, Jinan 250101, China
[2] School of Computer Science, Shandong University of Finance and Economics, Jinan 250014, China
[3] Shenlan Technology of Shandong Research Institute, Jinan 250300, China
* Correspondence: biand@sdu.edu.cn

Abstract: In recent times, there has been a swift advancement in the field of cryptocurrency. The advent of cryptocurrency has provided us with convenience and prosperity, but has also given rise to certain illicit and unlawful activities. Unlike classical currency, cryptocurrency conceals the activities of criminals and exposes their behavioral patterns, allowing us to determine whether present cryptocurrency transactions are legitimate by analyzing their behavioral patterns. There are two issues to consider when determining whether cryptocurrency transactions are legitimate. One is that most cryptocurrency transactions comply with laws and regulations, but only a small portion of them are used for illegal activities, which is related to the sample imbalance problem. The other issue concerns the excessive volume of data, and there are some unknown illegal transactions, so the data set contains an abundance of unlabeled data. As a result, it is critical to accurately distinguish between which transactions among the plethora of cryptocurrency transactions are legitimate and which are illegal. This presents quite a difficult challenge. Consequently, this paper combines mutual information and self-supervised learning to create a self-supervised model on the basis of mutual information that is used to improve the massive amount of untagged data that exist in the data set. Simultaneously, by merging the conventional cross-entropy loss function with mutual information, a novel loss function is created. It is employed to address the issue of sample imbalance in data sets. The F1-Score results obtained from our experimentation demonstrate that the novel loss function in the GCN method improves the performance of cryptocurrency illegal behavior detection by four points compared with the traditional loss function of cross-entropy; use of the self-supervised network that relies on mutual information improves the performance by three points compared with the original GCN method; using both together improves the performance by six points.

Keywords: cryptocurrency; mutual information; loss function; self-supervision; GCN

Citation: Zhao, K.; Dong, G.; Bian, D. Detection of Illegal Transactions of Cryptocurrency Based on Mutual Information. *Electronics* **2023**, *12*, 1542. https://doi.org/10.3390/electronics12071542

Academic Editors: Xiushan Nie, Guoqiang Zhong, Yongshun Gong, Bin Fan and Xin Li

Received: 27 February 2023
Revised: 18 March 2023
Accepted: 21 March 2023
Published: 24 March 2023

1. Introduction

People became increasingly interested in cryptocurrency on 16 February 2020, as Bitcoin broke through the USD 50,000 mark.Bitcoin is one of the most popular cryptocurrencies. Cryptocurrency is mainly characterized by its decentralized nature, as it facilitates peer-to-peer transmission for transactions. Unlike classic currency, cryptocurrency is issued by a proprietary algorithm rather than a specific organization. The cryptocurrency is then created through a lengthy and never-ending calculation. The cryptocurrency records and confirms all transaction activities via a distributed database composed of nodes across the entire peer-to-peer network, and employs cryptographic techniques to guarantee the security of every stage [1] in the currency circulation process. In general, cryptocurrency serves as a decentralized distribution system for currency and a network for currency circulation and settlement. It not only addresses the issue of excessive issuance of conventional credit currency, while enables inexpensive and nearly instantaneous currency settlement. Cryptographic currency has the characteristics of anonymity, decentralization, strong cross-border circulation, difficult supervision, and opaque transaction information, which brings

convenience but also many problems, such as money laundering, illegal transactions in secret networks, and online extortion. The emergence of cryptocurrency provides a natural umbrella for criminals, and criminals adore it. The emergence of cryptocurrency provides a natural umbrella for criminals, and criminals adore it. Furthermore, its value is rising precisely because it provides a safe and dependable trading method for criminals [2]. For example, when the once-famous online black market "Silk Road" was shut down, the price of Bitcoin fell precipitously. Cryptocurrency transactions are linked to one another. In recent years, graph neural networks (GNNs) get increasingly popular in the realm of deep learning, exhibiting impressive outcomes in the processing of graph structures. In graph neural networks, there are two common methods. One such type is the spatial-based graph convolutional network, which utilizes data from neighboring nodes to carry out graph convolutions. The other is a spectrum-based graph convolution network. As this methodology regards graph convolution as a mechanism for eliminating noise from an image signal, this paper regards graph convolutional network (GCN) as well as graph attention network (GAT) as the foundational models for monitoring illegal cryptocurrency transactions [3–5]. The number of transactions in the cryptocurrency trading network is extremely high, reaching tens of thousands per minute. It is extremely difficult to distinguish between legal and illegal addresses in these transactions [6,7]. Consequently, the training data comprises a significant volume of untagged data. Furthermore, the detection of cryptocurrency is an abnormal detection. Most cryptocurrency holders use it legally, but also a minority engaging in illicit transactions such as money laundering or extortion. Consequently, it is challenging to precisely classify a little quantity of unlawful transactions within an extensive array of cryptocurrency transactions. This is a common data imbalance issue. To tackle these problem, this article suggests two potential solutions. The first solution involves implementing a self-supervised approach grounded in mutual information to address the issue of a vast quantity of untagged data in the training dataset. Previous studies usually use classical methods to try to solve the second problem, including weight adjustments for legal and illegal usage categories and the replacement of the original loss function with Focal loss [8], and so on. To overcome the challenge posed by data imbalance, this study explores a loss function that relies on mutual information.

1. Aiming at a large amount of unlabeled data in the training data, a self-supervised method is used to alleviate it.
2. Data imbalance problem is addressed through application of the novel loss function by considering mutual information.
3. Experiment is conducted on real data sets.

The transaction record data type of cryptocurrency enables it to be modeled through the data structure of the graph. As a new method to process graph structure data, graph neural network is used in this paper. We verify the validity of our loss function and other loss functions through the model constructed by the graph neural network. At the same time, self-monitoring method is used to alleviate the imbalance between labeled data and unlabeled data. The experimental results verify the validity of our model.

With rapid development in the graph neural network, which provides better data feature extraction and analysis for cryptocurrency trading, this paper studies this method with two problems existing in the field of illegal transactions detecting in cryptocurrency. The paper is divided into five sections. The first chapter provides an in-depth investigation on the illegal behavior detection of cryptocurrency based on mutual info-graphic neural network from the aspects of background and significance. In Section 2, we review existing research methods for cryptocurrency violation detection, self-supervision, and data imbalance oriented improvements. In Section 3, we introduce the cryptocurrencies illegal detection using a novel loss function considering mutual information prior and the cryptocurrencies illegal detection based on mutual information self-supervised module learning. In Section 4, experiments are conducted on the cryptocurrency data sets and the superiority of the proposed model is verified. Section 5 summarizes and provides concluding remarks.

2. Related Work

The detection of cryptocurrency violations is critical to maintaining network financial security, controlling network financial chaos, and avoiding network financial risks [9–12]. At the moment, the detection of cryptocurrency violations is involved in a variety of fields, including finance, economics, law, and even politics. The detection of cryptocurrency violations has numerous applications and significant research significance in both the financial and scientific and technological fields. However, the issue of severe data imbalance in illicit cryptocurrency transactions persists, and the primary objective of researchers is to devise effective strategies to mitigate the challenge of data imbalance. Following that, this paper will provide a brief overview of existing research on cryptocurrency violations detection, self-supervised of graphs, and Enhancing loss functions has emerged as a promising approach to tackle the problem of data unbalance.

2.1. Detection in Cryptocurrency Violation

Akcora C [13] proposed a solution based on topological data analysis. The bitcoin net was divided into day scaled windows, and the features such as income and neighbor were extracted for each address. This approach is designed to identify new addresses associated with established ransomware families and predict the emergence of addresses linked to unknown ransomware families. The new tda-based tool significantly improves ransomware detection accuracy. Chen w et al. [14] used a model to address the class imbalance issue in phishing fraud identification, which involves a cascade feature extraction technique relied on transaction graph as well as the double sampling integrated method using lightGBM. Its goal is to improve whole blockchain [15] ecosystem. Additionally, provide user of early phishing fraud warning. Weber et al. [16] proposed several methods for authenticating cryptocurrency transactions, particularly for Bitcoin, to combat criminal activity. These efforts aim to improve human analysis and explanation capabilities while accelerate the global financial system more secure and reliable. Although neural networks were not tested in this research, their structure is suitable for modeling complex nonlinear data.

2.2. Graph Self-Supervised

Xiao liu [17] and colleagues summarized four major kinds of self-supervised methods based on production, the last in which is the hybrid of the first three methods, and the third, ae method, should be used more frequently in the field of graph learning. The auto-encoder model is similar to the principal component analysis method in that it maps the original feature into new dimensions then back to initial dimensions. This operation, like the principal component analysis method, must ensure that the mapped target retains some properties (nodes with high similarity should still have high similarity after mapping), while also reducing noise. Kaveh Hassani [18] put forward a technique for self-supervised, which aims to acquire nodes-hierarchy and graph-hierarchy expression through contrast configurable view in the graph. In contrast to visual representation learning, this paper contends that generating views quantity to exceed two or comparing multi-level coding will not enhance performance. Excellent performance can be obtained through comparing first order neighbor coding as well as graph diffusion. In the field of graph neural networks, Jie Zhong Qiu [19] developed a pre-training framework called graph comparison coding (GCC) as a self-supervised approach to learning representations of nodes and graphs by analyzing and capturing the underlying structural similarities across multiple networks. By using comparative learning, the graph neural network can learn intrinsic and transferable structural representations, with GCC's pre-training task specifically designed to distinguish subgraph instances within and between networks. Experimental results demonstrate that GCC pre-trained on various data sets could obtain excellent or superior performance for its particular tasks compared to training from scratch.

2.3. Improvement for Data Imbalance Problem

Cao k et al. [20] proposed a theoretically sound marginal ldam loss of label distribution that seeks to minimize the generalization boundary based on the edge. The proposed approach involves using a new loss function in place of the cross-entropy target when train the model. This loss function could be used with existing kind imbalance training polices such as re-weighting as well as re-sampling. Additionally, a novelty training policy is proposed where re-weighting is deferred until the initial stage. This enable the model to acquire the original expressions without the added complexity of re-weighting or re-sampling. Buda m et al. [21] conducted a systematic study on how sample imbalance affects model ability for classification assignments on three general datasets of different scales, namely MNIST, CIFAR-10, and ImageNet. They also explored various ways to take care of sample imbalance problem, including oversampling, downsampling, two-stage training, as well as threshold-based methods. The sample imbalance was discovered. At the moment, the mainstream method is primarily over-sampling; however, it is not always the case that over-sampling results in over-fitting for general networks. Lin t y et al. [8] introduced a loss function called focal loss, which was developed to address the issue of category imbalance and improve the performance of one-stage methods in comparison to two-stage methods. Focal loss operates by incorporating a modulating factor into the cross-entropy loss function that focuses on hard examples while reducing the weight of easy negatives. This model considering mutual information prior has significant implications for cryptocurrency illegal behavior detection, as it effectively addresses the serious sample imbalance Bitcoin, leading to improved precision as well as recall rates for illegal category.

The illegal detection of cryptocurrency is an anomaly detection [22] because of the data imbalance and large amount of unlabeled data. Existing studies usually use classical methods to solve the problem of data imbalance, such as changing legitimate weights as well as illegal cryptocurrency categories, or using more effective Focal loss fuction, etc. Instead, this article will consider using a loss function by considering mutual information solving data imbalance problem. At the same time, different from the existing work based on graph self-supervised learning, this paper uses the self-supervised method based on mutual information in graph neural network to address the issue of an abundance of unlabeled data in training dataset.

3. Method

3.1. Question Raised

The task of identifying illegal cryptocurrency transactions can be approached as a standard machine learning problem by treating it as a binary classification issue. This means the transaction could be classified as legal or illegal based on the information available to an unknown node.

In this paper, the problem is formally defined as:

Input: $P \in R^{|J|*K}$,where $|J|$ refers to node number and K refers to feature dimension. Output: $O \in \{0, 1\}$. Where 0 is the legal class and 1 is the illegitimate class.

To overcome the difficulties posed by lager amount of unknown nodes as well as severe data imbalance within legal and illegal categories, this study proposes using mutual information as the prior loss function through a self-supervised model.

3.2. Monitoring Illegal Transactions of Cryptocurrency Relied on GNN

The graph structure on cryptocurrency transactions is commonly modeled using graph structures, and graph neural network (GNN), which gain popularity in the field of deep learning, have proven effective in processing such structures. Therefore, in this study, the graph convolutional network (GCN), as the fundamental model of GNN, is considered as a model for monitoring illegal cryptocurrency transactions.

Graph Convolution Network

The GCN is a multi-layer graph convolution algorithm that, like the cognitive algorithm, takes spectral convolution to gather neighbor information. The cryptocurrency transaction graph from the data set is assumed to be $G = (X, Y)$, where X refers to node characteristics as well as Y refers to edge characteristics in the graph. Then, the inputs of each layer in the GCN are node characteristic matrix M as well as adjacency matrix N, while the output is matrix M updated by the weight matrix W. The adjacency matrix N represents the flow between two cryptocurrency transactions. A value of 1 in the matrix means there is a connection between the two transactions, and a value of 0 means there is no connection. H and W may differ at each level, but the adjacency matrix N remains constant. The mathematical formula for the graph convolution network is defined in Equation (1):

$$M^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{N}\tilde{D}^{-\frac{1}{2}}M^{(l)}W^{(l)}\right) \tag{1}$$

where l represents the layer number, and σ refers to Relu function. The input characteristic matrix adopts the characteristics of graph nodes, so $M^{(0)} = P$. $\tilde{N}$ Represent adjacency matrix plus self-linking to represent degree matrix, and the Formula (2) is, and the specific formula is: $\tilde{D}\tilde{N}\tilde{D}$

$$\tilde{N} = N + I, \tilde{D} = \text{diag}\left(\sum_j \tilde{N}_{ij}\right) \tag{2}$$

Hypothetically, a GCN is used for node classification, setting up a graph convolution l layer with ReLU as the activation function and a softmax output as the final output layer. The node's features $M^{(0)}$ and adjacency matrix A are first input. Convolution through layer l graph and finally output $M^{(l)}$ through softmax. The $M^{(l)}$ consists of the predicted probabilities. Each layer of the graph convolution varies from its feedforward counterpart solely by incorporating the product of the preceding $\hat{N}$. The layer could be perceived as a collection of transformed embeddings of neighboring nodes. The spectral filter drives the layer and is obtained by applying a linear function to the Laplacian matrix. The weight to be optimized is $W^{(l)}$. The 2-layer graph convolution network (GCN) is mathematically formulated in Equation (3).

$$M^{(2)} = \text{softmax}\left(\hat{N} \cdot \text{ReLU}\left(\hat{N}PW^{(0)}\right) \cdot W^{(1)}\right) \tag{3}$$

3.3. Self-Supervised Learning

Yann lecun said in their speech, "Analogously speaking, self-supervised learning forms the major portion of the cake in the field of artificial intelligence, while supervised learning acts as the icing on top. Reinforcement learning (RL) can be considered as the cherry on the cake." Although it cannot be said that what he said is completely correct, it can also show that Self-supervised learning has become increasingly important in artificial intelligence [23,24]. Self-supervised refers to the conversion of unsupervised machine learning problems to supervised machine learning problems, and then deal with it by using the supervised learning method. Generally speaking, pseudo-labels are constructed by using the characteristics and attributes of data sets to replace the labels set by human beings.

3.3.1. Mutual Information

Self-supervised learning extracts useful information from large-scale unlabeled data sets using a predefined auxiliary task, then creates its own labels using this information, feeds it into a neural network for training, and learns valuable information. Mon et aldim.'s theory can determine whether the learned information is valuable. The theory holds that the reconstruction error is small, which cannot account for the good learned features. Good features should be the samples' most distinct and specific information, and mutual information should be used. Mutual information is the statistical metric which quantifies the amount of dependence or correlation between two random variables, i.e., the degree

to which the uncertainty of O is reduced after a given P. If P and O have no relationship, value of mutual information is zero, and if the given random variable P could fully remove the nondeterminacy of another random variable O, and the value of mutual information between P and O is equal to the maximum entropy value in P.

Formula (4) provides the mathematical expression for the entropy of a discrete random variable.

$$S(A) = -\sum_{i=1}^{n} p(a_i) \log_b(p(a_i)) \tag{4}$$

In which event A has n states, I represents the state number, as well as base b is usually takes 2, and can also be set to 10 or e. For a set of random variables (a, v), the joint entropy is defined similarly to that of a single discrete random variable $S(A; V)$. Equation (10) provides formula for the joint entropy.

$$S(A; V) = -\sum_{v \in V} \sum_{a \in A} p(a, v) \log(p(a, v)) \tag{5}$$

where, the joint probability distribution function of two random variables A and V is denoted as $p(a, v)$, and the marginal probability distribution functions of A and V are denoted as $p(a)$ and $p(v)$. Equation (11) provides the equation for mutual information:

$$G(A; V) = S(A) + S(V) - S(A, V) \tag{6}$$

where $S(A)$,$S(V)$,$S(A, V)$ is greater than 0 and $G(A; V)$ must also be greater than 0.

By introducing the Formula (10) into the Formula (11), it can be obtained

$$\begin{aligned} G(A; V) &= \sum_{v \in V} \sum_{a \in E} p(a, v) \log \frac{p(a, v)}{p(a)p(v)} \\ &= \sum_{v \in V} \sum_{a \in A} p(v \mid a) p(a) \log \frac{p(v \mid a)}{p(v)} \end{aligned} \tag{7}$$

With Equation (12) it can be observed that mutual information value needs to max and $\frac{p(v|a)}{p(v)}$ will also be as large as possible, which means that $p(v)$ will be smaller than $p(v \mid a)$. The variable e takes for the input to the neural network, while v can be considered as the output, i.e., the learned features.

It can be stated that for each input e, the network is capable of identifying the unique feature v that corresponds to that input. Therefore we can also discriminate the original sample well by analyzing and learning only the learned feature v.

Mutual information quantifies the amount of information that input and output share, and if a model can directly learn to maximize mutual information, it can acquire more essential knowledge than just fitting conditional probability. Since mutual information is a measure that exposes the fundamental correlation between the input and output.

3.3.2. Maximize Mutual Information

Formula (8) can be obtained by changing the formula of mutual information.

$$G(A; V) = KL(p(a, v) \| p(a)p(v)) \tag{8}$$

Mutual information could be refered to KL divergence of the product of joint distribution of variables a, v and their edge distribution. In other words, maximizing mutual information is to maximize the distance of the product of joint distribution and edge distribution. However, there is no upper bound for KL divergence in theory, so Equation (14) converts KL divergence into JS divergence.

$$JS(R, T) = \frac{1}{2} KL\left(R \| \frac{R+T}{2}\right) + \frac{1}{2} KL\left(T \| \frac{R+T}{2}\right) \tag{9}$$

The upper bound of js divergence is log2/2, at this time, we change the problem of maximizing mutual information into maximizing js divergence.

It is very difficult to directly calculate the divergence of two probability distributions p and q, because generally speaking, we do not have a mathematical formula or expression that can describe or represent the two probability distributions r and t, and all we have is the samples obtained from the two distributions. Equation (15) is a generalized form of divergence.

$$D_f(R\|T) = \int r(A) f\left(\frac{r(A)}{t(A)}\right) dE \tag{10}$$

Sebastian nowozin et al. [25] proposed a method of estimating various kinds of divergence by using GAN, which is called f-GAN. By the approach suggested by Sebastian nowozin et al., the divergence is estimated as Formula (11).

$$\begin{aligned} D_f(R\|T) = \max_U \Big(& E_{a\sim p(a)}[U(a)] \\ & - E_{e\sim t(a)}[c(U(a))] \Big) \end{aligned} \tag{11}$$

where c denotes the conjugate of the function f, and the function $U(a)$ is implementable through a neural network. Equation (11) represents the sampling of two distributions. By calculating the expectation of $U(a)$ and $c(U(a))$, optimising U and maximising the discrepancy between $U(a)$ and $c(U(a))$, the final result is the estimate of the scatter.

Equation (12) is the estimation formula of js divergence with constant term removed:

$$\begin{aligned} JS(R,T) = \max_D \Big(& E_{a\sim r(a)}[\log \sigma(L(a))] \\ & + E_{a\sim t(a)}[\log(1-\sigma(L(a)))] \Big) \end{aligned} \tag{12}$$

Equation (13) is the objective function of maximizing mutual information:

$$\begin{aligned} & JS(r(a,v), r(a)r(v)) = \\ & \max_L \Big(E_{(a,v)\sim r(a,v)}[\log \sigma(L(a,v))] + \\ & E_{\tilde{a}\sim r(a),\tilde{v}\sim r(v)}[\log(1-\sigma(L(\tilde{a},\tilde{v})))] \Big) \Big) \end{aligned} \tag{13}$$

where, $\sigma(L(a,v))$ is the discriminant network, a and its corresponding v are the positive sample pairs, a and the randomly selected v are the negative sample pairs, and finally the likelihood function is maximized for them.

3.3.3. Maximizing Mutual Information of Graphs

Finally, maximize mutual information between local and global features because, in general, global features are better suited for reconstruction while local features are better suited for classification.

Each node in the graph contains rich information about nodes. The information associated with each node is fed into the GCN, and the information of the surrounding nodes is integrated together as $\vec{r}$. $\vec{r}$ is regarded as the local feature of the node.

Global features $\vec{c}$ are obtained by averaging all local features in the current graph.

In order to obtain suitable negative samples, we tried several construction methods. The first one is to change the order of local features without changing the global features. The second one is to replace the local features of the current graph with those of other graphs. The third one is to replace the global features of the current graph with those of other graphs. The fourth one is to replace the local features and global features of the current graph with those of other graphs. Through experiments, it is found that only the first and third ones converge.

Using a binary classifier to determine positive and negative, positive samples are local features and global features of the current graph, and negative samples are composed of

both local features that alter the arrangement of the current graph, as well as global features of the current graph. The positive sample pair consists of $(\vec{r}, \vec{c})$, whereas the negative sample pair consists of $(\vec{r}, \vec{c})$.

Finally, the loss function of the whole self-supervised part is shown in Equation (14).

$$B = \frac{1}{W+V}\left(\sum_{i=1}^{W} E_{(A,N)}\left[\log D\left(\overrightarrow{r_i}, \vec{c}\right)\right] + \sum_{j=1}^{W} E_{(\tilde{A},\tilde{N})}\left[\log D\left(\overrightarrow{\tilde{r}}_j, \vec{c}\right)\right]\right) \tag{14}$$

The equation involves several variables, including D which represents the discriminator, (A, N) which is the feature matrix and adjacency matrix of positive samples, $(\tilde{A}, \tilde{N})$ which is the feature matrix and adjacency matrix of negative samples, W which is the number of positive samples, and V which is the number of negative samples.

3.4. Mutual Information as Prior Loss

The loss function in statistics quantifies the error of a system. In a supervised learning model, the loss function measures the discrepancy between the predicted output of the model and the true label of a sample. This section will focus on how to enhance the loss function using mutual information to address the issue of class imbalance.

3.4.1. Cross Entropy

The formula of the cross entropy loss function is

$$H(r, t) = -\sum_{i} r(h_i) \log t(h_i) \tag{15}$$

where, the network's output, $t(h_i)$, represents the result obtained after inputting the samples into the neural network, whereas $r(h_i)$ is the distribution of the expected samples, i.e., the label of the actual data. This paper primarily focuses on label classification, assuming that there are k categories, with training data denoted as $(a, v) \sim D$ and the modeled distribution as $r_\theta(v \mid a)$. The optimization objective is to maximize likelihood or minimize cross-entropy, with Equation (16) providing the formula of minimizing cross-entropy.

$$\arg\min_{\theta} \mathbb{E}_{(a,v)\sim D}[-\log r_\theta(v \mid a)] \tag{16}$$

3.4.2. Improving the Loss Function Based on Mutual Information

The activation function used in the final layer of a neural network for binary or multi-classification tasks is often Softmax, which is preferred due to its normalization function and ease of computation. The generalized Softmax formula is presented in Equation (17):

$$t\left(h_j\right) = \frac{e^{z_j}}{\sum_{i=1}^{n} e^{z_i}} \tag{17}$$

The output of the previous layer in neural network, is denoted as z_j, while $t\left(h_j\right)$ represents the distribution form of the output of this layer. Moreover, e^{z_j} represents the sum of e^{z_j} within a batch.

Firstly, it is assumed that the logits are $f(a; \theta)$, which is the network's output. Equation (18) can be obtained by substituting this into the softmax formula.

$$r_\theta(v \mid a) = \frac{e^{f_v(a;\theta)}}{\sum_{i=1}^{K} e^{f_i(a;\theta)}} \tag{18}$$

The Equation (19) represents the loss function form of Equation (18).

$$-\log p_\theta(v \mid a) = -\log \frac{e^{f_v(a;\theta)}}{\sum_{i=1}^{K} e^{f_i(a;\theta)}} = \log\left[1 + \sum_{i \neq v} e^{f_i(a;\theta) - f_v(a;\theta)}\right] \tag{19}$$

The Equation (19) refers to the conventional softmax cross-entropy.
Equation (20) presents the model for mutual information:

$$\log \frac{r_\theta(v|a)}{r(v)} \sim f_v(a;\theta) \Leftrightarrow \log r_\theta(v \mid a) \sim f_v(a;\theta) + \log r(v) \tag{20}$$

The Equation (21) represents the softmax that has been re-normalized in the form on the right-hand side.

$$r_\theta(v \mid a) = \frac{e^{f_v(a;\theta)+\log r(v)}}{\sum_{i=1}^{K} e^{f_i(a;\theta)+\log r(i)}} \tag{21}$$

Equation (22) represents the loss function derived from Equation (21):

$$-\log r_\theta(v \mid a) = -\log \frac{e^{fv(a;\theta)+\log r(v)}}{\sum_{i=1}^{K} e^{f_i(a;\theta)+\log r(i)}} = \log\left[1 + \sum_{i \neq v} \frac{r(i)}{r(v)} e^{f_i(a;\theta) - f_v(a;\theta)}\right] \tag{22}$$

More generally, Equation (23) with the addition of the moderator τ:

$$-\log r_\theta(v \mid a) = -\log \frac{e^{f_v(a;\theta)+\log r(v)}}{\sum_{i=1}^{K} e^{f_i(a;\theta)+\log r(i)}} = \log\left[1 + \sum_{i \neq v} \left(\frac{r(i)}{r(v)}\right)^{\frac{1}{\tau}} e^{f_i(a;\theta) - f_v(a;\theta)}\right] \tag{23}$$

Although the model uses the same cross-entropy as the loss function, it is effectively to fit mutual information. Each logarithmic output receives an offset related to label prior (that is, the result before being activated by softmax).

Combined with the process of training in neural network Figure 1:

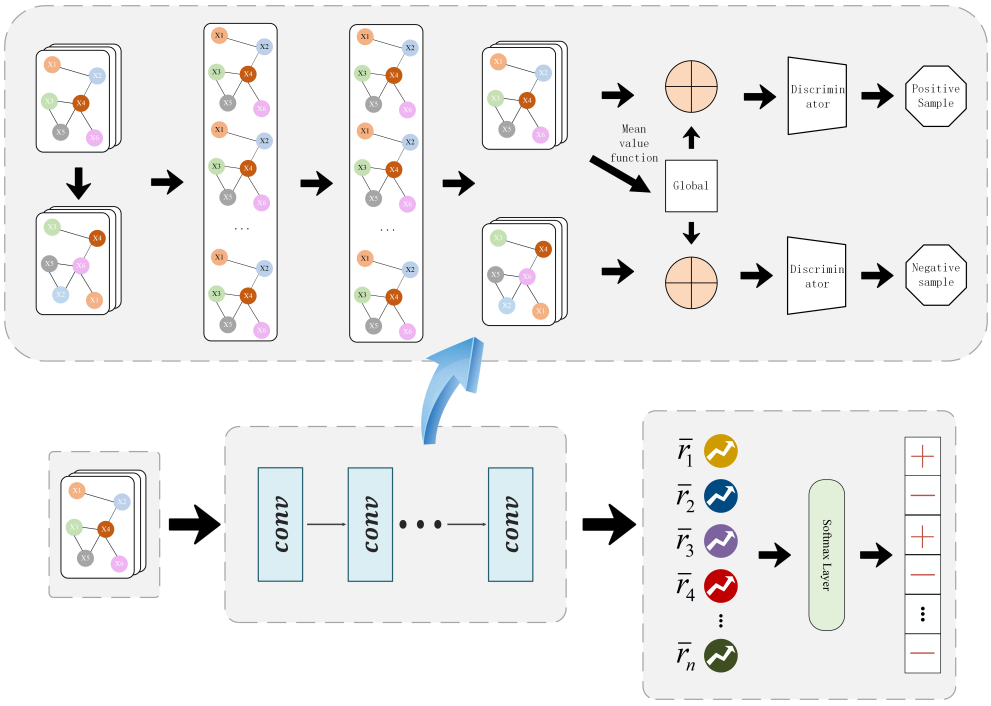

Figure 1. Training process in neural network.

The above procedure begins with randomly disrupting the original samples to obtain negative samples. After the creation of positive and negative samples, the local variables are then input into the convolutional neural network. The global variables are obtained by averaging the positive samples' local variables and coupling them with the positive and negative samples' local variables, respectively. In the model, the global variables are connected with the local variables of the both kinds samples. The trained network model's parameters are scored by feeding them into the discriminator. To clarify, the parameters of the trained neural network remain fixed, and the positive samples are inputted into the network to obtain a set of logits. These logits are then subtracted by the logarithm of the prior probability of each label, denoted by $\mathrm{Inp}(y)$, which is a form of regularization. The resulting values are then passed through a softmax activation layer to obtain the predicted output y. This process is often used to mitigate the impact of class imbalance in the training data. In the figure, $\bar{r}_i, i \in (1, n)$ represents the output combined with the vivid information of the label, n represents node numbers, $x_i, i \in (1, n)$ refers to the first node, positive sign means legal, negative sign means illegal, and the set of positive and negative signs represents the model's output y.

4. Experiments and Analysis

4.1. Dataset

Elliptic Company provided the data set used in this paper. The data set consists of 203,769 transactions nodes as well as 234,355 edges. Out of these transactions, approximately 21% (42,019) are labeled legal, 2% (4545) are labeled illegal, and the remaining transactions are labeled unknown, while they all have other associated characteristics. The dataset comprise 49 kinds graphs, and none of these graphs are related to each other. The node can take place of the transaction, and the flow of bitcoins is represented by an edge. There are currently two issues with this data set. The paper discusses two challenges in the given dataset. The first challenge is the presence of mass unlabeled nodes. The second challenge is the data imbalance problem, where the number of nodes labeled as illegal is only 10% of the nodes labeled as legal, which is a significant difference (Figure 2).

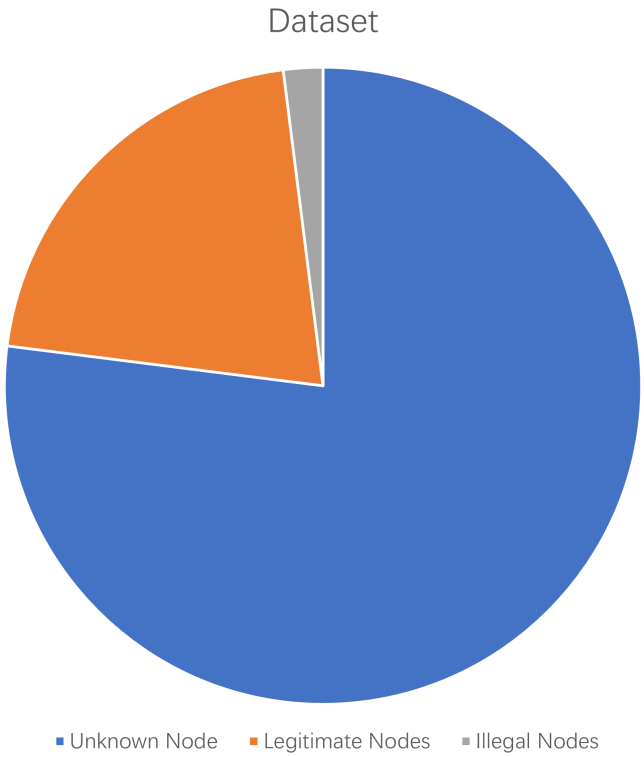

Figure 2. The proportion of different nodes in the data set.

4.2. Metrics

For the sake of assess approach superiority outlined in this paper in an unbiased manner, it is necessary to assess the performance when model has finished training. For comparative analysis, the evaluation indexs used include Precision, Recall, and F1-Score. The task at hand is binary, with a notable difference in the number of instances between the legal and illegal categories. The minority category (i.e., the illegal category) is more important for detecting cryptocurrency violations. As a result, the F1-Score of the illegitimate category will be the focus of this paper. The illegitimate category will be considered a positive category, while the legitimate category will be considered a negative category.

The precision calculation formula is:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{24}$$

Recall is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{25}$$

The formula of F1-score is:

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{26}$$

TP is true, which means that the correct number is classified as illegal. Meanwhile, *FP* refers to false positive, which means the wrong number is classified as illegal. *TN* (true negative) and *FN* (false negative) are the opposite of *TP* and *FP*.

4.3. Experiment with Result Analysis

The environment for this experiment consisted of an Ubuntu operating system, an Intel Xeon CPU, 318GB of memory, and three NVIDIA GeForce RTX 1080 8G graphics cards. The GPU acceleration library used was CUDA 10.0, and PyTorch was used as the deep learning framework. In Elliptic dataset, training set accounts for 60 percent, verification set accounts for 10 percent, test set accounts for 30 percent, where the first 31 graphs are take to train, the next 5 graphs are take to validation, and the final 13 graphs are take to test. The following are the experimental results conducted on the Elliptic dataset. Because all of the models following graph 43 are ineffective, it is discovered that the United States has severely cracked down on cryptocurrency crimes during that time period, with only one or two of the graphs from graph 43 to graph 49 being marked as illegal acts. As a result, from graph 37 to graph 42, this paper compares data from various models and loss functions. The training of the GCN model in this paper was performed using the Adam optimizer with a learning rate of 0.001 for a total of 1000 epochs. The model is composed of two hidden layers, each comprising 100 nodes. The mutual information-based loss proposed is compared to the classic solutions to data imbalance (over-sampling and under-sampling) and the recent solutions to data imbalance. Figure 3 summarizes the F1-score prediction results.

Figure 3 shows that traditional methods (over-sampling and under-sampling) have not improved the data imbalance in anti-money laundering, but have resulted in a decrease in results. Based on the experimental results, the mutual information-based loss proposed in the paper has shown improvement when compared to traditional solutions such as over-sampling and under-sampling, as well as recent solutions to data imbalance. Therefore, it can be concluded that the mutual information-based loss has a significant effect on improving data imbalance in anti-money laundering tasks. Figure 4 depicts a comparison of different loss functions after and before self-supervised.

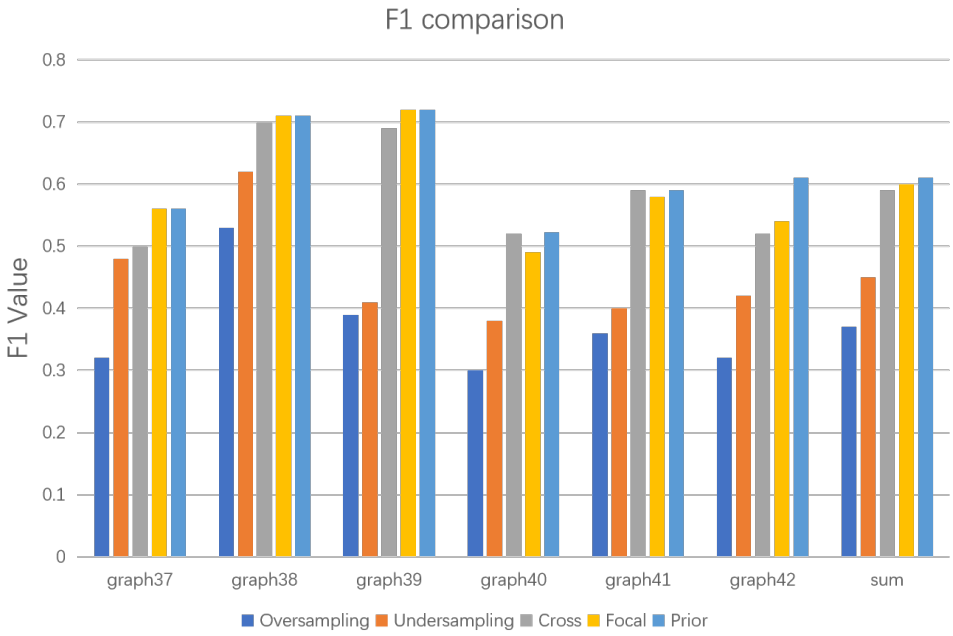

Figure 3. F1 score of different loss functions.

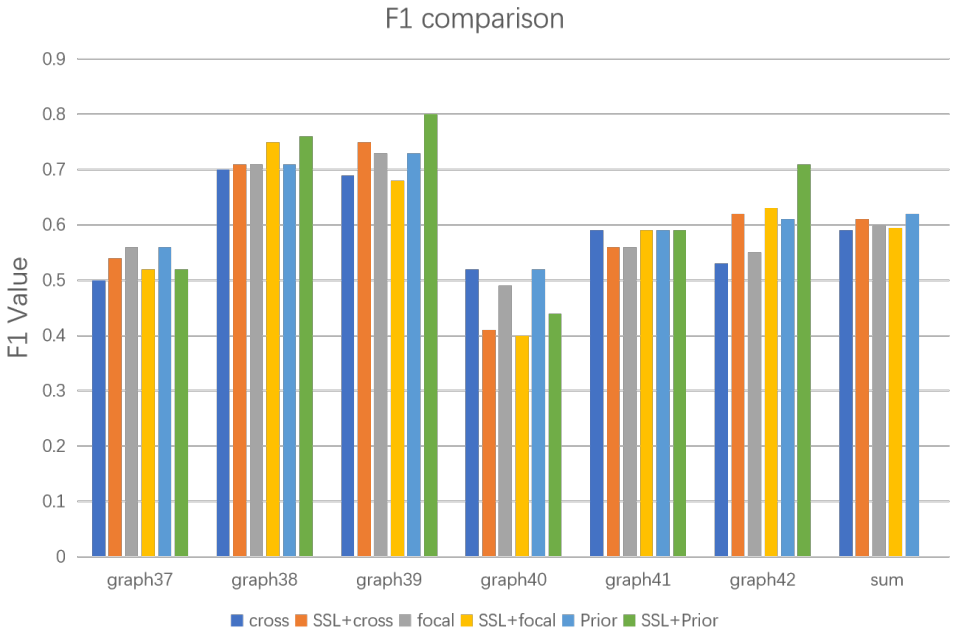

Figure 4. Experimental results after adding self-supervision.

From Figure 4, it can be seen that cross and prior have improved to some extent compared with those without self-supervised, while focal loss has declined slightly. The results indicate that incorporating a self-supervised mechanism can effectively address the challenge of abundant unlabeled data in cryptocurrency illegal detection.

Relied on the data presented from the Table 1, we can draw the following conclusions:

- Traditional methods for dealing with data imbalance are ineffective in detecting illegal cryptocurrency transactions. It can be seen that recall is relatively high regardless of over-sampling or under-sampling, but precision is relatively low, resulting in the final F1 value having a poor effect.
- Drawing from the presented table, we can infer that the mutual information-based loss function outperforms with cross entropy loss and focal loss by 4% and 2%, respectively,

in terms of F1-score when used with GCN. Thus, it can be inferred that the mutual information-based loss function has a significant positive impact on the detection of illegal cryptocurrency transactions.

- The implementation of self-supervised in GCN has alleviated the problem of illegal cryptocurrency detection. The F1-score of the cross-entropy loss with self-supervised have rised 3% compare with the F1-score of the cross-entropy loss without self-supervised. The loss function F1-score with self-supervised increased by 2% when compared to the loss function F1-score without self-supervised. There is no improvement and no intention of decreasing the F1-score value of focal loss. It demonstrates that the self-supervised mechanism can effectively reduce the existence of illegal cryptocurrency transactions.

The chart above provides a visual representation of the effectiveness of the mutual information-based loss function in addressing the issue of sample imbalance in illegal cryptocurrency activities detection. The introduction of self-supervised mechanisms can also be seen to alleviate the issue of mass unlabeled data in illegal cryptocurrency transactions.

Table 1. The experimental results of performace metrics on Graph37-42 with different methods and loss.

Model	Loss	Precision	Recall	F1
GCN	Oversampling	0.23	**0.86**	0.37
	Undersampling	0.31	0.81	0.45
	cross	**0.70**	0.52	0.58
	focal	0.68	0.55	0.60
	prior	0.69	0.58	**0.62**
SSL	cross	0.57	0.65	0.61
	focal	0.53	**0.69**	0.60
	prior	**0.66**	0.63	**0.64**

To address the severe class imbalance between legal and illegal categories in cryptocurrency transaction detection, our method combines mutual information as well as cross entropy loss function to obtain a novel loss function. On illegal cryptocurrency transactions dataset, we adopt the novel loss function considering prior mutual information and the classical method of data imbalance processing, Focal loss and cross entropy loss function, to conduct a comparative experiment.The results showed that F1 value was improved by using the proposed method by contrast of the previous cross entropy loss function as well as Focal loss. Meanwhile, we take the self-supervised learning method to deal with the problem of large amount of unlabeled data in cryptocurrency illegal behavior detection. On the cryptocurrency illegal transactions dataset, comparative experiments are conducted on whether mutual information prior loss function is used and self-supervised learning is used. The performance results demonstrate the traditional methods for dealing with data imbalance (oversampling and undersampling) are not effective in dealing with the cryptocurrency data set. After using the mutual information prior loss function, F1 value is improved by contrast of the original cross-entropy loss function and Focal loss. Compared with without self-supervised learning, the loss function performance is also improved. In summary, the comprehensive comparison shows that taking advantage both of the loss function with self-supervision and mutual information prior can achieve the greatest performance enhancement.

5. Conclusions

To tackle the imbalanced distribution of legal and illegal samples in detecting cryptocurrency violations, this paper proposes a solution that combines mutual information prior and cross-entropy loss functions, resulting in a novel loss function. Compared with traditional cross entropy loss, the novel loss function considering mutual information prior can locate the prior information of the label, and use the prior information to fit

with the output of the network. This greatly alleviates the problem regarding an excessively large number of legal samples and illegal samples in the illegal detection data set of cryptocurrency.Meanwhile, to tackle the issue concerning a substantial quantity of untagged data in the identification of cryptocurrency violations, the solution herein proposed employs a self-supervised method. Compared with the traditional graph neural network, the self-supervised model technology based on mutual information uses the self-supervised method to maximize the mutual information of global variables and local variables to alleviate issues regarding large amounts of unmarked data present in the illegal detection data set of cryptocurrency. This in turn, significantly improves the accuracy of cryptocurrency illegal detection. The use of mutual information prior loss function and the self-supervised method are compared on the elliptic data set. The results show that traditional over-sampling and under-sampling methods are ineffective in dealing with data instability. The F1 score of the proposed mutual information-based loss function has shown varying degrees of improvement compared to previous cross-entropy loss as well as focal loss functions. Additionally, all loss functions demonstrated improvement after the introduction of self-supervised learning. With the proposed loss function relying on self-supervised and mutual information prior, our method exhibits the highest level of improvement. Thus, the proposed loss function incorporating self-supervised and mutual information prior holds significance in detecting cryptocurrency violations.

Author Contributions: Conceptualization, K.Z. and G.D.; Data curation, K.Z.; Formal analysis, K.Z.; Investigation, K.Z.; Methodology, K.Z., G.D. and D.B.; Resources, K.Z.; Supervision, K.Z., G.D. and D.B.; Visualization, G.D. and D.B.; Writing—original draft, K.Z., G.D. and D.B.; Writing—review and editing, K.Z., G.D. and D.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Gai, K.; Qiu, M.; Sun, X. A survey on FinTech. *J. Netw. Comput. Appl.* **2018**, *103*, 262–273. [CrossRef]
2. Kruisbergen, E.W.; Leukfeldt, E.R.; Kleemans, E.R.; Roks, R.A. Money talks money laundering choices of organized crime offenders in a digital age. *J. Crime Justice* **2019**, *42*, 569–581. [CrossRef]
3. Fu, B.; Yu, X.; Feng, T. CT-GCN: A phishing identification model for blockchain cryptocurrency transactions. *Int. J. Inf. Secur.* **2022**, *21*, 1223–1232. [CrossRef]
4. Huang, T.; Lin, D.; Wu, J. Ethereum account classification based on graph convolutional network. *IEEE Trans. Circuits Syst. II Express Briefs* **2022**, *69*, 2528–2532. [CrossRef]
5. Gaihre, A.; Pandey, S.; Liu, H. Deanonymizing cryptocurrency with graph learning: The promises and challenges. In Proceedings of the 2019 IEEE Conference on Communications and Network Security (CNS), Washington, DC, USA, 10–12 June 2019; pp. 1–3.
6. Cui, W.; Gao, C. WTEYE: On-chain wash trade detection and quantification for ERC20 cryptocurrencies. *Blockchain Res. Appl.* **2023**, *4*, 100108. [CrossRef]
7. Ghosh, A.; Gupta, S.; Dua, A.; Kumar, N. Security of Cryptocurrencies in blockchain technology: State-of-art, challenges and future prospects. *J. Netw. Comput. Appl.* **2020**, *163*, 102635. [CrossRef]
8. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
9. Farrugia, S.; Ellul, J.; Azzopardi, G. Detection of illicit accounts over the Ethereum blockchain. *Expert Syst. Appl.* **2020**, *150*, 113318. [CrossRef]
10. Kumar, N.; Singh, A.; Handa, A.; Shukla, S.K. Detecting malicious accounts on the Ethereum blockchain with supervised learning. In Proceedings of the Cyber Security Cryptography and Machine Learning: Fourth International Symposium (CSCML 2020), Be'er Sheva, Israel, 2–3 July 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 94–109.
11. Gu, Z.; Lin, D.; Wu, J. On-chain analysis-based detection of abnormal transaction amount on cryptocurrency exchanges. *Phys. Stat. Mech. Its Appl.* **2022**, *604*, 127799. [CrossRef]

12. Ammer, M.A.; Aldhyani, T.H. Deep Learning Algorithm to Predict Cryptocurrency Fluctuation Prices: Increasing Investment Awareness. *Electronics* **2022**, *11*, 2349. [CrossRef]
13. Akcora, C.G.; Li, Y.; Gel, Y.R.; Kantarcioglu, M. Bitcoinheist: Topological data analysis for ransomware detection on the bitcoin blockchain. *arXiv* **2019**, arXiv:1906.07852.
14. Chen, W.; Guo, X.; Chen, Z.; Zheng, Z.; Lu, Y. Phishing Scam Detection on Ethereum: Towards Financial Security for Blockchain Ecosystem. In Proceedings of the IJCAI, Yokohama, Japan, 11–17 July 2020; Volume 7, pp. 4456–4462.
15. Gai, K.; Guo, J.; Zhu, L.; Yu, S. Blockchain meets cloud computing: A survey. *IEEE Commun. Surv. Tutorials* **2020**, *22*, 2009–2030. [CrossRef]
16. Weber, M.; Domeniconi, G.; Chen, J.; Weidele, D.K.I.; Bellei, C.; Robinson, T.; Leiserson, C.E. Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. *arXiv* **2019**, arXiv:1908.02591.
17. Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; Tang, J. Self-supervised learning: Generative or contrastive. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 857–876. [CrossRef]
18. Hassani, K.; Khasahmadi, A.H. Contrastive multi-view representation learning on graphs. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 4116–4126.
19. Qiu, J.; Chen, Q.; Dong, Y.; Zhang, J.; Yang, H.; Ding, M.; Wang, K.; Tang, J. Gcc: Graph contrastive coding for graph neural network pre-training. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual, 6–10 July 2020; pp. 1150–1160.
20. Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1567–1578.
21. Buda, M.; Maki, A.; Mazurowski, M.A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **2018**, *106*, 249–259. [CrossRef] [PubMed]
22. Gai, K.; Wu, Y.; Zhu, L.; Zhang, Z.; Qiu, M. Differential privacy-based blockchain for industrial internet-of-things. *IEEE Trans. Ind. Inform.* **2019**, *16*, 4156–4165. [CrossRef]
23. Misra, I.; Maaten, L.v.d. Self-supervised learning of pretext-invariant representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6707–6717.
24. Hendrycks, D.; Mazeika, M.; Kadavath, S.; Song, D. Using self-supervised learning can improve model robustness and uncertainty. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 15663–15674.
25. Nowozin, S.; Cseke, B.; Tomioka, R. F-Gan: Training generative neural samplers using variational divergence minimization. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 271–279.