

Data Fusion Approach for Collaborative Anomaly Intrusion Detection in Blockchain-based Systems

Wei Liang, Lijun Xiao, Ke Zhang, Mingdong Tang, Dacheng He, and Kuan-Ching Li *Senior Member, IEEE*

Abstract—Blockchain technology is rapidly changing the transaction behavior and efficiency of businesses in recent years. Data privacy and system reliability are critical issues that need to be addressed in the Blockchain environment. However, anomaly intrusion poses a significant threat to a Blockchain. Therefore, this article proposes a collaborative clustering-characteristic-based data fusion approach for intrusion detection in a Blockchain-based system. A mathematical model of data fusion is designed, and an AI model is used to train and analyze the data clusters in Blockchain networks. The abnormal characteristics in a Blockchain dataset are identified, a weighted combination is carried out, and the weighted coefficients among several nodes are obtained after multiple rounds of mutual competition among clustering nodes. When the weighted coefficient and a similarity matching relationship follow a standard pattern, an abnormal intrusion behavior is accurately and collaboratively detected. Experimental results show that the proposed algorithm has high recognition accuracy and promising performance in the real-time detection of attacks in a Blockchain.

Index Terms—Blockchain, intrusion detection, weighted combination, data fusion, similarity matching

I. INTRODUCTION

THE Blockchain network user community has witnessed a rapid exponential growth along with the development of Blockchain technology. Therefore, ensuring the security of Blockchain networks has become imperative[1][2]. A Blockchain is a point-to-point distributed ledger based on cryptography and a network-sharing system characterized by its disintermediation, transparency, and openness[3]. The security issue caused by the trust-based centralization model adopted by this technology needs to be addressed beforehand.

This work is supported by the National Natural Science Foundation of China under Grants 62072170, 61976087, 61872130, and 61872138, the Fundamental Research Funds for the Central Universities under Grant 531118010527, the Key Research and Development Project of China Hunan Science and Technology Department under Grant 2020GK2006, and 2020SK2066, and the Open Research Fund of Hunan Provincial Key Laboratory of Network Investigational Technology under Grant 2020WLZC001.

W. Liang is with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China, and also with Hunan Provincial Key Laboratory of Blockchain Infrastructure and Application, Changsha 410082, China (e-mail:weiliang99@hnu.edu.cn)

L. Xiao is with Big Data Development and Research Center, Guangzhou College of Technology and Business, Guangzhou 528138, China

K. Zhang is with the Department of School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

M. Tang is with School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou 510006, China

D. He is with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China

K.-C. Li is with the Department of Computer Science and Information Engineering, Providence University, Taichung 43301, Taiwan (e-mail:kuancli@pu.edu.tw) (Corresponding Author)

Transaction signatures, consensus algorithm, and cross-chain technology are utilized to ensure consistency in the distributed ledger of each transaction party, to achieve an automatic information disclosure, and to realize the principle of “account agrees with a document, account agrees with another account, and account agrees with the physical inventory.” In this way, the credit rating of a tradable product can be significantly improved, and its cost can be sharply reduced. Information users can obtain global information on company operations in real time, and their global access to such data signifies the large-scale growth of information. In this case, storing and extracting the value of information are critical. However, anomaly intrusion in Blockchain significantly threatens such information’s security and privacy; therefore, secure intrusion detection approaches should be developed.

Security detection technologies for Blockchain data have been widely used in various AI fields[4]. Nevertheless, the global financial system is exposed to security threats that may result in massive losses. For instance, some vulnerabilities have been detected in the function call of the smart contract in DAO, a crowdfunding project run by an Ethereum-based decentralized organization where 3,641,694 Ethereum coins (approximately \$7.9 million USD) were transferred to private accounts in 2016[5].

The currently available security technologies, such as identity authentication[6] [7], resource protection[8][9][10][11], and machine learning[12] can effectively address the security issues in Blockchain. The tamper-proofing environment of a Blockchain network requires a joint verification among all anonymous participants in any digital capital transaction. Many encryption algorithms are also utilized in Blockchain systems, and the transaction data in these systems are linked together to make the records traceable and unchangeable. Fig.1 shows the data transaction chart of consortium Blockchain, which has been used in various fields, including finance, traffic, and communication, to identify the normal and abnormal behavior of users. In these Blockchain-based applications, malicious third-party can invade the systems for their purposes.

Nevertheless, illegal attacks can use deception to terminate the transmission of data in high-frequency data transactions. Specifically, the miner’s calculation ability is required after data consensus for huge rewards, and greedy miners always attempt to enhance their mining ability through the system. In other words, many security vulnerabilities aim to improve miners’ calculation ability and increase their profit. With the increasing number of information leakage and security events over the past years, developing a secure way for third parties to collect and control a massive amount of private data has

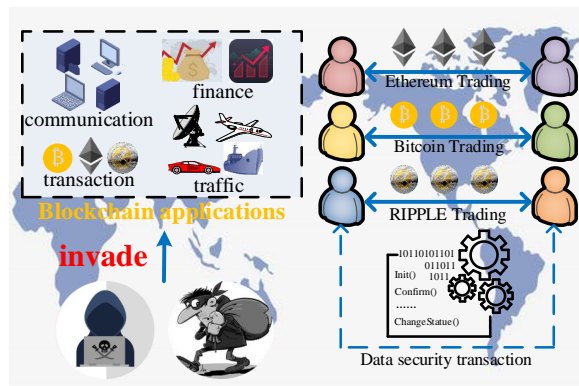


Fig. 1: Data transaction chart of consortium Blockchains

become imperative.

The available intrusion detection methods for Blockchain mainly consider data characteristic attributes, data characteristic models, and joint detection. Attribute-based detection usually creates a standard network behavior model and determines whether the current behavior accords with the standard model. An intrusion alarm is sent if the difference between these models exceeds a certain threshold. Despite being able to detect many new attacks, this method has a high false-alarm probability and cannot detect those intrusion attacks that pretend to be normal. Meanwhile, data characteristic models store the characteristics of known intrusion attacks in a database for misuse detection, which can lead to high detection accuracy and short response time. However, this method is unable to detect unknown intrusion attacks. The intrusion library should also be updated in real-time to ensure detection ability. Joint intrusion detection combines the advantages of misuse and anomaly detection and achieves a fairly accurate result[13]. Anomaly detection can recognize unknown attacks with a higher false alarm rate compared with misuse detection.

The analysis of these Blockchain-based intrusion detection technologies[14] reveals the following challenges:

- 1) The redundant transaction information of intrusion detection should be reduced to minimize the cost of decentralization. The optimal recognition of data clusters with high security and decentralization is selected in the Blockchain. Given that the Blockchain always requires a highly secure and energy-efficient data transaction verification at the cost of resources, obtaining the minimum number of clusters and consensus costs in intrusion detection presents a challenge.
- 2) The intrusion detection algorithm has limited storage and cannot accurately recognize the clusters of a dataset in any case. Therefore, this algorithm cannot identify the clusters of a dataset with low-frequency fusion characteristics, thereby affecting cross-chain technologies' security for intrusion detection.
- 3) Traditional network anomaly detection methods have low detection accuracy and speed. A high-speed, accurate anomaly detection is necessary to improve the real-time management of a Blockchain network.

This article proposes a data-fusion-based collaborative in-

trusion detection scheme to address data privacy and reliability issues in Blockchain-based systems. The data fusion characteristic is utilized for training the datasets in Blockchain-based systems.

The remaining of this article is organized as follows. Section II presents the related work, Section III introduces the matching detection model, and the data fusion approach for collaborative anomaly intrusion detection in Section IV. Section V evaluates the performance of the proposed algorithm, and finally, Section VI summarizes the study and presents directions for future work.

II. RELATED WORK

Research on intrusion detection technologies for Blockchain-based systems remains in its infancy. Characteristic behavior analysis is an essential component of security detection given that the frequency and scale of data transactions are critical to the security of a Blockchain network. Fig. 2 shows a high-frequency intrusion detection model for a Blockchain network and reveals that digital-characteristic-attribute-based intrusion detection has poor data transmission precision and real-time performance.

Given that the available intrusion detection technologies for Blockchain networks incur high detection costs for large networks, continuously measuring the entire network's performance incurs high communication costs and usually has poor timeliness. Previous studies have mainly focused on data consensus, completeness, privacy protection, and scalability and have, accordingly, proposed data consensus algorithms for Blockchain systems. However, the existing anomaly detection algorithms incur high calculation costs.

Large-scale network anomaly detection methods based on lightweight metric restoration have also been developed[15]. The singular value decomposition result of the last iteration is used to reduce the calculation cost for the current iteration. This approach realizes fast anomaly detection and is deemed more suitable than traditional anomaly detection methods for processing data in large-scale networks. Experiments show that the proposed algorithm can precisely detect the location of an anomaly in a Blockchain system and significantly reduce the calculation cost. The schemes presented in [16][17] combine deep reinforcement learning, and the authors propose a content caching technology based on Blockchain authorization to maximize system efficiency.

Several clustering methods [18][19][20][21] have also been developed in the past. In Blockchain data fusion, different algorithms are used to generate characteristic information for a dataset. These characteristics may not be repeated and can be used to match the clustering value of various fusion algorithms to obtain better clustering results[22]. For instance, the generation, exchange, and storage of private data in different devices in a Blockchain can be secured via the P2P feature of this Blockchain. Several operations, such as data creation, modification, and deletion, can also be registered and verified in a Blockchain to prevent illegal intrusion behaviors, including data tampering or misappropriation. Secure access control can also be implemented by customizing the Blockchain or

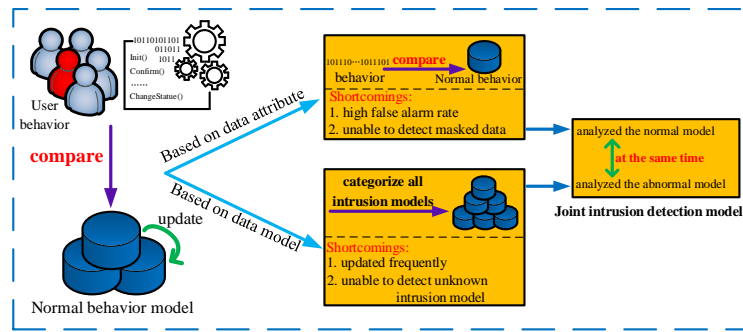


Fig. 2: High-frequent intrusion detection model for Blockchain network

by employing an access mechanism. Fig. 3 shows that in a Blockchain framework, several devices securely store data in different nodes without requiring human intervention, and the features of this Blockchain ensure the decentralization, authenticity, security, and privacy of these data.

The characteristic extraction of clustering data in a Blockchain depends on the amount of quantified information. Different fusion algorithms will generate various clustering characteristics, even for the same dataset. Moreover, a fusion algorithm for the same dataset will generate different clustering characteristic values with varying parameters. In this case, each fusion algorithm is designed based on a certain assumption[23][24][25], thereby limiting the application of these algorithms for the extraction of clustering characteristics in high-frequency transaction environments.

Given that the global distribution pattern of clustering data can be used to recognize the characteristic category of network data, Blockchain clustering-characteristic-based intrusion detection searches for a clustering structure from the sample data without the need for classification identification. Those samples within the same cluster share similar data characteristics, whereas those samples from different clusters show different characteristics. Therefore, the target data can be identified. Some studies have applied clustering-analysis-based intrusion detection. For instance, the clustering method [26] has been used in to connect the data in to a Blockchain network.

The methods mentioned above all assume that a cluster with few samples represents an anomaly. Unsupervised learning is then utilized to learn the normal behavior of a Blockchain network. In [27], a clustering method based on density and grid was used to discover abnormal data, whereas a self-adaptive learning method was used to adapt to a dynamic network change[28]. The authors in [29] proposed a gravitation-based clustering method that applies the concept of gravitation factor to measure the degree of clustering anomaly. To this end, a simple and effective method for calculating the clustering threshold is designed, and a novel network intrusion detection method is developed on the basis of this threshold. Meanwhile, the authors in [30] proposed an unsupervised clustering algorithm without an artificial parameter that is not affected by the order of input data. The shape of the cluster is arbitrary and accurately reflects the distribution of data. After comparing the distance between unlabeled training samples, those samples

with the closest distance are added to the same cluster. After each clustering step, the inter-class distance is re-compared, and the ratio of inner to total data is calculated to determine the anomaly cluster. These approaches can detect intrusion with a certain level of accuracy, but their reliability and efficiency performance warrant further improvements.

To address the above challenges, this research proposes an active intrusion detection approach based on data clustering characteristics. The weighted coefficient of data characteristics and the similarity matching relationship in the normal pattern are considered in this approach. The cluster positions are matched to minimize the cost of detecting clustering data, and the clustering characteristics are accurately recognized. Given the differences in abnormal behaviors and the detection costs of various characteristic data, the proposed intrusion detection approach's reliability and real-time performance can be optimized.

III. MATCHING DETECTION MODEL FOR HIGH-FREQUENT DATA CLUSTERING CHARACTERISTICS

For better evaluation of the user behavior in Blockchain-based system, we need to classify them with the data characteristic. Therefore, we design a matching detection model for high-frequent data clustering characteristic in this section, which can effectively match the characteristic of anomaly intrusion in a Blockchain-based system. Based on this model, a data fusion approach for collaborative anomaly intrusion detection is proposed. The detailed model is illustrated as follows.

In a Blockchain-based environment for large-scale data transaction, the data set for fusion in such a system is denoted by $Z = z_1, z_2, \dots, z_N$, where N is the number of sub-datasets, and M clusters of Blockchain data are generated for Z . There are g fuzzy classes for each Blockchain cluster, denoted by a membership function matrix $U^{(m)} = (u_j^{(m)}(x_i))_{N \times g}$. At this point, $u_j^{(m)}(x_i)$ represents the membership degree that x_i belongs to the j -th class of the m -th cluster $U^{(m)}$, $m = 1, 2, \dots, M$, and the combination cluster is $Q = (q_j(x_i))$. Herein, $q_j(x_i)$ represents the membership degree that x_i belongs to the j -th class of the combination cluster Q .

Definition 1. For any j in the Blockchain, the class at the j -th column of $U^{(m)}$ and that at the j -th column of Q are the

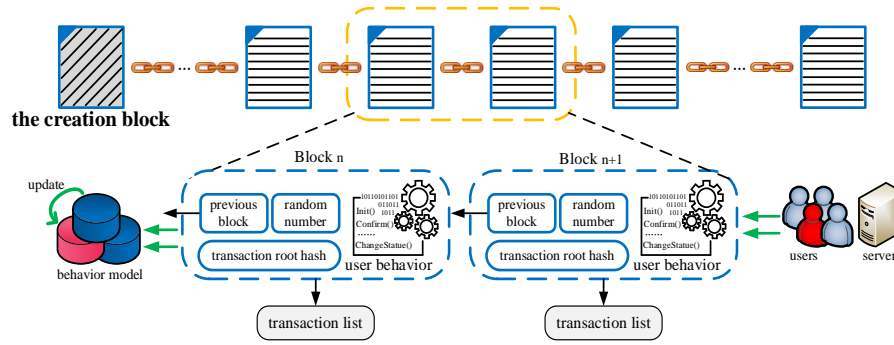


Fig. 3: The chart of Blockchain structure

same. The difference function $h(U^{(m)}, Q)$ of both clusters of $U^{(m)}$ and Q are defined as (1).

$$h(U^{(m)}, Q) \triangleq \min_{\Pi_m} \left(\frac{1}{N} \sum_{i=1}^N \left\| \Pi_m(u_i^{(m)}) - q_i \right\|^2 \right) \quad (1)$$

Herein, $u_i^{(m)}$ and Q are respectively the i -th of $U^{(m)}$ and Q , and Π_m is a column permutation of $U^{(m)}$.

Definition 2. If a combination cluster Q represents M clusters $U^{(m)}$, ($m = 1, 2, \dots, M$). The average value of difference function between Q and M clusters can be transformed into using Q representing the loss of M clusters, denoted by $f(U^{(1)}, U^{(2)}, \dots, U^{(M)}; Q)$.

$$f(U^{(1)}, U^{(2)}, \dots, U^{(M)}; Q) = \frac{1}{M} \sum_{m=1}^M h(U^{(m)}, Q) \quad (2)$$

Based on Definition 2, the combination cluster Q can be calculated through the following optimization model.

$$\min_q f(U^{(1)}, U^{(2)}, \dots, U^{(M)}; Q) = \min_q \left(\frac{1}{M} \sum_{m=1}^M h(U^{(m)}, Q) \right) \quad (3)$$

Formula (1) is substituted into (3), and the model is obtained as (4), which can solve the basic clustering combination model of the Blockchain.

$$\min_q f(U^{(1)}, U^{(2)}, \dots, U^{(M)}; Q) = \min_q \min_{\Pi_1, \Pi_2, \dots, \Pi_M} \left(\frac{1}{M} \sum_{m=1}^M \frac{1}{N} \sum_{i=1}^N \left\| \Pi_m(u_i^{(m)}) - q_i \right\|^2 \right) \quad (4)$$

Formula (5) can be optimized with (4), making (6) be satisfied:

$$\max_{\Pi_1, \Pi_2, \dots, \Pi_M} \frac{1}{N} \sum_{i=1}^N \|q_i\|^2 \quad (5)$$

$$\forall i, q_i = \frac{1}{M} \sum_{m=1}^M \Pi_m(u_i^{(m)}) \quad (6)$$

Data fusion in the Blockchain is to find the data node in the Blockchain network. According to (4), the target function of data fusion algorithm can be set as (7).

$$\min J_{FCM}(U, V; Z) = \sum_{i=1}^c \sum_{j=1}^N u_{ij}^{(m)} \|x_j - v_i\|^2 \quad (7)$$

The constraint condition is $u_{ij} \in [0, 1]$, $\sum_{i=1}^c u_{ij} = 1$, and $0 < \sum_{j=1}^N u_{ij} \leq N$. Herein, U is the membership matrix, V is matrix of cluster center, Z is data collection and c is the number of cluster. Assuming that there are n fusion algorithms C_i ($i = 1, 2, \dots, n$). The effectiveness metric of m clusters is denoted by T_j ($j = 1, 2, \dots, m$), the number of clusters that recognize X by each fusion algorithm (C_i, T_j) is denoted by k_{ij} , and $u_{ij}^{(m)}$ is the element of i -th row and j -th column of membership matrix.

- 1) Calculate the metric value $T_j = iv_j$ ($j = 1, 2, \dots, m$) of each candidate, then combine them as a vector.

$$\alpha = \{iv_1, iv_2, \dots, iv_m\} \quad (8)$$

- 2) The vector is transformed into the vector of equivalent coefficient. Assuming that the area $[\lambda_{min}, \lambda_{max}]$ of system vector has several mutually disjoint sub-intervals E_1, E_2, \dots, E_9 . At this point, $\lambda_{min} = \min\{\alpha\}$, $\lambda_{max} = \max\{\alpha\}$. For $\forall j \in 1, 2, \dots, m$, if $iv_j \in E_p$, the equivalent coefficient will be $r(iv_j) = q$. In this case, an equivalent coefficient vector is generated as follows to indicate each metric.

$$\beta = (r(iv_1), r(iv_2), \dots, r(iv_m)) \quad (9)$$

- 3) The equivalent coefficient vector is applied to construct a comparative matrix $B = (b_{st})_{m \times m}$, so thus, we have

$$b_{st} = r(iv_t)/r(iv_s) \quad (10)$$

- 4) The local prior equivalent coefficient Q_j of each candidate is calculated as

$$\min \sum_{s=1}^m \sum_{t=1}^m (Q_s - b_{st} Q_t)^2 \quad \text{s.t.} \sum_{j=1}^m Q_j = 1 \quad (11)$$

- 5) The global prior equivalent coefficient can be used in clustering level analysis. The weighted least squares

are used to solve the local prior equivalent coefficient $TQ_{ij}(t)(j = 1, 2, \dots, m)$ as (9). The weighted coefficient of each cluster number is updated by

$$Tw_j(t+1) = \sum_{i=1}^n Aw_i(t) TQ_{ij}(t) \quad (12)$$

Hereupon, Tw_j represents the weighted coefficient of j -th cluster number T_j , and Cw_i is the weighted coefficient of i -th fusion algorithm C_i . The error absolute value of each fusion algorithm is calculated when j increases from 1 to m , so the error absolute vector is constructed.

$$e_{I_j} = \{|k_{1j} - K(t)|, |k_{2j} - K(t)|, \dots, |k_{nj} - K(t)|\} \quad (13)$$

At this point, (14) is obtained.

$$K(t) = \sum_{i=1}^n \sum_{j=1}^m Cw_i Iw_j k_{ij} \quad (14)$$

- 6) The local prior equivalent coefficient $CQ_{ij}(t)(i = 1, 2, \dots, n)$ of each distributed fusion algorithm is calculated, as also the distributed weighted coefficient of (C_i, T_j) updated. The cluster detection time can be calculated as follows.

$$w_{ij}(t+1) = Cw_i(t+1) Tw_j(t+1) \quad (15)$$

Here, $i = 1, 2, \dots, n; j = 1, 2, \dots, m$. The weighted sum K of $k_{ij}(i = 1, 2, \dots, n; j = 1, 2, \dots, m)$ is updated.

$$K(t+1) = \sum_{i=1}^n \sum_{j=1}^m w_{ij}(t+1) k_{ij} \quad (16)$$

The ending condition $\epsilon(t)$ of iteration is updated.

$$\epsilon(t) = \max\{|K(t+1) - K(t)|\} \quad (17)$$

Finally, the weighted detection method is used to select the optimal cluster number c_{opt} from K_1, K_2 , where K_1 is the maximum natural number less than K and K_2 is the minimum natural number larger than K .

IV. DATA FUSION FOR COLLABORATIVE ANOMALY INTRUSION DETECTION

There are errors from existing clustering characteristic detection algorithms in network data transaction and communication, which decrease the accuracy and real-time performance in intrusion detection, making the accurate number of clusters unavailable. Therefore, it is unstable and unreliable to recognize the clusters with the efficiency of a single cluster. As shown in Fig.4, this work combines several clustering characteristics for mutual recognition among clusters. In this case, different fusion-based intrusion detection algorithms can accurately recognize the abnormal behavior in the Blockchain network.

In the Blockchain network, a data fusion-based intrusion detection algorithm first recognizes the number of clusters. Next, each cluster that is singly recognized by the fusion

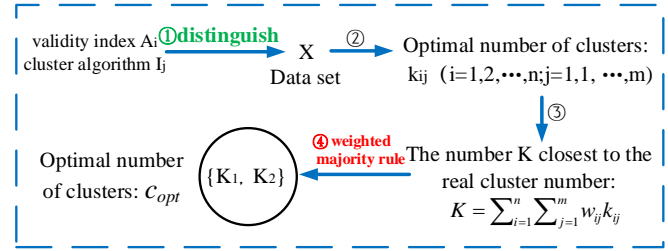


Fig. 4: The characteristic model of data clustering in Blockchain system

algorithm is combined into a value K by weighted sum as (16). Assuming that K is the number of clusters closest to the real value, so two integers closest to K will be selected as the optimal number of clusters using weighted voting. The proposed algorithm is concretely illustrated as follows.

As initial step, m characteristic metrics $T_j(j = 1, 2, \dots, m)$ and n fusion algorithms $C_i(i = 1, 2, \dots, n)$ of AI clusters are selected, and construct next mn pairs of effectiveness metric and fusion algorithm (C_i, T_j) . Given a set $Z = (z_1, z_2, \dots, z_N), z_i \in R^s$, fusion algorithm (C_i, T_j) can automatically recognize the cluster number k_{ij} of Z . In this work, a new collaborative intrusion detection method is proposed based on the Blockchain, and it recognizes the optimal number of clusters for a data set with the multiple-pair fusion algorithm through the following steps:

- 1) Clustering algorithm (C_i, T_j) uniquely recognizes the optimal number of clusters $k_{ij}(i = 1, 2, \dots, n; j = 1, 2, \dots, m)$ of Z .
- 2) k_{ij} is combined as a cluster number K closest to the real value via (16). Here, the weighted coefficient w_{ij} is calculated by (15), Cw_i represents the weighted coefficient of C_i , and Iw_j is the weighted coefficient of T_j .
- 3) The weighted voting method is used to select the optimal cluster number c_{opt} from K_1, K_2 , with K_1 is the maximum natural number less than K and K_2 the minimum natural number larger than K . As depicted in Fig.5 the concrete detection procedure.

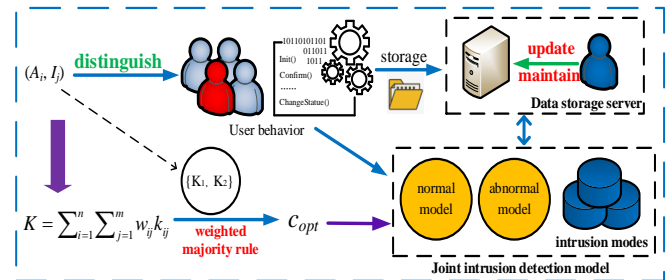


Fig. 5: Clustering characteristic based detection procedure

To avoid repeated calculation on the number of clusters, the characteristic selection algorithm is described in algorithm 1. With the fusion algorithm (C_i, T_j) , the optimal cluster number k_{ij} can combine as number K closest to the real

TABLE I: The symbols of Blockchain parameters

Symbol	Notation
Z	Data set
C_i	The i -th fusion algorithm
T_j	Effectiveness metric of the j -th cluster
(C_i, T_j)	Effectiveness combining C_i and T_j
$TV_{ij}(c)$	The value of effectiveness metric for the j -th class of i -th cluster
k_{ij}	The number of clusters recognized by (C_i, T_j)
TQ_{ij}	Priority coefficient of the j -th cluster given by i -th fusion algorithm
CQ_{ij}	Priority coefficient of the i -th fusion algorithm given by the j -th cluster
Tw_j	The weight coefficient of effectiveness metric T_j for j -th cluster
Cw_i	The weight coefficient of i -th fusion algorithm C_i
$vn(c)$	The weighted voting number of the cluster number c
G	Attack strength
c_{opt}	The optimal number of clusters

Algorithm 1 Calculation algorithm on the number of optimal clusters.

Input:

Clustering parameter C_i , fusion algorithm (C_i, T_j) for Z ;

Output:

Optimal clustering metric $TV_{ij}(c)$;

```

1: for  $i = 1$  to  $n$  do
2:   for  $c = 2$  to  $c_{max}$  do
3:     Generate a cluster  $U^{(i)}(c)$  with  $c$  classes for data set
        $X$  with fusion algorithm  $A_i$  ;
4:   for  $j = 1$  to  $m$  do
5:     Calculate the value  $TV_{ij}(c)$  of effectiveness metric
        $T_j$  of the cluster  $U^{(i)}(c)$  ;
6:   end for
7: end for
8: end for
9: return  $TV_{ij}(c)$  .

```

cluster number, which procedure can effectively recognize the clustering characteristic behavior of each fusion algorithm (C_i, T_j) . The principle of characteristic recognition is to make the weighted sum of the difference between the minimum k_{ij} and K , as calculated by (18).

$$\min_{w_{ij}} \left\{ \sum_{i=1}^n \sum_{j=1}^m w_{ij} (k_{ij} - K)^2 \right\} \quad (18)$$

$$s.t. \quad \sum_{i=1}^n \sum_{j=1}^m w_{ij} = 1$$

It is proposed in this work a design of a clustering characteristic matching algorithm for Blockchain-based system and solves an approximate solution of w_{ij} . With $w_{ij} = Cw_i Tw_j$, the solution of w_{ij} is transformed into solving Cw_i and Tw_j . Herein, Cw_i is the weighted coefficient of $C_i (i = 1, 2, \dots, n)$ and Tw_j is the weighted coefficient of $T_j (j = 1, 2, \dots, m)$, and Cw_i, Tw_j and K are initialized. Each fusion algorithm C_i is the candidate, $e_{I_j} = |k_{ij} - K| (j = 1, 2, \dots, m)$ the evaluation metric, so the metric value for each C_i is $|k_{ij} - K| (i = 1, 2, \dots, n)$. The hierarchical analysis is used to allocate priority coefficient of each fusion algorithm and update $Cw_i (i = 1, 2, \dots, n)$. After that, the effectiveness metric $T_j (j = 1, 2, \dots, m)$ of each data cluster is regarded as the candidate and $e_{A_i} = |k_{ij} - K| (i = 1, 2, \dots, n)$ as the evaluation metric. The value of T_j is $|k_{ij} - K| (j = 1, 2, \dots, m)$, so the effectiveness metric $Tw_j (j = 1, 2, \dots, m)$ of clustering

characteristic is updated with priority coefficient, and finally, $w_{ij} = Cw_i Tw_j$ is used to update the weighted coefficient characteristic value w_{ij} of the fusion algorithm $(C_i, T_j) (i = 1, 2, \dots, n; j = 1, 2, \dots, m)$. Reversely, the matched w_{ij} is used to update K , iterated and terminated the matching when the absolute value of the difference of w_{ij} and K in two adjacent iterations is less than the preset value, generating the approximate optimal matching solution of w_{ij} . The pseudocode is presented in algorithm 2.

Each data fusion algorithm $(C_i, T_j) (i = 1, 2, \dots, n; j = 1, 2, \dots, m)$ in Blockchain selects an optimal data cluster c_{opt} from K_1 and K_2 by using the weighted voting method, as presented in algorithm 3.

In the selection algorithm on the optimal number of clusters, two clusters are generated for a data set z with K_1 classes and K_2 classes, respectively. The characteristic values of all clusters will be calculated, and after that, each fusion algorithm votes the optimal cluster number. For example, despite most votes agree $K_1 = 2$ to be the optimal cluster number, the weighted vote $vn(K_1)$ for $K_1 = 2$ is less than that for $K_2 = 3$. Based on the principle of the weighted majority vote, $K_2 = 3$ is the optimal cluster number.

In the proposed algorithm, the correct recognized cluster number of data set is calculated by weighted voting of all fusion algorithms, since it can effectively evaluate each data cluster's performance in Blockchain and improve the recognition accuracy of clusters.

V. EXPERIMENTS AND ANALYSIS

A. Experimental Environment

A server composed of one AMD X4 630 processor and 64G memory is used to compare the performance of the proposed algorithm against other algorithms. The smart contract utilizes Hyperledger Fabric as the framework, Golang as the development language, and KDD CUP1999 [31] as the training dataset. To analyze and evaluate the effectiveness of the proposed algorithm, four datasets are selected from the databases of machine learning and support vector machine for testing[32], namely, *the iris*, *lymphography*, *vehicle*, and *glass* datasets. Each of these datasets has different scales of clusters, and each of their records shows a real pattern of network connections with 41 parameters that can be marked as either an attack or a normal connection. A complete

Algorithm 2 Clustering characteristic-based intrusion detection algorithm.

Input:
Weighted coefficient $w_i (i = 1, 2, \dots, n)$;

Output:
Matching matrix B ;

- 1: **for** $i = 1$ to n **do**
- 2: Evaluate the objects with the i -th block of Blockchain via (8)-(11);
- 3: Assign local priority coefficient $Q_{ij} (j = 1, 2, \dots, n)$;
- 4: **end for**
- 5: Calculate the global priority coefficient $GQ_j (j = 1, 2, \dots, n)$ with Q_{ij} ;
- 6: Initialize $Cw_i(t) = 1/n, Tw_j(t) = 1/m$;
- 7: **repeat**
- 8: Calculate $K(t) = \sum_{i=1}^n \sum_{j=1}^m Cw_i Tw_j k_{ij}$;
- 9: update $K(t), w_{ij}(t)$;
- 10: **until** $(\epsilon(t) = 1)$
- 11: **while** $\epsilon(t) \geq 0.0001$ **do**
- 12: **for** $i = 1$ to n **do**
- 13: Calculate the absolute value of relative errors of the clustering effectiveness metric for each block and construct the error vector $e_{Ai} = |k_{i1} - K(t)|, |k_{i2} - K(t)|, \dots, |k_{im} - K(t)|$;
- 14: **if** $|k_{ij} - K| \in E_k^{(C_i)}$ **then**
- 15: Assign a priority number $r_j^{(C_i)} = k$;
- 16: Generate priority coefficient vector $r^{(C_i)} = (r_1^{(C_i)}, r_2^{(C_i)}, \dots, r_m^{(C_i)})$;
- 17: **end if**
- 18: Construct matching matrix $B = (b_{st})_{m \times n}, b_{st} = r_t^{(C_i)} / r_s^{(C_i)}$;
- 19: **end for**
- 20: **end while**
- 21: **return** B .

dataset includes almost 5 million records, with each record representing a connection comprising 41 characteristics and can be marked as either normal or abnormal. Therefore, the intrusion attack records of all normal connections are included in these datasets.

The detection data set is corrected testing data set and training data set. To balance the calculation complexity and the selected fusion algorithm, more similar fusion algorithms will increase the calculation complexity in detection. The recognition accuracy will also be affected. Accordingly, the fusion algorithm with different characteristics is used to recognize the dataset clusters.

B. Performance Evaluation

The performance of an intrusion detection method in a Blockchain network is evaluated based on its recognition accuracy, detection efficiency, and reliability. The evaluation results are then compared with those of DLANID[33], CID[34] and CBSigCID[35].

1) *Recognition Accuracy*: During the extraction, most of the Blockchain data clustering characteristics cannot be accu-

Algorithm 3 Selection algorithm of the optimal number of cluster c_{opt} .

Input:
Clustering algorithm $(C_i, T_j), K_1, K_2$;

Output:
Optimal number of cluster c_{opt} ;

- 1: Generate two clusters respectively with K_1 and K_2 classes;
- 2: **for** $i = 1$ to n **do**
- 3: Set the parameters of fusion algorithm C_i expect the number of clusters;
- 4: **end for**
- 5: **for** $i = 1$ to n **do**
- 6: **for** $j = 1$ to m **do**
- 7: Record the vote number of K_1 and K_2 from (C_i, T_j) as $\delta_{ij}(K_1)$;
- 8: **if** $IV_{ij}(K_1) > IV_{ij}(K_2)$ **then**
- 9: $\delta_{ij}(K_1) = 1$;
- 10: **else**
- 11: $\delta_{ij}(K_1) = 0$;
- 12: **end if**
- 13: **end for**
- 14: **end for**
- 15: Calculate weighted vote number
 $vn(K_1) = \sum_{i=1}^n \sum_{j=1}^m w_{ij} \delta_{ij}(K_1)$,
 $vn(K_2) = \sum_{i=1}^n \sum_{j=1}^m w_{ij} \delta_{ij}(K_2)$;
- 16: Calculate $c_{opt} = \arg \max_{k \in \{K_1, K_2\}} \{vn(k)\}$;
- 17: **return** c_{opt} .

TABLE II: Experimental results of clustering characteristic

Data set	C_{real}	R	K	C_{opt}	TP	FP
<i>lymphgraphy</i>	12	99%	4.951	4	93%	2.2%
<i>vehicle</i>	18	97%	4.892	18	97%	1.3%
<i>glass</i>	28	98%	5.424	28	96%	1.2%

rately recognized because each dataset has a different recognition accuracy. Table II presents the experimental results, where C_{real} is the real number of clusters, R is the recognition accuracy, and K is the optimal number of cluster characteristics. These experimental results demonstrate that the fusion algorithms show significant differences in their data recognition accuracy. The best recognition accuracy is reported in the *lymphgraphy* dataset.

Table III lists the clusters recognized for the vehicle dataset. A total of 18 clustering characteristics are considered in a clustering characteristic matching algorithm. As depicted in Table III, only the cluster numbers of (FCM, F_Sil), (FCM, FS), and (PFCM, F_Sil) are accurately recognized, while all others occur error in evaluation. The recognition results show significant deviations due to the employed characteristic matching algorithm.

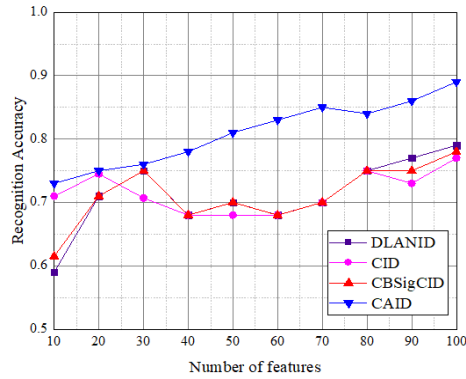
A total of 28 clusters are present in the *glass* dataset, and the evaluation results are listed in Table IV. Both (FCM, SVI) and (FCM, BS) are recognized, and the number of clusters is closest to the real value. The effectiveness metric and recognized cluster number are below the real value of 28. (AFCM, fpbm) recognizes the cluster number 32 for the glass

TABLE III: Recognition of clusters for data set *vehicle*

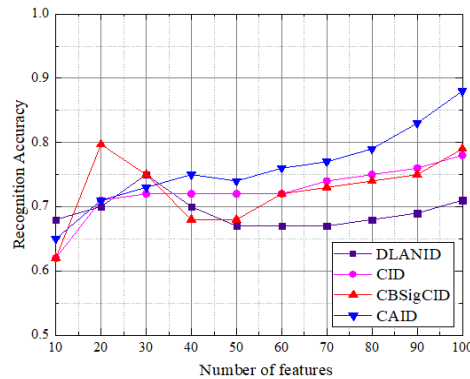
	FS	V_t	SVI	V_w	V_{gd}	V_{sc}	f_{pbm}	SCG	BS	F_{Sil}
AFCM	6	4	6	2	3	2	5	3	4	18
FCM	18	12	17	2	2	2	5	4	17	28
PFCM	6	4	3	2	3	2	8	3	4	18
PCA	22	4	3	3	3	3	4	3	32	11

dataset, and this number exceeds the real value. In sum, the proposed algorithm can accurately recognize the clustering characteristic of the glass dataset.

The proposed algorithm also shows high recognition accuracy, low false alarm rate, and promising efficiency. Given that both the *lymphography* and *glass* datasets contain high-dimensional data, they should be transformed into low-dimensional data for the characteristic detection to reduce the storage space, simplify learning in the intrusion detection model, and subsequently improve accuracy.



(a) Comparison of recognition accuracy for data set *lymphography*



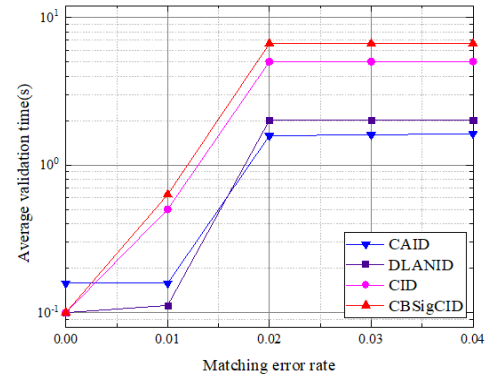
(b) Comparison of recognition accuracy for data set *glass*

Fig. 6: Comparison of recognition accuracy for various algorithms

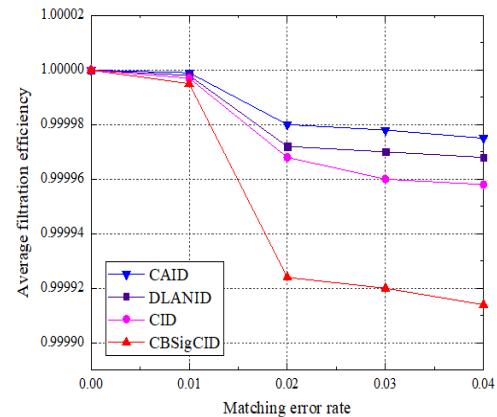
Figs. 6(a) and 6(b) show the recognition accuracy curves of three algorithms for the *lymphography* and *glass* datasets, respectively. In these curves, the X and Y axes show the number of clustering characteristics and recognition accuracy, respectively. The proposed algorithm accurately matches the number of clustering characteristics to the neighboring structure of the Blockchain with the *lymphography* and *glass*

datasets. The recognition accuracy curve of this algorithm is smoother than those of DLANID, CID, and CBSigCID, thereby suggesting that this algorithm has higher recognition accuracy compared with the other methods. This result can be ascribed to the performance deviations observed among different algorithms for the same sample, which can disturb the network anomaly detection. Therefore, the proposed algorithm is highly suitable for Blockchain data characteristic detection.

2) *Detection Efficiency*: The detection time of the proposed algorithm in a Blockchain-based environment is compared with that of DLANID, CID, and CBSigCID. Matching error rate and averaging filtration efficiency are used as evaluation metrics, and the results are shown in Fig. 7. All algorithms show similar curves, with the proposed algorithm having low average matching error rate and filtration efficiency.



(a) Comparison of average validation time



(b) Comparison of average filtration efficiency

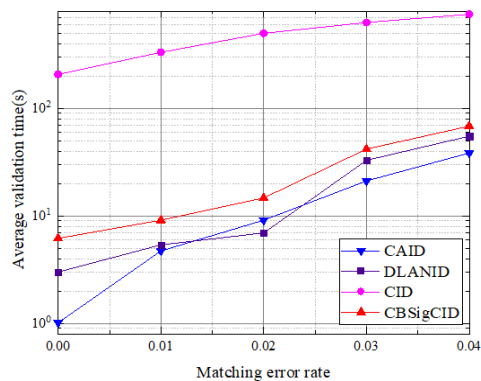
Fig. 7: Comparison of detection efficiency for various algorithms

The real-time detection performance of the proposed algorithm on the *lymphography* and *glass* datasets is evaluated based on its average validation time as shown in Fig. 8. By

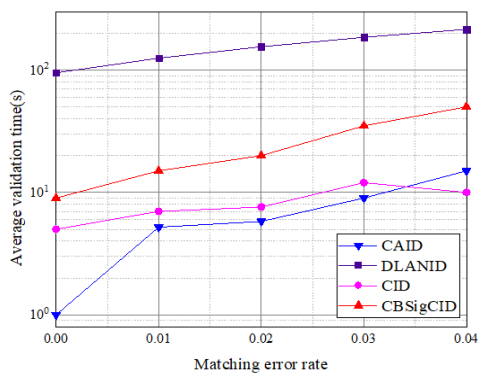
TABLE IV: Recognition of clusters for data set *glass*

	FS	V_t	SVI	V_w	V_{gd}	V_{sc}	f_{pbm}	SCG	BS	F_{Sil}
AFCM	16	14	16	13	13	12	32	13	14	18
FCM	16	12	14	12	12	12	15	14	27	24
PFCM	16	14	27	14	13	12	14	13	14	17
PCA	22	14	13	13	13	13	14	13	23	19

training different Blockchain data, the recognition accuracy of each algorithm in these datasets before and after improvement is checked. Data objects in the training data set to mark the matching error rate, so thus, the average validation time of normal data set after improvement is a lower value, despite that the proposed algorithm is higher.



(a) Result for data set *lymphography*



(b) Result for data set *glass*

Fig. 8: Comparison of average validation time with various matching error rates

Average validation time is classified into normal and abnormal based on a certain threshold value. By using the k value of the distributed data fusion algorithm, the proposed algorithm achieves a better effect than CID and CBSigCID, avoids the artificial error caused by the empirical value, and shows improvements in its accuracy.

3) *Reliability Performance*: If an abnormal transaction information exists in the Blockchain network, then the state is considered abnormal. Blockchain business transactions are periodic procedures that increase the Blockchain number by increasing the number of transaction states. If the data characteristics of a transaction in a Blockchain system are rapidly extracted, then the false alarm rate and recognition accuracy

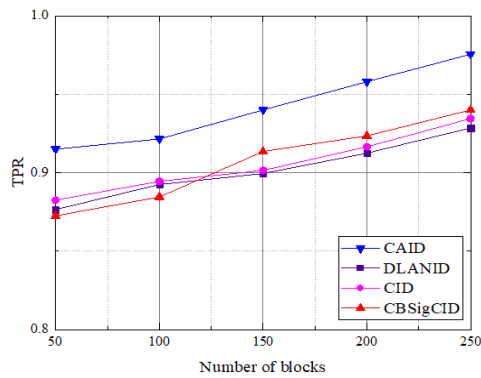
after an attack can be accurately analyzed. On this basis, the security and reliability of the entire Blockchain system can be evaluated.

In a Blockchain data transaction, the transaction states and number of successfully intercepted attacks are used to verify the effectiveness of the proposed algorithm. Experiments are conducted with the control of different attack intensity G , and the security evaluation aims to determine the probability of hiding the clustering characteristics in the Blockchain network and to evaluate the security of the Blockchain. A total of 300 blocks are generated for the experiment, and the attack intensity is set to 0.4 and 0.6. The relationship between detection time and hiding probability is analyzed along with an increasing number of blocks. Figs. 9(a) and 9(b) present the evaluation results for true positive rate (TPR) and false positive rate (FPR) when $G = 0.4$, whereas Figs. 9(c) and 9(d) present the evaluation results for TPR and FPR when $G = 0.6$. Increasing the number of blocks also increases the FPR and TPR. The TPR of the proposed algorithm obviously increases. When the attackers obtain additional resources, both attack density and intensity increase. In this case, the intrusion detection algorithm becomes highly sensitive to abnormal network behavior. Overall, the proposed algorithm outperforms the others in terms of FPR and TPR across all selected testing datasets.

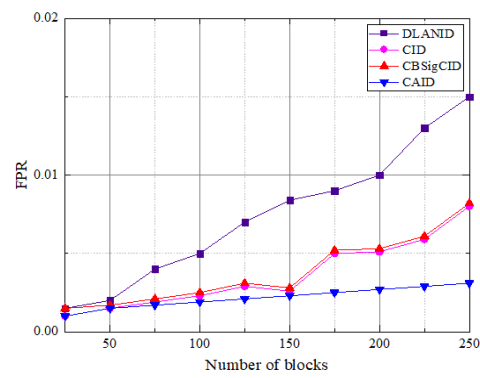
The relationship between attack intensity and average hiding probability is then evaluated, and the result is shown in Fig. 10. In this figure, the X-axis represents the attack intensity (changing from 0-0.9), whereas the Y-axis represents the average hiding probability. As the attack intensity increases, the average hiding probabilities of four comparative algorithms also increase. Therefore, when the attack intensity exceeds 0.6, the average hiding probability decreases. The proposed algorithm is only suitable for anomaly detection when the attack intensity is relatively low. The normal distribution in Fig. 10 shows that the average hiding probability of the abnormal data clustering characteristic reaches its peak when the attack intensity is 0.6. The ability against attacks is increasingly better.

VI. CONCLUSIONS AND FUTURE WORK

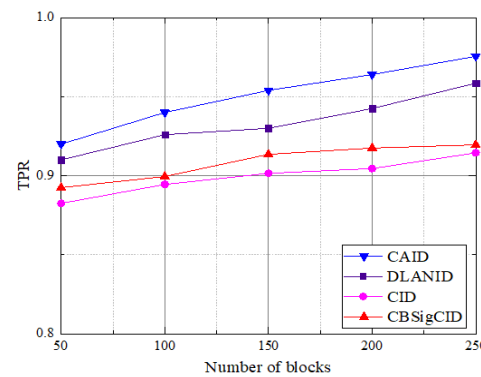
An analysis of the characteristic behavior of clustering data in a Blockchain network reveals that using a single clustering characteristic for anomaly detection is not ideal. To address this security issue, this article proposes a clustering-characteristic-based collaborative intrusion detection algorithm for a Blockchain that can rapidly recognize the clustering characteristics in Blockchain data transactions. A mathematical model is also created, based on which a clustering-characteristic-based intrusion detection protocol is designed to accurately detect the number of clustering characteristics



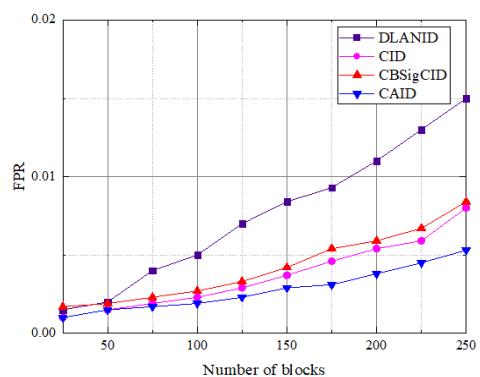
(a) Evaluation of TPR with $G=0.4$



(b) Evaluation of FPR with $G=0.4$



(c) Evaluation of TPR with $G=0.6$



(d) Evaluation of FPR with $G=0.6$

Fig. 9: Evaluation of FPR and TPR for various algorithms

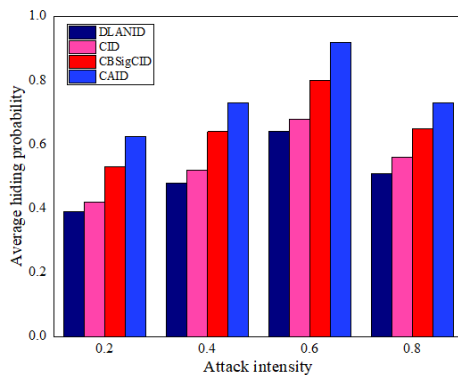


Fig. 10: The relationship between attack intensity and average hiding probability

in an abnormal network. Experimental results show that the proposed algorithm has excellent recognition accuracy and time overhead.

As future work, we plan to further optimize the clustering model in a Blockchain network. Given the variability of network attacks, an anomaly detection algorithm has many application scenes, including detecting anomaly in a data packet and in network traffic. As another direction for future research, Blockchain and encryption technologies may be combined to realize a secure identity authentication of Blockchain data. The controllability of data anonymity and privacy protection in a

Blockchain network can also be considered to improve the real-time performance and security of intrusion detection.

REFERENCES

- [1] I.-C. Lin and T.-C. Liao, "A survey of blockchain security issues and challenges," *IJ Network Security*, vol. 19, no. 5, pp. 653–659, 2017.
- [2] K. Zhang, Y. Zhu, S. Maharjan, and Y. Zhang, "Edge intelligence and blockchain empowered 5g beyond for industrial internet of things," *IEEE Network*, vol. 33, no. 5, pp. 12–19, 2019.
- [3] W. Liang and M. T. et al., "A secure fabric blockchain-based data transmission technique for industrial internet-of-things," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3582–3592, 2019.
- [4] K. Zhang, S. Leng, X. Peng, L. Pan, S. Maharjan, and Y. Zhang, "Artificial intelligence inspired transmission scheduling in cognitive vehicular communications and networks," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1987–1997, 2019.
- [5] C. Jentzsch. (2019) Decentralized autonomous organization to automate governance. [Online]. Available: <https://download.slock.it/public/DAO/WhitePaper.pdf>
- [6] X. Li and J. M. et al., "A novel smart card and dynamic ID based remote user authentication scheme for multi-server environment," *Mathematical and Computer Modelling*, vol. 58, no. 1-2, pp. 85–95, 2013.
- [7] X. Li and J. N. et al., "An enhanced smart card based remote user password authentication scheme," *Journal of Network and Computer Applications*, vol. 36, no. 5, pp. 1365–1371, 2013.
- [8] W. Liang, D. Zhang, X. Lei, M. Tang, K. Li, and A. Zomaya, "Circuit copyright blockchain: Blockchain-based homomorphic encryption for IP circuit protection," *IEEE Transactions on Emerging Topics in Computing*. [Online]. Available: <http://dx.doi.org/10.1109/TETC.2020.2993032>, 2020.
- [9] W. Liang and Y. F. et al., "Secure data storage and recovery in industrial blockchain network environments," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 10, pp. 6543–6552, 2020.

- [10] W. Liang and W. H. et al., "Deep reinforcement learning for resource protection and real-time detection in IoT environment," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6392–6401, 2020.
- [11] D. Han, N. Pan, and et al., "A traceable and revocable ciphertext-policy attribute-based encryption scheme based on privacy protection," *IEEE Transactions on Dependable and Secure Computing*. [Online]. Available: <https://doi.org/10.1109/TDSC.2020.2977646>
- [12] K. Zhang, J. Cao, H. Liu, S. Maharjan, and Y. Zhang, "Deep reinforcement learning for social-aware edge computing and caching in urban informatics," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5467–5477, 2019.
- [13] K. Xie, X. Li, X. Wang, J. Cao, G. Xie, J. Wen, D. Zhang, and Z. Qin, "On-line anomaly detection with high accuracy," *IEEE ACM Transactions on Networking*, vol. 26, no. 3, pp. 1222–1235, 2018.
- [14] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys and Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.
- [15] K. Xie and X. L. et al., "Quick and accurate false data detection in mobile crowd sensing," in *international conference on computer communications*, Paris, France, 2019, pp. 2215–2223.
- [16] K. Zhang, Y. Zhu, S. Leng, Y. He, S. Maharjan, and Y. Zhang, "Deep learning empowered task offloading for mobile edge computing in urban informatics," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7635–7647, 2019.
- [17] H. Ayad and M. S. Kamel, "Cumulative voting consensus method for partitions with variable number of clusters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 1, pp. 160–173, 2008.
- [18] T. Hu and Y. Y. et al., "Maximum likelihood combination of multiple clusterings," *Pattern Recognition Letters*, vol. 27, no. 13, pp. 1457–1464, 2006.
- [19] A. L. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 835–850, 2005.
- [20] Z. He, X. Xu, and S. Deng, "A cluster ensemble method for clustering categorical data," *Information Fusion*, vol. 6, no. 2, pp. 143–151, 2005.
- [21] Y. Dai and G. W. et al., "Conceptual alignment deep neural networks,"
- [34] H. Bowen and Z. C. et al., "A collaborative intrusion detection approach using blockchain for multimicrogrid systems," *IEEE Transactions on Systems Man & Cybernetics Systems*, pp. 1–11, 2019.
- Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 3, pp. 1631–1642, 2018.
- [22] A. Topchy, A. K. Jain, and W. F. Punch, "Clustering ensembles: Models of consensus and weak partitions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1866–1881, 2005.
- [23] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, no. 3, pp. 583–617, 2003.
- [24] E. Dimitriadou, A. Weingessel, and K. Hornik, "A combination scheme for fuzzy clustering," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 16, no. 07, pp. 901–912, 2002.
- [25] G. Yang and X. Y. et al., "An intrusion detection algorithm for sensor network based on normalized cut spectral clustering," *PLoS ONE*, vol. 14, no. 10, p. e0221920, 2019.
- [26] L. Portnoy, "Intrusion detection with unlabeled data using clustering," *Columbia University*, 2000.
- [27] M. Yingchi, Z. Haishi, and Q. Hai, "An adaptive trajectory clustering method based on grid and density in mobile pattern analysis," *Sensors*, vol. 17, no. 9, 2013.
- [28] W. Liang and J. L. et al., "A fast defogging image recognition algorithm based on bilateral hybrid filtering," *ACM Transactions on Multimedia Computing, Communications, and Applications*. [Online]. Available: <https://doi.org/10.1145/3391297>
- [29] B. M. A. Shahri and S. K. e. a. Zadeh, *Comparative Analysis of Gravitational Search Algorithm and K-Means Clustering Algorithm for Intrusion Detection System*, 2013.
- [30] W. Liang, K. Li, J. Long, X. Kui, and A. Zomaya, "An industrial network intrusion detection algorithm based on multi-characteristic data clustering optimization model," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 3, pp. 2063–2071, 2019.
- [31] "KDD cup 1999 data," <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>.
- [32] "ML database," <http://archive.ics.uci.edu/ml/index.php>.
- [33] N. Shone and T. N. N. et al., "A deep learning approach to network intrusion detection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 1, pp. 41–50, 2018.
- [35] W. Li and S. T. et al., "Designing collaborative blockchained signature-based intrusion detection in iot environments," *Future Generation Computer Systems*, vol. 96, 2019.