Hindawi
Journal of Applied Mathematics
Volume 2023, Article ID 8489644, 12 pages
https://doi.org/10.1155/2023/8489644

Research Article

Subspace-Based Anomaly Detection for Large-Scale Campus Network Traffic

Xiaofeng Zhao and Qiubing Wu

Book and Information Center, Anhui University of Finance and Economics, Bengbu 233030, China

Correspondence should be addressed to Qiubing Wu; wuqb@aufe.edu.cn

Received 26 August 2022; Revised 20 June 2023; Accepted 28 August 2023; Published 16 September 2023

Academic Editor: Oluwole D. Makinde

With the continuous development of information technology and the continuous progress of traffic bandwidth, the types and methods of network attacks have become more complex, posing a great threat to the large-scale campus network environment. To solve this problem, a network traffic anomaly detection model based on subspace information entropy flow matrix and a subspace anomaly weight clustering network traffic anomaly detection model combined with density anomaly weight and clustering ideas are proposed. Under the two test sets of public dataset and collected campus network data information of a university, the detection performance of the proposed anomaly detection method is compared with other anomaly detection algorithm models. The results show that the proposed detection model is superior to other models in speed and accuracy under the open dataset. And the two traffic anomaly detection models proposed in the study can well complete the task of network traffic anomaly detection under the large-scale campus network environment.

1. Introduction

With the continuous upgrading of the network traffic bandwidth, the amount of network users and the network information amount increase exponentially. The constantly expanding amount of information makes the content and form of network communication more abundant [1]. Although the expansion of healthy and high-quality online information is adorable, the number of disastrous cyberattacks is also increasing. How to deal with more high-frequency cyber threats has become a problem for everyone nowadays [2]. In particular, campus networks are relatively more frequently attacked; therefore, campus networks need to strengthen the means to cope with the problem of cyber threats after being attacked, in addition to the need for continuous reinforcement at the protection level [3]. The anomaly detection of information carriers in campus network belongs to the anomaly detection of large-scale information carriers. In addition to the accuracy of detection, the efficiency and cost of anomaly detection should also be considered when detecting the campus network [4]. The problem of campus network traffic anomaly detection is a large-scale traffic anomaly detection problem, and the efficiency and cost of anomaly detection should be considered in addition to the accuracy of detection when performing detection on campus networks [5]. In this regard, literature [6] proposed a network intrusion detection model based on wavelet neural network; literature [7] proposed a traffic anomaly detection method combining feature dimensionality reduction; and literature [8] proposed an active intrusion detection framework based on density perception and feature deviation of network traffic flow. However, the existing research has not solved these problems well.

To further improve the detection efficiency and accuracy, the research is based on subspace combined with information entropy traffic matrix and clustering algorithm to explore large-scale campus network traffic anomaly detection. Combining density anomaly weight and clustering idea, a network traffic detection model based on subspace anomaly weight clustering is proposed to distinguish normal and abnormal network traffic. We have achieved network traffic anomaly detection in a large-scale campus network environment and extended this method to anomaly detection in other large-scale information carriers, which has

certain promotion and application value. In practical applications, this model and its extensions can be used in anomaly detection tasks of large-scale information carriers such as campus networks to improve the level of network security and ensure the safe and reliable dissemination of network information. In addition, this study also has a certain driving effect on in-depth research in the fields of subspace decomposition, information entropy calculation, and clustering algorithms.

2. Related Work

To help maintain the safe and stable operation of the campus network, the study of anomaly detection strategies for large-scale campus network traffic has become a focal topic nowadays. This topic has been thoroughly studied by researchers at home and abroad. Amoozegar et al. conceived an anomaly tracker in online subspace. The tracker can perform more accurate anomaly detection for data with relatively low rank and relatively sparse components and can better deal with sudden changes in subspace. Experimental evaluation using datasets with different sparsities shows that the accuracy of this method is better than others [9]. Zhong et al. conceived a new preprocessing method using subspace similarity detection to solve the problem of low quality of data mining in the Internet of Things. This method can flexibly handle sporadic sensor data of similar types in nature, which exists in many environmental sensing applications. Through evaluation experiments, it is proved that this method has high accuracy [10]. Guo et al. conceived a deep learning method for detecting network carrier information abnormalities to enhance the security of data in network transmission. This method combines multiple convolutional automatic encoders with long-term and short-term memory networks to reduce the feature loss. The loss values in CAE (convolutional automatic encoder) and SAE (stacked automatic encoder) training in the experiment are compared, and the results are better than other studies [11]. Song et al. constructed a Hurst (H) parameter for evaluating network traffic to detect network information carrier anomalies, formed a set of information features on this basis, formalized the normal and abnormal behaviors of the system, and conducted static analysis based on the flow. The experimental results show that using this algorithm in the IOT communication infrastructure can significantly improve the level of information security and reduce the risk of information loss [12]. Zhong et al. organically combined a variety of deep learning neural networks to build a brand-new anomaly information detection framework. The framework trains LSTM with data with anomaly score tags and obtains the final anomaly score by weighting. The experimental results show that the framework has stronger adaptability and accuracy in anomaly information detection than other detection frameworks [13].

Zhang et al. have constructed a solution for anomaly detection of power grid information carrier. This method originates from the multilayer echo state network (ML-ESN). It uses Pearson and Gini coefficient methods to calculate the statistical distribution and correlation of network traffic characteristics first, and then ML-ESN is used to classify network anomaly attacks. Experiments are conducted on a public network security dataset, which shows that the accuracy of this method is extremely high, significantly higher than other methods [14]. Peng et al. conceived a network information carrier anomaly detection method, which relies on a Mahout classifier. The experimental results show that this method performs well in network information carrier anomaly detection and has higher accuracy and adaptability than other existing algorithms [15]. Wu et al. considered a hybrid feature selection method to improve the speed and accuracy of feature selection in anomaly detection of network information carriers. This method converts the interaction information between features into the redundancy of features and uses the method of deleting before and after random to test the influence of features on detection. The experimental results show that this method can quickly screen out feature subsets with better detectability and lower dimension [16]. Wang et al. have considered a method to solve the problem of dynamic variability of training model and deployment environment. This method converts network carrier information into numerical information and projects it into different dimensional spaces to form detection vectors. This method is capable of processing high-dimensional data and has strong generalization ability which makes anomaly detection easier. This method can effectively reduce the training cost, complete the detection task well, and has a high accuracy [17]. Aliguliyev et al. proposed a more accurate multiclassifier model for anomaly detection of network information carriers based on big data. Weka was used to conduct experiments on NSL-KDD datasets. The experimental results show that the model performs very well in anomaly detection accuracy [18]. Khatibzadeh et al. considered a mutation theory-based network information carrier anomaly detection model, which can effectively describe the process of abrupt changes in the network due to the dynamic nature of clouds. This model was compared with the general mutation theory model, and the results verify that the proposed method model has great advantages in accuracy and detection rate [19].

Through the in-depth research of many domestic and foreign researchers on the anomaly detection of network information carriers, it can be seen that the research on anomaly detection of large-scale campus network information carriers can be carried out using subspace method combined with traffic matrix and clustering algorithm. Therefore, the study proposes a large-scale campus network traffic anomaly detection method based on subspace, which has a certain role in promoting the research of network traffic anomaly check.

This study used three different datasets for testing and validation. When selecting and adjusting datasets, the following standards were followed: firstly, the dataset must contain normal network traffic and different types of network attacks; secondly, the size of the dataset should be large enough to cover different environments and scenarios; and the final dataset must have widespread use and benchmark value. The final selected datasets for the study were NSL-KDD, CSE-IC-IDS2018, and CIC-DDoS2019. The NSL-KDD dataset is one of the widely used datasets in the field of network

intrusion detection. It is an improved version of the KDD Cup 1999 dataset, which includes four different types of attacks and normal network traffic. This dataset is used to evaluate the effectiveness of different network traffic anomaly detection methods. The CSE-IC-IDS2018 dataset contains different types of network traffic anomalies and attacks, such as DDoS attacks, SQL injection, and crosssite scripting attacks. This dataset contains a relatively small number of samples but covers various types of attacks and can be used to evaluate the accuracy and robustness of algorithms. The CIC-DDoS2019 dataset contains large-scale DDoS attacks and normal network traffic. This dataset is currently one of the largest available for evaluating large-scale DDoS attack detection algorithms. During the model testing process, research also needs to select and adjust specific parameters and thresholds in order to obtain the best detection results.

In addition, when adjusting hyperparameter, the following strategies are used: first, preprocess the dataset, such as missing value filling, data standardization, and normalization. This step can help us better understand and analyze the features of the dataset and reduce the impact of errors and noise. Secondly, select the hyperparameter, for example, the square prediction error threshold selected in this study. Once again, it is necessary to determine evaluation indicators to adjust the square prediction error threshold and study selecting indicators such as F1 score, precision, and recall for evaluation and adjustment. Afterwards, research will use methods such as grid search or random search. When choosing different combinations of hyperparameter and thresholds for adjustment, attention should be paid to avoiding overfitting or underfitting. Finally, evaluate the anomaly detection model. Then, according to the evaluation results, the hyperparameter can be fine-tuned again.

3. Research on Subspace-Based Anomaly Detection Method for Large-Scale Campus Network Traffic

3.1. Information Entropy Traffic Matrix Combined with Subspace Analysis for Network Traffic Anomaly Detection. Both subspace-based and information entropy-based network traffic anomaly detection methods are commonly used in detecting large-scale network carrier information systems. The study uses the traffic matrix to organically combine the two and construct a comprehensive model of information entropy traffic matrix combined with subspace methods for network traffic anomaly detection of large-scale campus network traffic [20]. The traffic matrix can facilitate the observation of the traffic distribution of the whole network system, then integrate and analyze the traffic distribution with the information of the routers so that the operation of the whole network link can be observed intuitively. When the traffic matrix produces huge fluctuations or changes, it represents an abnormality in the network, and the manager can trace the cause through that abnormality and complete the handling of the abnormality. If D represents the traffic matrix of a network and the network has n nodes, the expression of the traffic matrix D of the network is shown in

$$D = \left[d_1, d_2, \cdots, d_i, d_{n+1}, d_{n(n-1)}\right]^T. \tag{1}$$

In equation (1), each element represents a node pair and d_j represents no. j node pair. A node can generate links with $n-1$ nodes to generate node pairs, so there are $n(n-1)$ node pairs in a traffic matrix. Let C be the value of traffic on a link, r be the number of links, and b be the number of node pairs. In this network, there are n nodes, and the amount of node pairs is $b = n(n-1)$. When the node pair d_j passes through the link i, the value of x_{ij} is 1; otherwise it is 0. The routing matrix $X = (x_{ij})_{rb}$ is a $0-1$ matrix, and each column in the matrix represents all the links that the node pair passes through. $C = XD$, i.e., link load = routing matrix × traffic matrix. The link load and routing matrix obtained by the conventional method is seriously affected by the antecedent information, and the estimation results obtained have large errors, especially in the large-scale network environment. Therefore, the traffic matrix is analyzed by the information entropy method, which can describe the changes of environmental information more accurately, find abnormal conditions in the changes, and classify them. The self-information formula obtained by defining d_i with information entropy is shown in

$$I(d_i) = \log_2 \frac{1}{P_i} = -\log_2 P_i. \tag{2}$$

In equation (2), P_i denotes the probability of generation of x_i, when the expression of information entropy is shown in

$$H(D) = \sum_{i=1}^{N} P(d_i) I(d_i) = -\sum_{i=1}^{N} P(d_i) \log_2 P(d_i). \tag{3}$$

Equation (4) is obtained according to the definition of information entropy.

$$P(d_i) = \frac{n_i}{\sum_{i=1}^{N} n_i}. \tag{4}$$

Bringing equation (4) into equation (3), equation (5) is obtained.

$$H(D) = -\sum_{i=1}^{N} \left(\frac{n_i}{S}\right) \log_2 \left(\frac{n_i}{S}\right). \tag{5}$$

In equation (5), S represents the amount of times a specific attribute appears in the overall flow system, which is calculated as shown in

$$S = \sum_{i=1}^{N} n_i. \tag{6}$$

Use NetFlow to collect network data information such as destination port number, source IP address, total network data packet, protocol type, and total network data

flow. These collected data are treated as information sources, and their attributes are considered random events and used to calculate the information entropy of all attributes. The observation of network traffic anomalies can be achieved by observing the entropy anomaly of the source or destination port number information and the source or destination IP address information. When their entropy values are in a smooth state and converging to 0, it indicates that the network data are in an ordered and discrete form. When the entropy value is fluctuating and the absolute value tends to a larger value, it indicates that the network data is in a disordered and concentrated form. By observing the change of information entropy curve, we can determine whether there is an abnormal situation in the network. As shown in Figure 1, the part of the SrcIP information entropy curve has obvious changes, which means an abnormal situation.

As shown in Figure 1, a large change in entropy value is produced when the time series is about 170 and 605, indicating that a network attack may occur at that moment. Taking the worm invasion as an example, when the attack occurs, the infected host will keep spreading randomly in the network area. When observing the network as a whole, its total network traffic packets and packet count may not change significantly. However, when observing the destination IP address, its information entropy changes dramatically as shown in Figure 1. The number of IP addresses increases rapidly due to the addition of many infected hosts in the network, and its information entropy value then decreases rapidly. By observing whether there are drastic fluctuations in information entropy can determine whether the network has abnormal conditions. The information entropy traffic matrix is combined with the subspace analysis method. The subspace analysis method is a very effective method to study anomaly detection. Its basic idea is to map the network traffic data into a subspace. Normal network traffic will be fully mapped into some subspace $\widehat{S}$, while abnormal network traffic will deviate out and cannot be mapped into that subspace. Instead, abnormal network traffic will go to another subspace $\tilde{S}$. Denote a traffic flow by y and decompose it as shown in

$$y = \hat{y} + \tilde{y}. \tag{7}$$

In equation (7), $\hat{y}$ represents normal network traffic, $\tilde{y}$ represents abnormal network traffic, and the magnitude of $\tilde{y}$ value represents the degree of traffic data abnormality. In the normal network traffic subspace, the principal components in it are processed by principal component analysis, as shown in

$$\begin{cases} \hat{y} = PP^T y = Cy, \\ \tilde{y} = \left(1 - PP^T\right) y = \tilde{C}y. \end{cases} \tag{8}$$

In equation (8), the columns of the matrix P represent the feature vectors, $\hat{y}$ is mapped to the normal network traffic subspace $\widehat{S}$, and $\tilde{y}$ is mapped to the abnormal network traffic subspace $\tilde{S}$. The squared prediction error (SPE) measure is used to measure $\tilde{y}$ values, as shown in

$$y_{SPE} = \|\tilde{y}\|^2 = \left\|\tilde{C}y\right\|^2. \tag{9}$$

In equation (9), y_{SPE} represents the square prediction error value. The squared prediction error y_{SPE} is controlled for the conditions shown in equation (10) when the network traffic is in a normal state.

$$y_{\mathrm{SPE}} \le \sigma_a^2, \tag{10}$$

In equation (10), σ_a^2 represents the threshold of the square prediction error, which is set according to different network situations. The anomaly space is judged again to find the real outliers, and the traffic y is represented as equation (11) in the one-dimensional anomaly detection.

$$y = y^* + \theta_i f_i. \tag{11}$$

In equation (11), y^* represents the nonanomalous flow sample vector, f_i represents the magnitude of the anomalous flow sample, and θ_i represents the correlation vector of the anomalous flow sample. The best value estimation for the anomaly vector is calculated using the minimum value of the distance in the anomaly direction $\tilde{s}_{\min}$, and the result is shown in

$$\tilde{f}_i = \arg\min \|\tilde{y} - \theta_i f_i\|. \tag{12}$$

Combining vector relations is shown in

$$\begin{cases} \tilde{y} = \tilde{C}y, \\ \tilde{\theta} = \tilde{C}\theta_i, \\ \tilde{f} = \left(\tilde{\theta}_i^{\,T}\tilde{\theta}_i\right)^{-1}\tilde{\theta}^{T}{}_i\tilde{y}. \end{cases} \tag{13}$$

Combining the above three equations leads to the optimal valuation is shown in

$$y_i^* = y - \theta_i \tilde{f}_i = \left(1 - \theta_i\left(\tilde{\theta}^T{}_i\tilde{\theta}_i\right)^{-1}\tilde{\theta}_i^{\,T}\tilde{C}\right)\cdot y. \tag{14}$$

The optimal y_i^* is selected to be mapped to the minimum value space of $\tilde{s}_{\min}$. The anomalous value of each anomaly sample is calculated and compared, and if $\tilde{f}_j = \arg\min \|\tilde{C}y_i^*\|$ exists, the flow is judged to be anomalous. Check all node pairs to determine if there are any abnormal traffic conditions. The individual anomaly parameters θ_i are replaced into the form of a matrix to perform similar detection operations and complete the task of anomaly detection for the whole traffic system. For a complex large-scale network system, a single characteristic information entropy value cannot judge all anomalies, so it is also necessary to combine information entropy values of multiple dimensions for network traffic. After serializing the information entropy

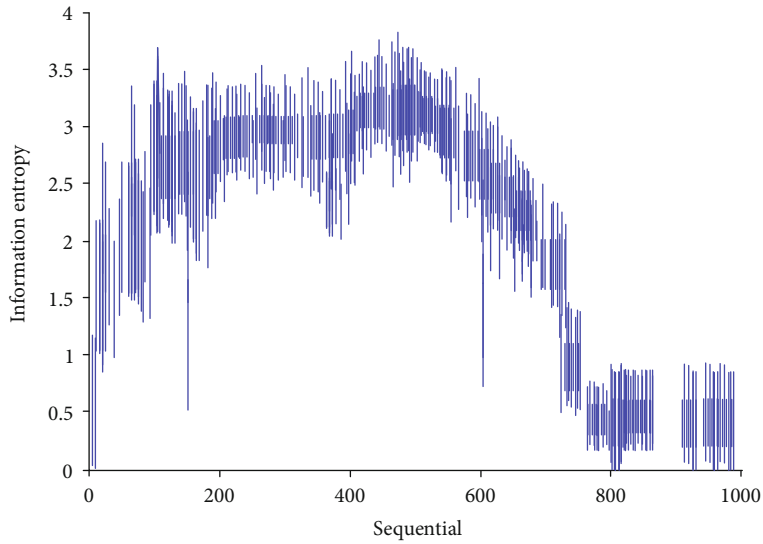

Figure 1: Information entropy time series diagram of SrcIP.

values of multiple dimensions, the corresponding traffic matrix is derived, and then combined with subspace analysis to detect abnormal traffic among them. Using the threshold value as the basis for judgment, the flow of this network information carrier abnormality detection is shown in Figure 2.

3.2. Research on Network Traffic Anomaly Detection Based on Subspace Combined with Density Anomaly Weights and Clustering Ideas. Judging anomalous information based on the value domain only will lead to a large error. Introducing the idea of clustering in the subspace and combining the density anomaly weights to judge the anomalous information can greatly improve the ability of the algorithm [21]. Considering the elements in the subspace as different clusters, define the local density ρ_j of a sample point j and the shortest distance γ_j of this sample point from all sample points having higher density values. Both variables can be expressed using the point distance d_{jk}, when the local density ρ_j is calculated as shown in

$$\rho_j = \sum_k \chi\left(d_{jk} - d_c\right). \tag{15}$$

In equation (15), d_c represents the truncation distance, when $x < 1$, the value of $\chi(x)$ is 1. When $x \geq 1$, the value of $\chi(x)$ is 0. In most cases, the number of points of the sample point j in the neighborhood of d_c is the same as the local density value, when the shortest distance γ_j is calculated as shown in

$$\begin{aligned} \gamma_j &= \min\left(d_{jk}\right), \\ j &: \rho_k > \rho_j, \end{aligned} \tag{16}$$

Taking the 2-dimensional space as an example, the sample points are represented by circles, and the numbers indicate the relative density of the sample points. The local density distribution of a certain sample space is shown in Figure 3.

Observing Figure 3, it can be seen that there are two class clusters in the distribution map with sample point 10 and sample point 1 as the clustering centers, respectively. The information of each sample point in the distribution map is extracted, the local density is used as the horizontal coordinate, and the shortest distance is used as the vertical coordinate to construct a decision map for the selection of sample point clustering centers. From the figure, it can be seen that sample point 9, although closer to sample point 10 in terms of local density, is far apart in terms of shortest distance, so they do not belong to the same class of clusters. It can be observed from the decision diagram that the points with both higher local density and larger shortest distance values can only be the clustering centers of the sample class clusters. Sample points that deviate from the group are shown to have a larger shortest distance and a lower relative density. Based on the above analysis, the anomaly weights of local density are used to describe as shown in

$$M_i = \frac{\gamma_i}{\max\left(d_{jk}\right)} e^{-\alpha\rho_i}. \tag{17}$$

In equation (17), α denotes a smoothing parameter taking values between 0 and 1, which is used to adjust the data samples adapted to different situations. d_{jk} is the distance between two sample points j and k. Applying the method to the subspace, the subspace clustering is viewed as a kind of data feature selection in 3D space, as shown in Figure 4.

There are several datasets in 3D space, as shown in Figure 4. The datasets are clustered into 4 class clusters, and the class clusters exist in 2-dimensional space; red and green two-class clusters exist on dimensions a and b. The third dimension belongs to noise. The blue and yellow class clusters exist on dimensions b and c. Let the number of

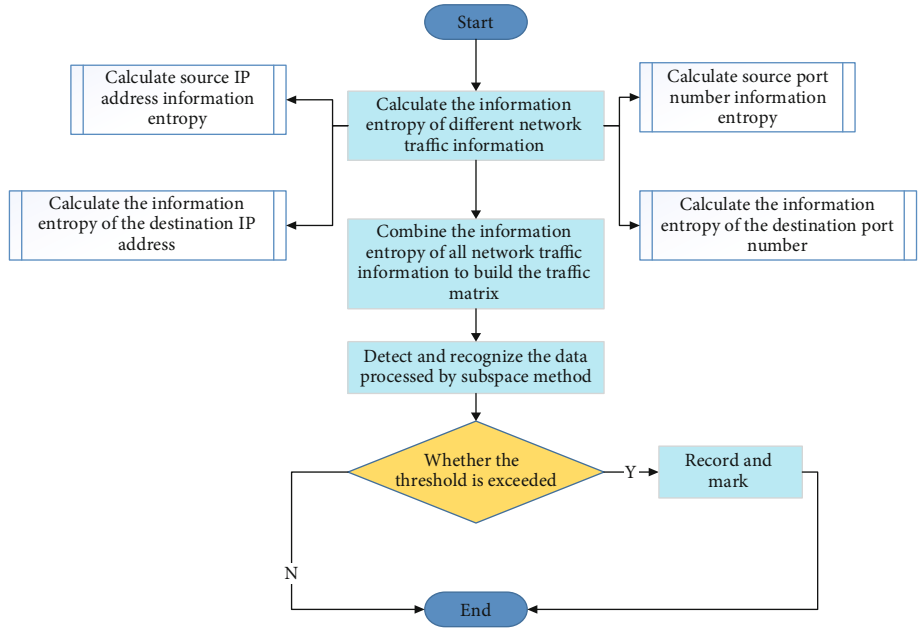

Figure 2: Network information carrier anomaly detection process based on subspace and traffic matrix.

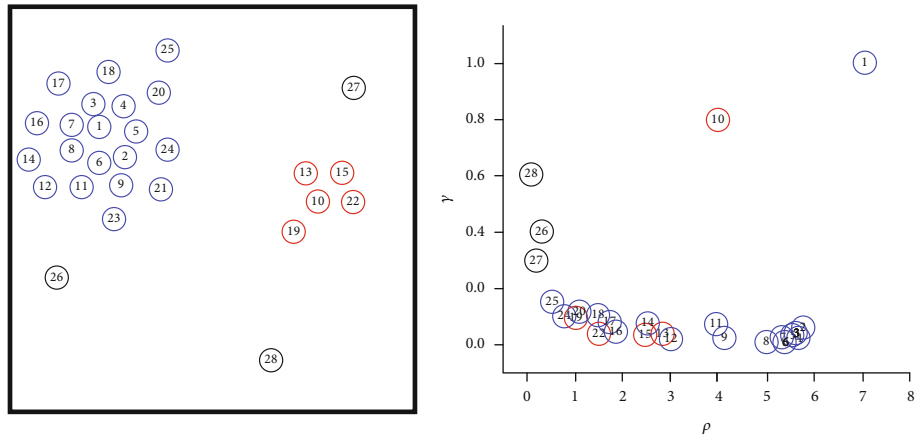

Figure 3: Sample local density distribution diagram and cluster center selection decision diagram.

feature values of original data feature space E be $l$and map them all to u dimensional subspace. The number of u dimensional subspace is o and $u < l$. At this time, the subspace is represented as $E_i(i = 1, 2, \cdots, o)$, and the abnormal weight of each sample data in the subspace is m_i^j, $(j = 1, 2, \cdots, |E_i|)$. Calculate the abnormal weight values of all subspaces and perform corresponding processing. Record the abnormal weight values of each sample in the original data space sample according to the results. The algorithm model framework of network traffic anomaly detection based on subspace combined with density abnormal weight values and clustering idea is shown in Figure 5.

As shown in Figure 5, the sustainable and adaptive wireless charging (SAWC) model first obtains the campus network traffic data information from the campus network router, then applies feature selection on these data information and assigns them to the original space E. Then the data of the original feature space E is mapped to the subspaces of n, and the anomaly weights are assigned to different subspaces in turn. Finally, the anomaly weights of each number of traffic are obtained, and they are integrated and sorted to complete the network traffic anomaly detection. The pseudocode of the SAWC model is as follows:

4. Experiments and Analysis of Subspace-Based Anomaly Detection Method for Large-Scale Campus Network Traffic

This experiment selects the public dataset NSL-KDD and a sample dataset of one day's data flow collected from a university campus network collected by router NetFlow for the experiment. The spatiotemporal interaction networks with factorization machines (SINFM) network traffic anomaly detection model and the SAWC network traffic anomaly detection model proposed in the study are compared with the extreme learning machine (ELM) network traffic

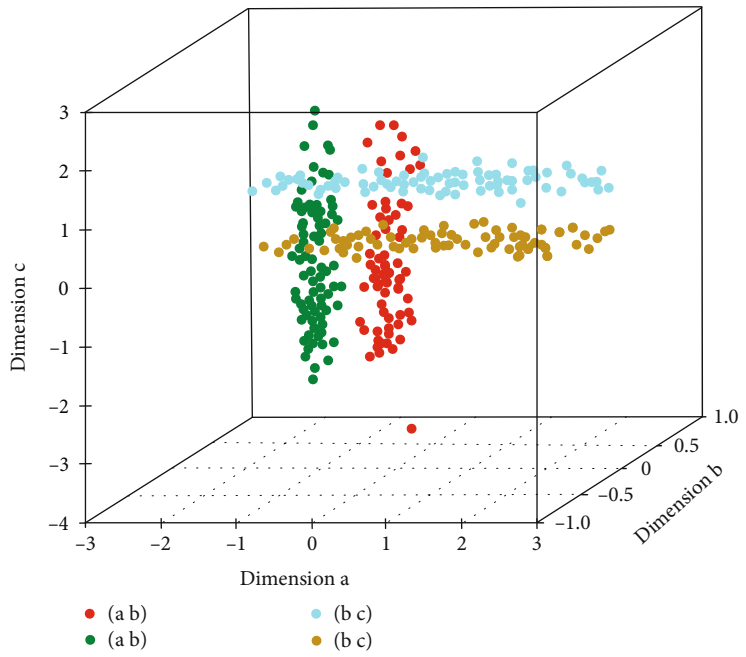

Figure 4: Three-dimensional spatial clustering diagram of data samples.

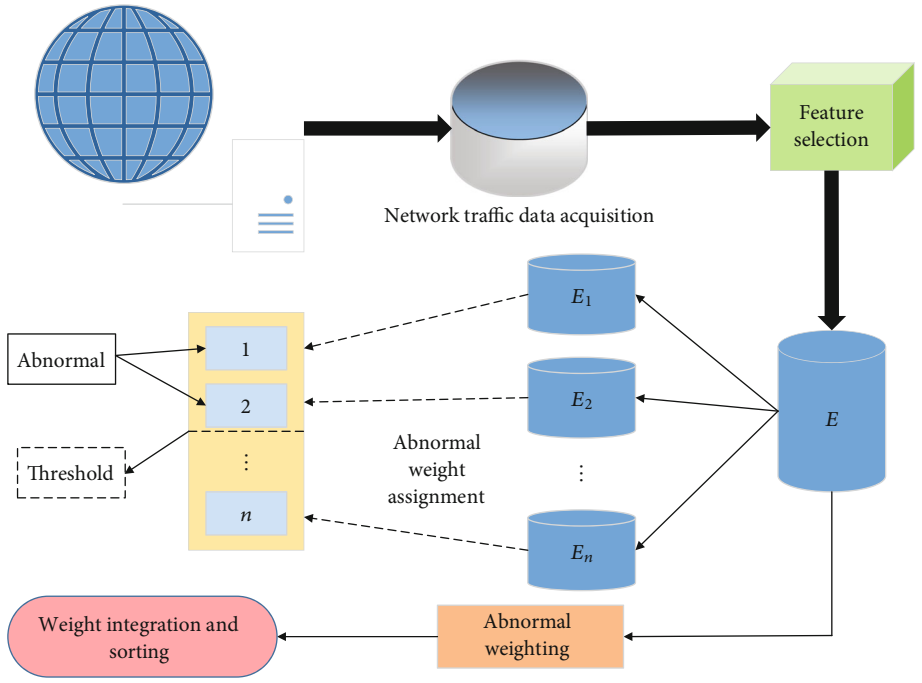

Figure 5: Frame diagram of anomaly detection algorithm based on subspace clustering and anomaly weight.

anomaly detection model proposed in literature [5] and the support vector machine (SVM) network traffic anomaly detection model proposed in literature [20]. The SINFM pseudocode is as follows:

The analysis experiments are conducted. The parameters of each model are tuned before the experiment to ensure the accuracy of the experiment. The square prediction error threshold of SINFM σ_a^2 is set to 99.3%, the smoothing parameter of SAWC α is set to 0.28, and the anomaly weight threshold of data traffic θ is set to 0.835. There are 7638 pieces of attack data, which is 13.43% of the total data. These data were normalized in the range of [0, 1], and the types of attacks, types of attacks, number of attacks, and percentage of experimental data are shown in Table 1.

Comparing the accuracy of each algorithm model for network traffic anomaly detection under partial NSL-KDD dataset, the obtained experimental results are shown in Figure 6.

```
for all nodes v in the network:
    if v's energy level <= threshold:
        if there exists C in charging stations that can charge node v:
            Move charging battery C to node v;
        else:
            Move charging battery to nearest charging station;
    if v's energy level >= optimal_threshold:
        if there exists C in charging station that can be charged by node v:
            Move charging battery C to charging station
for charging station g:
    if g's energy level<= threshold:
        if there exist v in nodes that can charge charging station g:
            Move battery from node v to charging station g
```

Pseudocode 1: SAWC pseudocode.

```
Input: spatiotemporal interaction networks with factorization machines (SINFM) dataset D = (X, Y, F, G, T)
Output: prediction result
1: Train dataset D with factorization machines
2: Calculate the embeddings for the temporal dataset
3: Calculate the embedding for each region
4: Calculate the embedding for each topic
5: Define the temporal interaction network
6: Define the spatial interaction network
7: Define the topic interaction network
8: Generate the three submatrices based on the three interaction networks
9: Train the interaction terms for spatiotemporal data in each submatrix
10: Calculate the scores of each testing item for each submatrix
11: Aggregate the scores of each testing item in the three submatrices
12: Output the final prediction result
```

Pseudocode 2: The SINFM pseudocode.

Table 1: Partial NSL-KDD dataset information.

Data type	Included attack types	Number of pieces	Proportion
Normal	/	49252	87.23%
Dos	Neptune, pod, back, land, smurf, teardrop	2885	6.14%
Probes	Ipsweep, satan, portsweep, namp	4163	7.58%
R2L	mltihop, ftp_write, warezclient, phf, warezmasyer, spy, guess_password, imap	261	0.32%
U2R	Loadmudule, perl, buffer_overflow	508	0.83%

Figure 6 shows that all four network information carrier anomaly detection models have good anomaly detection functions, with overall accuracy rates of more than 90%. The best performers are SAWC and SVM, with an overall accuracy rate of 97.1%, which is 5% higher than ELM. For normal data traffic, SVM has the highest accuracy rate, followed by SAWC. Since normal traffic occupies the highest percentage of data, accurate release of normal traffic is an important means to effectively improve the accuracy of the traffic anomaly detection model. SINFM and ELM need to be improved in this regard. The best detection accuracy among the four types of attacks, DOS, probes, R2L, and U2R, is SAWC, and this result is in line with expectations, proving that SAWC has a good detection effect on network traffic anomalies. The performance of the SINFM model is not outstanding, and the detection accuracy of all types is at a medium level. This result is consistent with the expectation that the SINFM model is positioned to prioritize large-scale traffic processing, and its accuracy performance will not be particularly outstanding. The best overall performance for attack detection is achieved by SAWC, while SVM performs best in the detection of normal traffic but has a relatively low accuracy in the detection of all types of attacks. To continue comparing the detection speed of each model, 10 separate experiments were conducted on each model to record its data detection time, and the results obtained are shown in Figure 7.

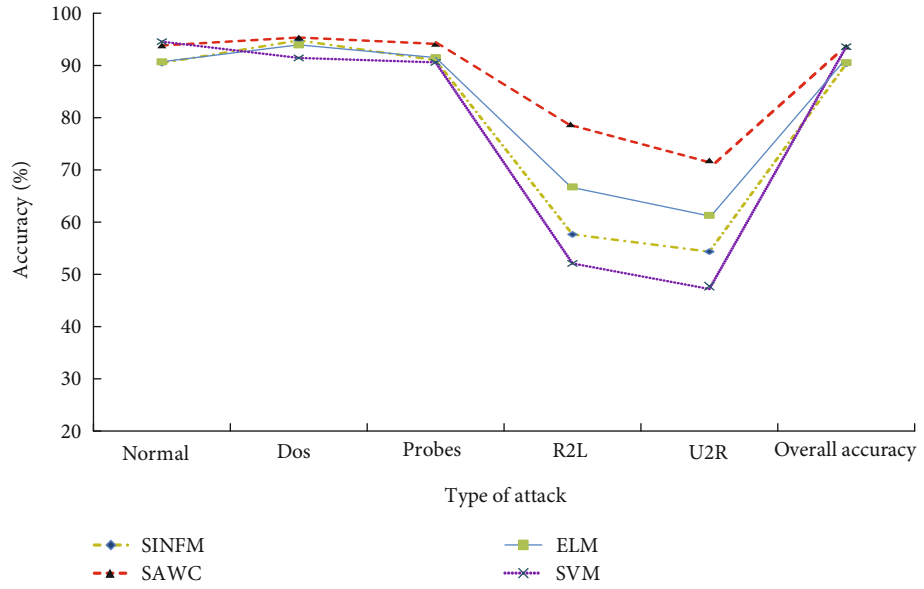

Figure 6: Accuracy of each detection model in specific attack types.

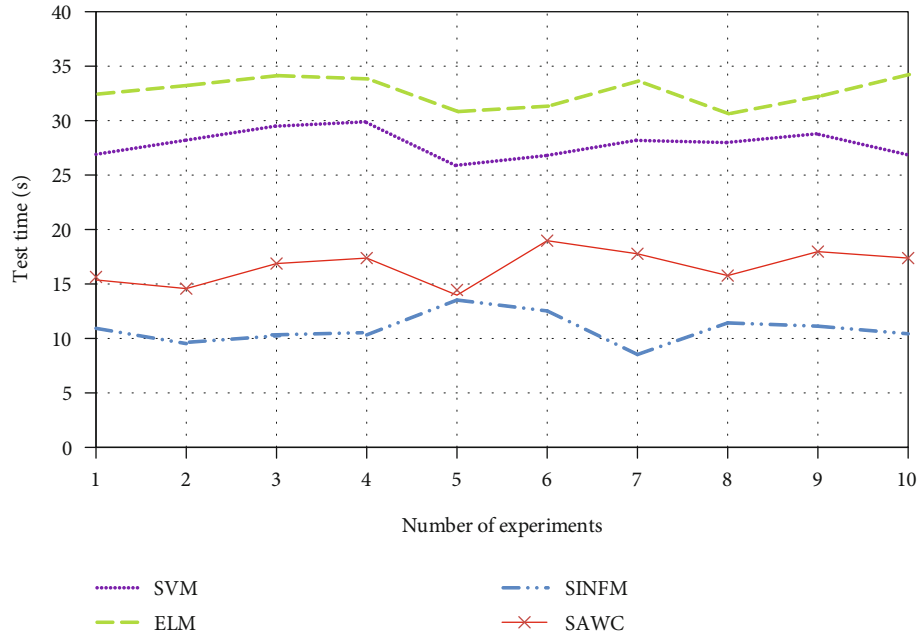

Figure 7: Comparison of detection time of NSL-KDD data by each detection model.

Figure 7 shows that the average detection time of the proposed SINFM detection model is 11.67 seconds. The average detection time of the SAWC detection model proposed in the study is 15.72 seconds, which is faster than the comparison models. This experimental result is in line with expectations; that is, both models proposed in the study are designed to adapt to anomaly detection of large-scale campus network information carriers and require high data processing speed. During the construction, complex operations and judgment processes were reduced, and the operational efficiency of the model was improved. In terms of operational efficiency, the SINFM detection model has the best performance in the NSL-KDD dataset, followed by the SAWC detection model. In terms of detection accuracy, the SAWC detection model has the best performance, followed by the SVM detection model. Combining these two aspects, the SAWC detection model performs better overall, with faster detection speed while maintaining higher accuracy. Apply these four models to real-world scenarios and conduct experiments using CSE-IC-IDS2019 (large-scale campus network data), which is a sample dataset of one-day data streams collected from university campus networks through router NetFlow. The experimental results obtained are shown in Figure 8. This dataset contains

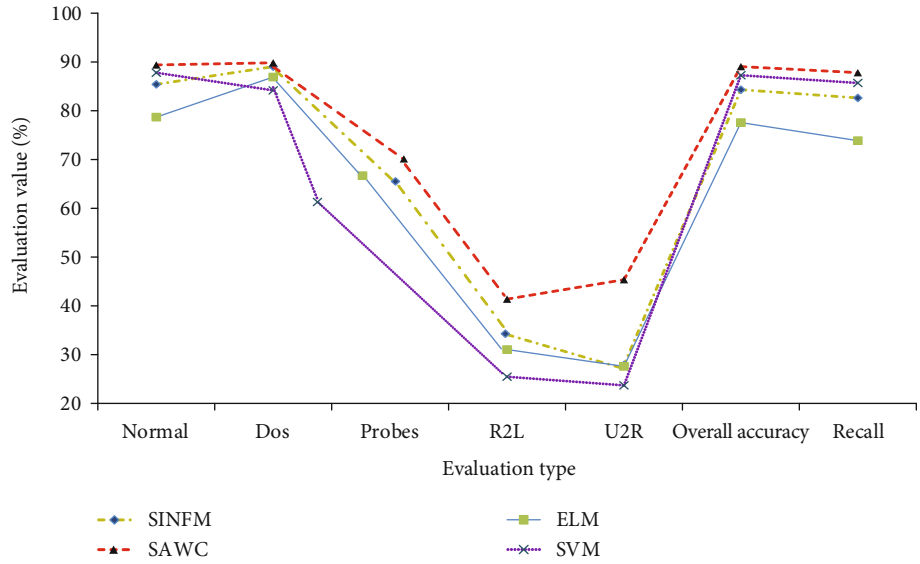

Figure 8: Experimental results of each algorithm model under CSE-CIC-IDS2019 dataset.

Table 2: Comparison of experimental results of this method and related work on various indicators.

Algorithm	Accuracy rate (%)	Precision rate (%)	Recall rate (%)	F1-score (%)
ELM in the literature [5]	86.16	85.42	86.41	83.75
SVM in the literature [16]	87.98	87.71	89.26	87.29
K-SVM in the literature [22]	92.23	88.59	92.42	88.99
MRA S-dDCA [23]	99.72	98.97	99.54	99.32
The SAWC proposed in this paper	99.81	99.22	99.56	99.49

583495 data, of which 59135 are abnormal data, accounting for 10.13%.

The accuracy of the SAWC detection model is 97.81%, which is 12.99% higher than the ELM detection model and 2.01% higher than the SVM detection model. The recall rate of the SAWC detection model is 96.72%, which is 15.77% higher than the ELM detection model and 2.39% higher than the SVM detection model. Compared with the NSL-KDD dataset, the accuracy and recall rate of the SINFM detection model have been improved, indicating that the SINFM model is also well adapted to large-scale campus network information carrier anomaly detection tasks and has a faster detection speed. Reference [22] used a K-means hybrid model combined with the SVM algorithm for training on the dataset CSE-IC-IDS2018. Based on the tuning results of the models in various literatures, the corresponding model parameters were adjusted to continue the comparative experiment. The comparison results of the accuracy, precision, recall rate, and F1 score performance indicators of each model under the CIC-DDoS2018 dataset were obtained, as shown in Table 2.

As shown in Table 2, the SAWC detection model proposed in this paper performs the best in anomaly detection under the dataset CIC-DDoS2018, followed by the combined model of MRA S-dDCA. This experimental result proves that the clustering algorithm can play a good role in traffic anomaly detection, which has a great improvement on the accuracy and efficiency of the detection algorithm. Comprehensive analysis of the appeal experimental results shows that both the SAWC detection model and the SINFM model proposed in the study are better able to accomplish traffic anomaly detection in a large-scale campus network traffic environment, and the SINFM detection model is capable of the detection task, but the detection capability still needs to be improved. The SAWC detection model performs stably under multiple data, and both outperform similar algorithmic models.

This type of model needs to be able to handle large-scale datasets well and have good universality and adaptability. At the same time, the model should have high accuracy and robustness and be able to effectively detect and recognize under different types of attacks. In addition, the detection speed of the model should also be fast enough to adapt to real-time detection needs [24, 25]. The results obtained from this study meet these requirements.

Firstly, SAWC and SVM perform well in detecting normal traffic, which is related to their strong performance in handling classification problems and their ability to adapt well to large-scale datasets. Secondly, SAWC performs best in four types of attack detection, indicating that the SAWC model has better ability to identify and handle attack traffic. In addition, the performance of the SINFM model is not

outstanding, which may be due to its greater focus on large-scale traffic processing without optimizing model performance. Finally, the SAWC detection model has faster detection speed and higher accuracy, indicating its superior performance in handling large-scale campus network information carrier anomaly detection tasks. In summary, the SAWC detection model has high feasibility and applicability in large-scale campus network information carrier anomaly detection tasks.

5. Conclusion

Aiming at the problem of traffic anomaly detection methods in large-scale campus networks, a SINFM detection model combined with information entropy traffic matrix and a SAWC model combined with density anomaly weight and clustering idea based on subspace analysis are proposed. The results showed that under the NSL-KDD dataset, the accuracy of the SAWC model was as high as 97.1%. The average detection time of SINFM is 11.57 seconds. The average detection speed of SAWC is 15.72 seconds. 13.22 s. Under the CSE-IC-IDS2019 dataset, the accuracy of the SAWC detection model is 97.81%, and the recall rate is 96.72%. Under the CIC-DDoS2018 dataset, the accuracy of SAWC is 99.72%, the accuracy is 99.22%, the recall rate is 99.56%, and the F1 score is 99.49%. Compared with other methods proposed in the literature, the SAWC model performs better than all other detection algorithm models. The experimental results meet expectations and demonstrate that the model proposed in the study can perform traffic anomaly detection in large-scale campus network environments. There are also some shortcomings in this study, such as insufficient sample richness in the experimental dataset, relatively simple actual campus network datasets, and insufficient meticulous data processing. It is expected that in future research, more campus network information datasets can be tested and adapted to more campus network environments to promote more comprehensive and accurate research.

Data Availability

The data used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Y. Ma, J. Feng, and J. Li, "Method for identifying abnormal access to sensitive data based on network flow," *Journal of Physics: Conference Series*, vol. 1646, no. 1, article 012067, 2020.

[2] M. Roshanaei, "Resilience at the core: critical infrastructure protection challenges, priorities and cybersecurity assessment strategies," *Journal of Computer and Communications*, vol. 9, no. 8, pp. 80–102, 2021.

[3] S. Feng, W. Ying, and W. Wu, "Analysis of DNS security threats on campus network," *Journal of Physics Conference Series*, vol. 1601, article 052018, 2020.

[4] R. H. Hwang, M. C. Peng, C. W. Huang, P. C. Lin, and V. L. Nguyen, "An unsupervised deep learning model for early network traffic anomaly detection," *IEEE Access*, vol. 8, pp. 30387–30399, 2020.

[5] M. K. Hooshmand and D. Hosahalli, "Network anomaly detection using deep learning techniques," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 2, pp. 228–243, 2022.

[6] Y. Humid, F. A. Shah, and M. Sugumaran, "Wavelet neural network model for network intrusion detection system," *International Journal of Information Technology*, vol. 11, no. 2, pp. 251–263, 2019.

[7] P. Ding, J. Li, M. Wen, L. Wang, and H. Li, "Efficient BiSRU combined with feature dimensionality reduction for abnormal traffic detection," *IEEE Access*, vol. 8, pp. 164414–164427, 2020.

[8] B. Li, Y. Wang, K. Xu, L. Cheng, and Z. Qin, "DFAID: Density-aware and feature-deviated active intrusion detection over network traffic streams," *Computers & Security*, vol. 118, p. 102719, 2022.

[9] M. Amoozegar, B. Minaei-Bidgoli, M. Rezghi, and H. Fanaee-T, "Extra-adaptive robust online subspace tracker for anomaly detection from streaming networks," *Engineering Applications of Artificial Intelligence*, vol. 94, article 103741, 2020.

[10] Y. Zhong, S. Fong, S. Hu, R. Wong, and W. Lin, "A novel sensor data pre-processing methodology for the Internet of Things using anomaly detection and transfer-by-subspace-similarity transformation," *Sensors*, vol. 19, no. 20, p. 4536, 2019.

[11] S. Guo, Y. Liu, and Y. Su, "Network traffic anomaly detection method based on CAE and LSTM," *Journal of Physics: Conference Series*, vol. 2025, no. 1, article 012025, 2021.

[12] W. Song, M. Beshley, K. Przystupa et al., "A software deep packet inspection system for network traffic analysis and anomaly detection," *Sensors*, vol. 20, no. 6, pp. 1637–1637, 2020.

[13] Y. Zhong, W. Chen, Z. Wang et al., "HELAD: a novel network anomaly detection model based on heterogeneous ensemble learning," *Computer Networks*, vol. 169, no. 14, article 107049, 2019.

[14] S. T. Zhang, X. B. Lin, L. Wu, Y. Q. Song, N. D. Liao, and Z. H. Liang, "Network traffic anomaly detection based on ML-ESN for power metering system," *Mathematical Problems in Engineering*, vol. 2020, Article ID 7219659, 21 pages, 2020.

[15] H. Peng, L. Liu, J. Liu, and J. R. Lewis, "Network traffic anomaly detection algorithm using mahout classifier," *Journal of Intelligent and Fuzzy Systems*, vol. 37, no. 1, pp. 137–144, 2019.

[16] H. Wu, B. Zhang, and S. Dong, "A hybrid feature selection method for network traffic anomaly detection," *Journal of Physics: Conference Series*, vol. 1395, no. 1, article 012015, 2019.

[17] R. Wang, J. Fang, Z. Yang, and H. Li, "Multi feature selection based network traffic anomaly detection method," *Journal of Physics Conference Series*, vol. 1288, no. 1, article 012003, 2019.

[18] R. M. Aliguliyev, M. S. Hajirahimova, and S. Office, "Classification ensemble based anomaly detection in network traffic," *Review of Computer Engineering Research*, vol. 6, no. 1, pp. 12–23, 2019.

[19] L. Khatibzadeh, Z. Bornaee, and A. G. Bafghi, "Applying catastrophe theory for network anomaly detection in cloud computing traffic," *Security and Communication Networks*, vol. 2019, Article ID 5306395, 11 pages, 2019.

[20] N. Liao, Y. Song, S. Su, X. Huang, and H. Ma, "Detection of probe flow anomalies using information entropy and random forest method," *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 1, pp. 433–447, 2020.

[21] Y. Shi and H. Shen, "Anomaly detection for network flow using immune network and density peak," *International Journal of Network Security*, vol. 22, no. 2, pp. 337–346, 2020.

[22] M. Jain, G. Kaur, and V. Saxena, "A K-means clustering and SVM based hybrid concept drift detection technique for network anomaly detection," *Expert Systems with Applications*, vol. 193, article 116510, 2022.

[23] D. Limon-Cantu and V. Alarcon-Aquino, "Multiresolution dendritic cell algorithm for network anomaly detection," *PeerJ Computer Science*, vol. 7, article e749, 2021.

[24] S. Dong, Y. Xia, and T. Peng, "Network abnormal traffic detection model based on semi-supervised deep reinforcement learning," *IEEE Transactions on Network and Service Management*, vol. 18, no. 4, pp. 4197–4212, 2021.

[25] Y. Xia, S. Dong, T. Peng, and T. Wang, "Wireless network abnormal traffic detection method based on deep transfer reinforcement learning," in *2021 17th International Conference on Mobility, Sensing and Networking (MSN)*, pp. 528–535, Exeter, United Kingdom, 2021.