Network Security Detection Method Based on Abnormal Traffic Detection

Tao Xiao*, Yang Ke, Hu YiWen, Wang HongYa
State Grid Jiangxi Electric Power Co, Ltd. Training Center, Nanchang 330013, China

***Abstract*—To discover potential risks and vulnerabilities in the network in time and ensure the safe operation of the network, a network security detection method based on abnormal traffic detection is studied. Construct network security detection architecture from several aspects, including the front-end interface module, control center module, network status extraction module, anomaly detection module, alarm module, and database module. Use NetFlow technology to capture network traffic from the network in the form of flow, and use the KNN algorithm in the traffic filtering submodule to filter network traffic packets and eliminate duplicate traffic data. After filtering traffic, the traffic data is transmitted to the feature selection sub-module. PCA-TS algorithm is used to reduce the dimension of the network traffic data and select the network traffic characteristics, and then it is input into the SVM classifier. The improved SVM multi-classification algorithm is used to classify normal and abnormal traffic, complete abnormal traffic detection, and achieve network security detection. Experimental results show that the time for feature selection of this method does not exceed 3.0s, and the G score in the detection process also remains above 0.70, indicating that this method has strong network security detection capability.**

***Keywords*—Abnormal traffic; network security detection; data dimensionality reduction; flow characteristics; traffic capture; alarm module**

I. INTRODUCTION

Network security detection refers to a comprehensive security assessment and inspection of the computer network system to find potential security risks and vulnerabilities and take corresponding measures to protect the security and integrity of the network system [1]. Its significance lies in preventing potential threats, protecting important data, maintaining business continuity, improving user trust and complying with regulatory requirements. By preventing potential threats, vulnerabilities and weaknesses in the network system can be found and repaired in time to avoid security attacks [2]. At the same time, security incidents such as hacker intrusion, data leakage and malware infection can be avoided by timely finding and solving security problems [3]. Network security detection can help enterprises and individuals protect important business and personal data. Data security and confidentiality can be ensured by detecting vulnerabilities and risks in the network system [4]. It can also ensure the normal operation of the network, reduce business interruption and loss caused by security vulnerabilities and attacks, find and repair the vulnerabilities in the network system in time, and ensure the continuity and stability of the business [5]. By improving network security through network security detection, users can use online services more confidently without worrying about personal information leakage or account theft [6]. Therefore, network security detection is of great significance.

With the popularization of computer network applications and services, the number of Internet users continues to increase, and the demand for Internet information sharing continues to expand [7]. The threat of network security attacks has become more serious, and network anomaly detection has become an increasingly important task in network security research [8]. There are many reasons for network exceptions, such as network overload, worm network intrusion, routing policy modification, and distributed denial of service attacks. Network traffic anomaly is the most common threat in network anomaly. Abnormal network traffic may reduce the central network speed or even cause network paralysis, which will cause serious damage to the network environment [9]. Weihai caused by abnormal network traffic is generally characterized by bandwidth occupation, network blocking, failure to send normal information on time, network packet loss, etc. [10]. For computer systems, servers, and clients, the harm caused by abnormal network traffic is shown as occupying a large amount of memory space, and data responses are transmitted to the server normally with different responses [11]. Many scholars have studied network security detection methods to solve these threats to network security and achieve timely hazard warnings and other functions.

Wozniak M et al. [12] studied the cyclic neural network model for threat detection of the Internet of Things and network malware. This method classifies the information in the network through the cyclic neural network to detect malicious threat information, but this method cannot achieve security alarm in the detection process, and the detection results of multiple attacks are not clear enough; Steno P et al. [13] studied uses deep learning to detect threat objects in security screening. This method uses deep learning network and cross-entropy loss calculation to realize risk screening in network objects, but this method cannot detect the degree of flow fluctuation in the network, resulting in a small detection range; Gaber T et al. [14] studied an injection attack detection method for intelligent Internet of Things applications using machine learning. This method detects network attack traffic in smart cities, uses constant removal and recursive feature elimination methods to achieve feature selection, and uses machine learning classifiers to classify attack traffic. Although the accuracy of this method is as high as 99%, this method needs a lot of time in feature extraction and detection.

The abnormal traffic detection method identifies and detects abnormal situations inconsistent with normal network traffic behavior by analyzing the network traffic changes. This

detection method can help find abnormal activities in the network, such as network attacks, malware propagation, data leakage, etc., to ensure the network's security and stability [15]. There are many methods to detect abnormal network traffic, such as traffic analysis methods, machine learning methods, etc. The traffic analysis method determines whether the traffic is abnormal by capturing and parsing the network data packets, analyzing the source address, destination address, protocol type and other information of the data packets, as well as the size, frequency and other characteristics of the data packets. For example, an exception may exist if an IP address sends many data packets quickly or a port receives abnormally large data traffic. The machine learning method uses algorithms to analyze and model network traffic to identify abnormal traffic. Machine learning can automatically identify abnormal data packet size, abnormal connection behavior, etc. [16] by learning the characteristics of normal traffic behavior. Therefore, this paper proposes an abnormal traffic detection method based on support vector machine (SVM). The innovation lies in the construction of a network security detection architecture, which uses NetFlow technology to capture traffic data from the network and uses KNN algorithm to remove duplicate data. In the feature selection submodule, PCA-TS algorithm is used to reduce the dimensionality of network traffic and select features. An improved SVM multi classification algorithm is used to classify normal and abnormal traffic, achieving efficient abnormal traffic detection and network security detection.

II. Design of Network Security Detection Method

A. Construction of Network Security Detection Architecture

To improve the network security and operation status, this paper studies the network security detection architecture, shown in Fig. 1. The architecture is divided into a foreground interface, control center, network status extraction, anomaly detection, alarm, and database modules. Feature selection and anomaly detection modules are the architecture's main parts.

The specific contents of the network security detection architecture are as follows:

1) Front-end interface: The main functions of the front-end interface module include network topology node display function, user information display and user login function, system working status display function, log information display function and abnormal flow alarm display function. The main function of the foreground interface is to provide users with a good and beautiful interface display function and provide users with a good interaction function. The network topology node display function can view the current network's working nodes and the current network's topology. The alarm display function can more intuitively see the attacked node. The log information display function can view the system log, including user login information and running status record information on the architecture.

2) Control center: The main functions of the control center module include user authority management function, configuration function, alarm function, task allocation function, etc. The user authority management function can ensure that the authority is not abused. Only super users have administrator authority to modify sensitive content such as network configuration information. The alarm function makes it possible to quickly give feedback to the foreground interface to prompt network exceptions when the detection architecture is abnormal and even locate the attacked node in time. The task allocation function mainly includes switching between the network status extraction function, normal detection status, and log viewing function.

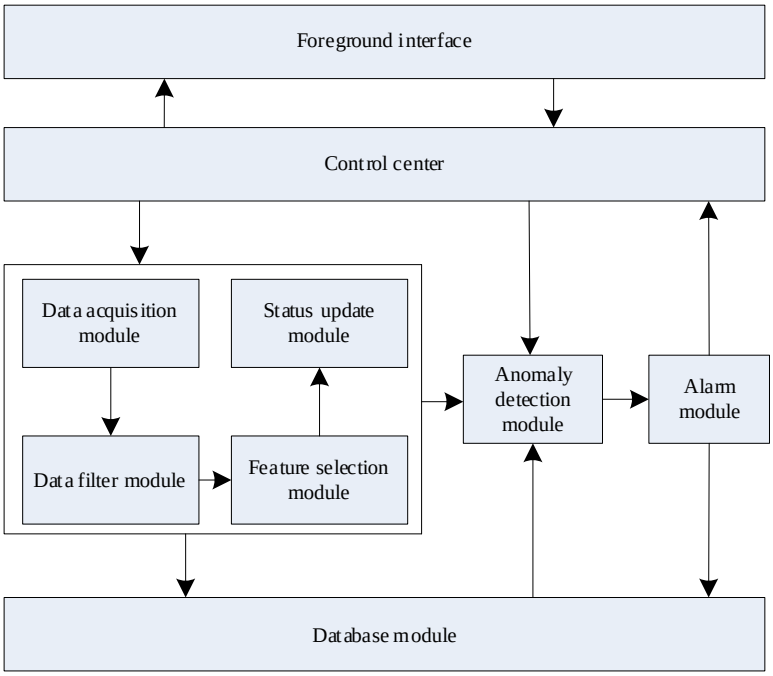


Fig. 1. Overall design of network security detection architecture.

3) Feature selection module: The main function of the feature selection module is to capture network traffic through the function of collecting data and selecting the features of the captured traffic to achieve feature extraction. In the safe working mode, the module saves the extracted feature information to the database. In the normal working mode, the extracted features are transmitted to the abnormal flow detection module to judge the abnormal flow of the network. In the update mode, the extracted features are updated to the database. Feature selection module is one of the most important functional modules, including four sub-modules: traffic capture module, traffic filtering module, feature selection module and network status update module.

The traffic capture submodule's main function is to capture each node's data packets and then save the data packets to the packet queue. The main function of the traffic filtering sub-module is to filter the redundant content and duplicate content captured and then wait for the feature selection sub-module to extract features. The feature selection sub-module is the core module of the architecture. It receives the data package of the filter sub-module, selects its features, and saves the extracted state information to the database. The status update sub-module mainly updates the recorded standard status information.

4) Anomaly detection module: The anomaly detection module is mainly responsible for judging whether the network has been attacked by abnormal traffic and passing the detection results to the alarm module. At the same time, the detection results are saved in the database as log management, which can facilitate viewing historical record information. The algorithm implements this module. At the same time, this module is one of the most important functional modules and is the key judge of whether an intrusion occurs.

5) Alarm module: The main function of the alarm module is to send an alarm to the user when the abnormal flow detection module judges that an intrusion has occurred. The alarm module receives the result of the abnormal flow detection module. If the result is true, the alarm information will be written into the database and sent to the control center simultaneously. The alarm can be sent out an alarm tone or pop up a window light on the foreground interface.

6) Database module: The main function of the database module is to record various types of information, including log information, user information records, alarm information records, etc. The database module is the information center of the entire architecture. All data extraction, information exchange and record-keeping between modules are completed in this module. The database module requires the physical support of the database software, or it can be deployed to an independent server separately.

B. Network Traffic Capture

Based on NetFlow technology, the traffic capture sub-module under the feature selection module in the network security detection architecture captures network traffic in flow. Use the V9 NetFlow technology as a sniffer or probe in the network to transmit the traffic records to a data collector with a specified IP address. The output package format of NetFlow V9 is shown in Fig. 2. FlowSet represents the collection of traffic records. The Template Flowset in Fig. 2 is the template for subsequent data records, enhancing network traffic records' flexibility.

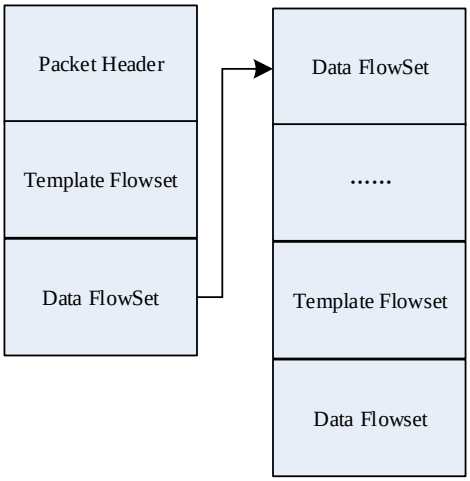


Fig. 2. Analysis of the output package format of NetFlowV9.

Encapsulates the network traffic records into UDP packets and transmits them to the collector with UDP protocol to ensure high efficiency when transmitting a large number of traffic records. To prevent NetFlow from generating a large amount of data and causing network congestion, a dedicated link is designed for the traffic record output to the collector in the congestion-sensitive network. When the collector cannot be placed at the router's next hop or the transmission link cannot be exclusive to NetFlow, a special link needs to be designed to handle the large amount of data NetFlow generates.

C. Traffic Filtering of KNN Classification Algorithm

After capturing the network traffic data, due to the huge amount of network traffic data [17], and some botnet traffic and duplicate content, effective measures must be taken to filter the traffic. Through reasonable filtering means, the traffic data can be more conducive to subsequent network security detection [18]. This paper uses the KNN algorithm to filter traffic in the traffic filtering sub-module under the feature selection module.

1) Filtering analysis: KNN algorithm realizes classification by measuring the distance between different eigenvalues to achieve a data filtering effect. If most of the k, that the most similar samples belong to a category, the samples also belong to that category. The most similar definition here is that the eigenvalue of the sample is the nearest in the feature space. Among K is an integer less than or equal to 20. In the KNN algorithm, the most similar samples are correctly divided into corresponding classes [19]. When this method is used for classification, the possible classification of samples is determined according to the classification of the nearest one or several samples. Fig. 3 shows the operation mode of the algorithm.

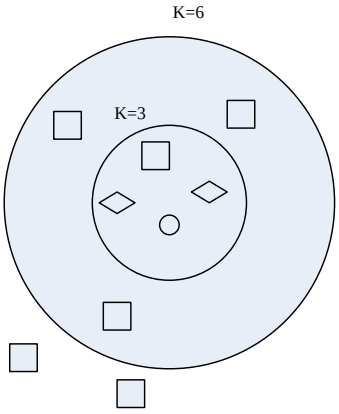


Fig. 3. Analysis of KNN algorithm operation mode.

It can be seen from Fig. 3 that when $K = 3$, the grey filled circle shall belong to the hollow triangle type. When $K = 6$, the grey filled circle shall be classified as a hollow square. Therefore, it can be concluded that the selection of values K affects the filtering results; the optimal value can make the filtering effect the best.

2) Flow filtration: Due to the large difference between normal traffic and abnormal traffic caused by attacks, the external parameters of abnormal traffic can be filtered out from the external parameters of mixed traffic, including normal and abnormal traffic, using the KNN method to achieve faster traffic data processing [20]. The implementation process is:

The obtained labelled data (packet length, URL length) and unlabeled data (packet length, URL length) are regarded as vectors, and their Euclidean distances are calculated. European distance can be calculated by Formula (1):

$$dist(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (1)$$

In Formula (1), $dist(X,Y)$ represents a data set between European distance of X and Y; x_i and y_i respectively represent i of the labeled data and unlabeled data; n indicates the number of traffic data. This formula can measure the absolute distance between points in multidimensional space.

3) The first one closest to the unlabeled data is counted K the labelled data with the most occurrences among the data will be marked with the same label as the unlabeled data. Obtain a batch of labelled data, which can support subsequent work in terms of data quantity.

D. Design of Network Traffic Feature Selection Method

Real network traffic contains many feature attributes, and the existing anomaly traffic detection methods based on feature analysis cannot meet the real-time requirements of high-dimensional feature analysis [21], [22]. Therefore, when this paper filters the traffic data through the KNN algorithm, the feature selection submodule uses the traffic feature selection algorithm based on principal component analysis (PCA) and tabu search (TS) to conduct feature reduction and near-optimal feature subset selection for high-dimensional features through PCA-TS, providing reliable feature data for subsequent abnormal traffic detection.

1) Dimension reduction of traffic data: The dimensionality of the traffic data filtered by the KNN algorithm is reduced to facilitate the subsequent feature selection [23]. Principal component analysis is an effective method of analyzing data in statistics, mainly used for feature extraction and data dimension reduction. The idea is to reduce the dimension of a data set with high dimension and correlation by using the feature space transformation of statistical properties of the data set [24]. PCA transforms the original space into a new principal component space, and the principal components are unrelated.

Assume that the network traffic data set contains N samples $N = \{x_1, x_2, \ldots, x_m\} \in R^n$, where,$R^n$is the feature space,$m$ is the characteristic dimension. Find variable space $Z = \{z_1, z_2, \ldots, z_k\}$, satisfied $k < m$ and $cov(z_i, z_j) = 0$, through transformation k new variables Z can represent most of the information of m original variables X, as shown in Formula (2):

$$Z = \kappa N dist(X,Y) \qquad (2)$$

In Formula (2),κ is a one $m \times m$ orthogonal matrix and is the covariance matrix of the eigenvalue matrix of data samples $C = \frac{1}{N}\sum_{i=1}^{N}(x_i - u)(x_i - u)^T$, where, $u = \frac{1}{N}\sum_{i=1}^{N} x_i$. Therefore, it is transformed into solving the eigenproblem as shown in Formula (3):

$$\lambda_i = CPZ \qquad (3)$$

In Formula (3), λ_i is characteristic value of C, P is the corresponding eigenvector. Principal component analysis selects several characteristic values with a high contribution rate $\lambda_i$corresponding eigenvector Pas the principal component, to achieve the purpose of dimension reduction. The characteristic contribution rate is shown in Formula (4):

$$\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{p}\lambda_i} = \frac{\sum_{i=1}^{m}\lambda_i}{n} \geq R \qquad (4)$$

In Formula (4), R is the threshold value of the feature contribution rate, feature dimension m is selected according to R to determine the general choice R85%~95%. When using PCA for analysis, different variables in the data often have different dimensions, leading to a large difference in the dispersion of the values of each variable, thus affecting the calculation accuracy. To eliminate the possible impact of different dimensions, the variables need to be standardized first, and then the dimension can be reduced by PCA.

2) Feature selection based on the tabu search algorithm: After the dimensionality reduction of traffic data through the PCA algorithm, feature selection can be done. Tabu Search (TS) algorithm is a heuristic global optimization search method, which obtains the global optimal solution by marking the searched local optimal solution and avoiding repeated search in the iterative calculation [25]. The main idea of the algorithm is first to determine an initial effective solution z, for each solution z define a neighborhood $Y(z)$, determine

several candidate solutions from the neighborhood of the current solution, and select the best candidate solution from them. Selecting the best candidate solution is a search process. To avoid the search process being limited to cycles, TS avoids the local optimization of the search algorithm by constructing a tabu table and defining stop rules. Tabu list before saving n the second taboo length avoids returning to the original solution, thus improving the search ability of the solution space; Stop rule defines that when the optimal solution cannot be improved within several iterations, the algorithm stops. In addition, neighborhood, tabu list, tabu length, amnesty rule and initial solution in the tabu search algorithm will directly affect the search optimization results.

Feature selection based on tabu search is an optimization problem constrained by the objective function, and the appropriate objective function improves the quality of search and optimal feature selection. A good feature solution should guarantee as much classification information as possible on the minimum number of features. In information theory, the greater the information gains of an attribute, the greater the amount of information it contains [26]. Based on the information gained, the classification information of feature vectors can be effectively evaluated. Therefore, this paper selects information gain as the objective function and defines the objective function as shown in Formula (5):

$$G_T = R\sum_{i=1}^{m} C(i) \times \frac{\sum_{j=1}^{n} G(A_j)}{n} \quad (5)$$

In Formula (5), $C(i)$ represents sample i whether it is correctly classified, m is the number of samples; $G(A_j)$ is information gain of features i. Ensure that the maximum classification information is guaranteed with a small number of features through Formula (5), and select divided by n that can ensure faster tabu search speed and avoid overfitting.

The selection of the initial solution in tabu search greatly impacts the effect of tabu search. In the calculation process of other optimal feature selection algorithms, due to the large feature dimension of actual network traffic, it will affect the efficiency of the tabu search algorithm, and feature redundancy will also affect the selection of the optimal feature set. Therefore, the initial solution of tabu search has an important impact on search efficiency and quality.

Generally, the larger the feature, the higher the accuracy of the analysis is. However, in practice, too large a feature space will cause two problems: (1) The huge feature space not only needs higher storage space but also increases the measurement time, which is difficult to apply to real-time traffic analysis; (2) In some applications, such as anomaly detection, service classification, etc., the characterization of different network services requires different feature attribute vectors. If all features are used to represent different service flows, not only the learning effect is reduced, but also the learning time is increased. So, feature selection is to mine the best feature set to describe network traffic; best and tabu search provides a near-optimal solution.

3) Feature selection design based on PCA-TS algorithm: The statistical characteristics of network traffic refer to the characteristics of extracting ports and protocols from the attributes of packets or flows. Such as message length, arrival interval, number of messages, flow duration, number of messages in the flow, etc. Feature vectors represent these statistical characteristics. Such as a network flow F, the characteristic description based on the flow can be expressed as $F = \{y_1, y_2, \ldots, y_n\}$, where y_i represents the value of the feature. The feature set of a flow may contain as many as hundreds of features. Finding a small number of optimal feature subsets to describe the flow is important to improve learning efficiency.

Therefore, this paper makes full use of the feature that PCA can perform fast and effective feature reduction on high-dimensional data, and improves the efficiency of solving the optimal solution of the tabu search method by eliminating feature redundancy and reducing the dimension space. To this end, this paper selects network traffic characteristics by combining PCA and TS algorithms. The flow chart of the PCA-TS feature selection method is shown in Fig. 4.

In Fig. 4, the specific implementation steps of the PCA-TS algorithm are as follows:

1) The tabu table is empty, and initialization parameters are set: tabu length $L_J = 13$, maximum iterations $D_{max} = 600$, maximum improvement times $\bar{D}_{max} = 100$.

2) Use PCA to reduce the original network traffic characteristics and obtain the reduced feature collection $G'_T = \{T_1, T_2, \ldots, T_p\}$, p is the number of feature sets after reduction.

3) To feature set G'_T perform binary coding to obtain the initial solution R_{init}.

4) Set termination conditions, when getting $\bar{D}_{max}$, the search stops; When the best solution cannot be improved by passing R_{init}, stop searching.

5) Judge whether the termination conditions are met. If the termination conditions are met, end the operation and output the optimal flow feature subset. Otherwise, go to the next step.

6) Initial solution R_{init} brings into the neighborhood structure to calculate the neighborhood solution, and the best candidate solution is selected through the objective function.

7) Judge whether the candidate solution meets the amnesty rule. If yes, update the optimal solution in the tabu list and go to step (4), otherwise go to the next step.

8) Calculate the tabu attribute of the candidate solution, select the initial value of the optimal replacement tabu table for non-tabu objects, and go to step (4).

9) End, output the optimal flow characteristic subset G'_R.

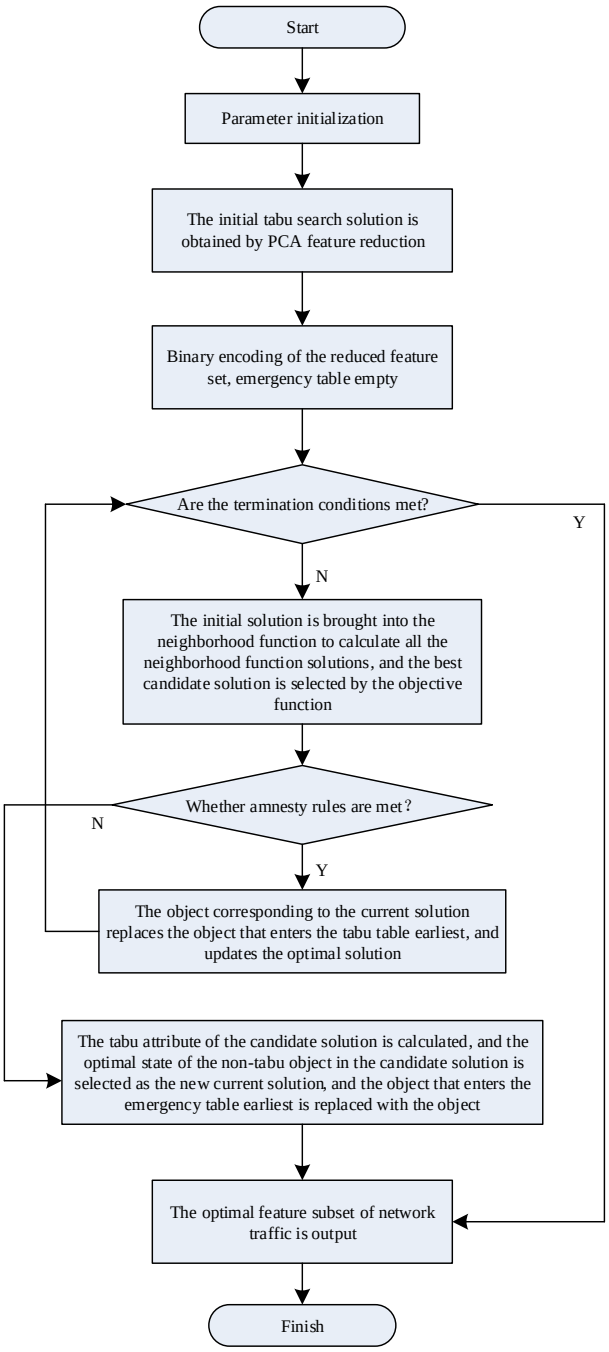


Fig. 4. Process of feature selection method based on PCA-TS.

In Fig. 4, the specific implementation steps of the PCA-TS algorithm are as follows:

1) The tabu table is empty, and initialization parameters are set: tabu length $L_J = 13$, maximum iterations $D_{max} = 600$, maximum improvement times $\bar{D}_{max} = 100$.

2) Use PCA to reduce the original network traffic characteristics and obtain the reduced feature collection $G'_T = \{T_1, T_2, \ldots, T_p\}$, p is the number of feature sets after reduction.

3) To feature set G'_T perform binary coding to obtain the initial solution R_{init}.

4) Set termination conditions, when getting $\bar{D}_{max}$, the search stops; When the best solution cannot be improved by passing R_{init}, stop searching.

5) Judge whether the termination conditions are met. If the termination conditions are met, end the operation and output the optimal flow feature subset. Otherwise, go to the next step.

6) Initial solution R_{init} brings into the neighborhood structure to calculate the neighborhood solution, and the best candidate solution is selected through the objective function.

7) Judge whether the candidate solution meets the amnesty rule. If yes, update the optimal solution in the tabu list and go to step (4), otherwise go to the next step.

8) Calculate the tabu attribute of the candidate solution, select the initial value of the optimal replacement tabu table for non-tabu objects, and go to step (4).

9) End, output the optimal flow characteristic subset G'_R.

E. Abnormal Flow Detection

After selecting network traffic characteristics, you can use the selected traffic characteristics to detect abnormal traffic through the network security architecture abnormal traffic detection module. Improve the detection efficiency of abnormal flow and improve the detection accuracy. This paper uses the SVM algorithm to detect abnormal traffic, assuming there is k type of samples, and then it is necessary to construct k two class classifiers. Each classifier is used to separate one class from the rest. During training, please take one of them as positive, and the rest $k-1$ class is negative. When judging, the sequence of the tested samples passes through k, the total of two class classifiers k output values is $f_i(x) = sgn(g_i(x)), i = 1,2,\ldots,k$. If the decision result contains only one+1, the corresponding classifier's sample class to be detected is the positive class. Suppose there is more than one+1 in the decision result, that is, classification overlap. In that case, it is also necessary to compare the decision function value of the classifier whose output is+1, and the positive class of the classifier with the largest value represents the class of the sample to be detected. If the judgment result is -1, the sample is considered to be indivisible. Therefore, this paper proposes an improved SVM multi-classification algorithm. The idea of class distance in clustering analysis is used as the basis for sorting the second-class classifiers in the detection model.

About k class flow characteristic samples, calculate the center distance from each class to other classes, and then calculate the average distance from each class to other classes. The class with the largest average distance is the class with the most obvious specificity, and such class is preferred as the positive class of the second-class classifier ranking first. The relevant definition of distance is:

Definition 1: Center distance. The center distance of the flow characteristic samples of class i and j is defined as the Euclidean distance in the space of the spherical center that can contain all class i traffic characteristic samples to the spherical center that can contain all samples of class j traffic characteristics, recorded as d_{ij}.

Definition 2: Average distance. The mean distance between the class $i$traffic characteristics and the remaining categories is defined as the mean of the center distance from the class i traffic characteristics to the other samples of the traffic characteristics, recorded as γ_i, and meet:

$$\gamma_i = \left(\frac{G'_R}{k-1}\right)\sum_{j=1}^{k} d_{ij}(i \neq j).$$

The specific implementation steps are as follows:

1) Calculate the center distance of $d_{ij}(i, j = 1,2,\ldots,k, i \neq j)$between various flow characteristic samples and other flow characteristic samples according to definition 1.

2) Calculate the average distance of $\gamma_i(i = 1,2,\ldots,k)$ between various flow characteristic samples and other flow characteristic samples according to definition 2.

3) Compare the size of step $2\gamma_i$, and then follow the categories that numbered in descending orderγ_i.

4) Construct one by one according to the sample number sequence obtained in step (3) the k two class classifiers. The sample with the first number is the positive class of the first second-class classifier, the sample with the second number is the positive class of the second-class classifier, and so on.

During abnormal flow detection, let the sample to be tested pass through each two-class classifier in turn. If the decision result of the sample to be tested in a classifier is+1, then it is determined that the sample is the positive class of the corresponding classifier, and the detection of this sample is terminated. If the output result of samples passing through all the second-class classifiers in turn is - 1, it is determined as unknown flow, added to the set to be verified, and waiting for re training. The sample decision process is shown in Fig. 5.

The distance first SVM multi-classification algorithm can improve the classifier's detection accuracy. Therefore, although k class two classifiers were constructed when building model sets. However, in the improved SVM multi-classification algorithm, when a classifier is judged as+1, the judgment is terminated to shorten the sample detection time.

At this time, according to the specific process of abnormal network traffic detection method, abnormal traffic detection can be realized by the following methods:

1) NetFlow technology captures network traffic packets through flow.

2) The captured data packets are used in the KNN algorithm to filter traffic and eliminate duplicate traffic packets.

3) After data filtering, the traffic packets are input into the PCA-TS algorithm for data dimensionality reduction and feature selection.

4) After obtaining the traffic characteristics, the structure contains N sample a set of traffic, including normal traffic and $N-1$ abnormal flow with obvious difference in the distribution characteristics of three kinds of flow.

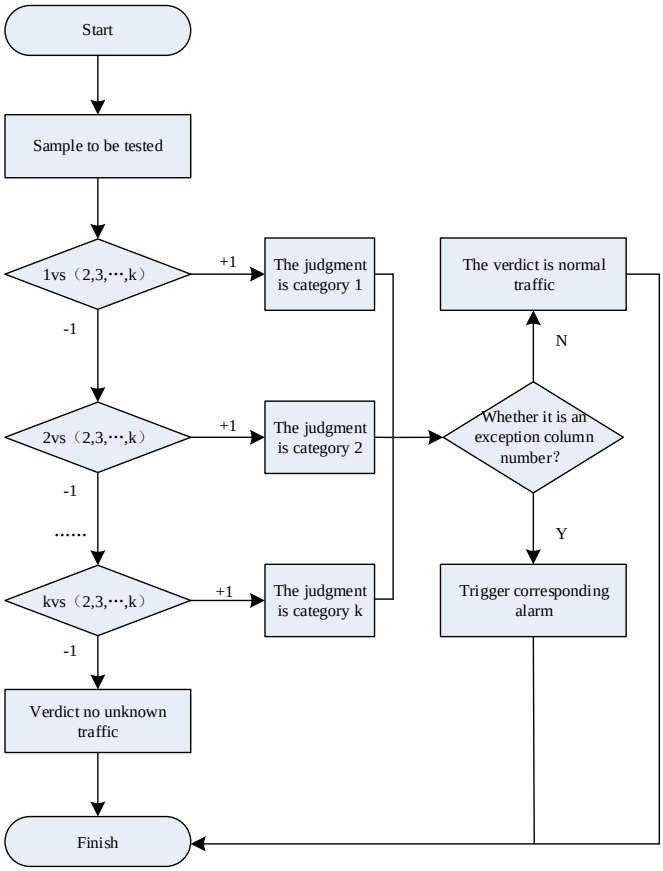


Fig. 5. Sample decision process.

5) The improved SVM multi - classification method is used to input the characteristic samples of the traffic to be measured into the SVM multi-class classifier. If the second-class classifier can recognize the characteristic samples of the traffic to be measured, it is determined that the traffic of the corresponding category is detected. If the detected flow is normal, continue; if it is abnormal, send an alarm. Repeat step (5). If the flow characteristic sample to be detected is determined to be an unknown flow, perform step (6).

6) Add unknown traffic to the collection to be verified. If the traffic in the set to be verified can be clustered and is significantly different from the normal traffic, it can be considered that a new anomaly has occurred.

7) Add new exceptions to the training samples in step (1) for re training to obtain a new model set, and repeat step (5) to achieve network security detection.

III. Experimental Analyses

To verify the effectiveness of the network security detection method in this paper, this paper constructs a simulation experiment through the NS2 simulation platform and shows the network topology in Fig. 6.

In Fig. 6, R1, R2 and R3 are routers, of which R2 is the "key router". The link between R2 and R3 is the bottleneck link, with a bandwidth of 10Mbps and a delay of 30ms. All other links have a bandwidth of 100Mbps and a delay of 15ms. The network contains 25 legitimate TCP connections, 10 of which are background traffic.

Meanwhile, attack parameters are designed: attack cycle is 1s, attack pulse duration is 150ms or 200ms or 250ms, and attack pulse intensity is 30Mbps or 40Mbps. The observation time window WS duration is 90s, and the attack packet types are UDP, ICMP, and invalid TCP.

The experimental data set uses the Honeynet Challenges data set provided by the HoneyPot Project. HoneyPot Project is a non-profit network security research institution committed to studying the latest network attacks and developing open-source security tools to improve the network environment. The organization has volunteers from all over the world. On its official website, many open-source security tools are developed to improve the network security environment. Table I shows some network traffic attributes of the experiment.

Analyze this method's abnormal feature selection ability before abnormal flow detection, and analyze the time required for different abnormal flow feature selections. The analysis results are shown in Table II.

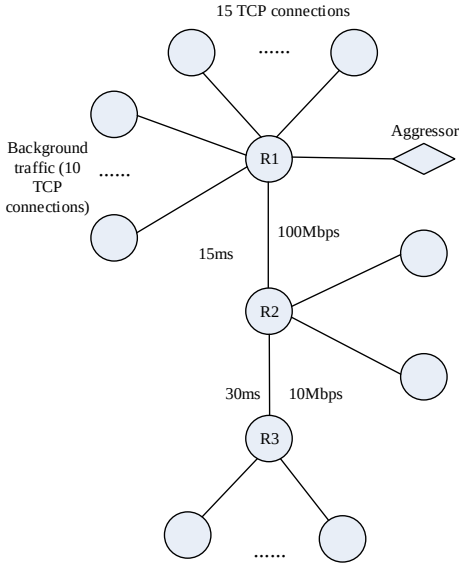


Fig. 6. Topology structure of the experimental network.

TABLE I. Attribute List of Network Traffic Status

Serial number	Symbolic representation	Feature description
1	outPckets	Total number of egress packets
2	outBytes	Total size of the egress packet
3	inPckets	Total number of incoming packets
4	inBytes	Indicates the total packet size of the entry
5	minOutpktLen	Indicates the minimum Byte of the egress packet
6	maxOutpktLen	Indicates the maximum Byte of the egress packet
7	meanLenOutsm	Average Byte of the egress packet
8	minInpktLen	Indicates the maximum Byte of the incoming packet
9	maxInpktLen	Indicates the maximum Byte of the incoming packet
10	meanLenInsm	Average Byte of an incoming packet
11	Outtotalduration	The total duration of the exit flow
12	Outavgduration	The average duration of the exit flow
13	Intotalduration	The total duration of the inlet stream
14	Inavgduration	The average duration of the inlet stream
15	OutavePktPerSecond	Average outbound packet size per second
16	InavelPktPerSecond	The average size of an entry packet in seconds
17	InnerOneToMulty	Indicates the number of one-to-many internal IP addresses
18	OutOneToMulty	Indicates the number of one-to-many external IP addresses
19	stdLenOutqsm	Mean square error of the egress message
20	stdLenInqsm	Mean square error of the incoming message

TABLE II. TIME REQUIRED FORABNORMAL FEATURE SELECTION

Feature sequence number	Feature type	Specific description	Time required for feature selection /s
1	duration	Connection duration	2.4
2	service	The network service type of the target host	2.6
3	src_bytes	Number of bytes of data from the source host to the destination host	2.1
4	land	Determine whether the connection is coming from the same host or port	1.5
5	urgent	Number of urgent packets	1.7
6	num_failed_logins	The number of failed login attempts	1.5
7	num_compromised	Compromised frequency	1.7
8	num_access_files	The number of times the control file was accessed	1.5
9	is_hot_login	Whether the login belongs to the hot list	1.6
10	protocol_tyoe	Protocol type	2.6
11	flag	The connection status is normal or incorrect	1.4
12	dst_bytes	The number of bytes of data from the destination host to the source host	2.1
13	hot	The number of times to access system-sensitive files and directories	2.5
14	wrong_fragment	The number of incorrect segments	1.9
15	dst_host_diff_srv_rate	Among the top 100 connections, the proportion of connections that have different services to the same destination host as the current connection	1.5
16	dst_host_srv_diff_host_rate	Among the top 100 connections, the number of connections that have the same destination host as the current connection and the number of connections that have a different source host from the current connection	1.5
17	srv_count	Number of connections that have the same service as the current connection in the last two seconds	1.6

According to Table II, this method can effectively select multiple features during feature selection to provide reliable data for subsequent network security detection. At the same time, in the feature selection process, the time for this method to realize feature selection does not exceed 3.0s. Therefore, this method has a strong feature selection ability, which provides a reliable guarantee for subsequent abnormal traffic detection.

This paper uses the G score to evaluate this method's abnormal traffic detection effect. G score is defined as:

$$G = \sqrt{precision \times recall} \quad (6)$$

In Formula (6), $precision$ and $recall$ indicate the accuracy rate and recall rate in turn. The higher the G score, the stronger the method's ability to detect abnormal traffic is. Test the G scores of different numbers of packets when they are attacked by different traffic and show the experimental data results through Fig. 7.

As shown in Fig. 7, with the increase in the number of test packets, the G scores obtained by this method begin to decline under attacks of different attack data types. Among them, when being attacked by UDP, the G score obtained by the test is the lowest among the three attack types, but not less than 0.70. It shows that this method can maintain high performance in the detection process. This paper can get a higher G score for detecting ICMP and invalid TCP attacks. It can be seen that this detection method can effectively detect multiple types of attacks.

This study selected three different scenarios for verifying abnormal traffic detection. In scenario B1, there are no attacks in the network, or the attacks present in the network have no direct impact on TCP data traffic. In the B2 scenario, the attacks present in the network have a direct impact on TCP data traffic, but there are no serious attacks. In the C3 scenario, there are serious attacks in the network. Verify the detection effectiveness of our method by analyzing the degree of traffic fluctuations in different scenarios. The analysis results are shown in Fig. 8.

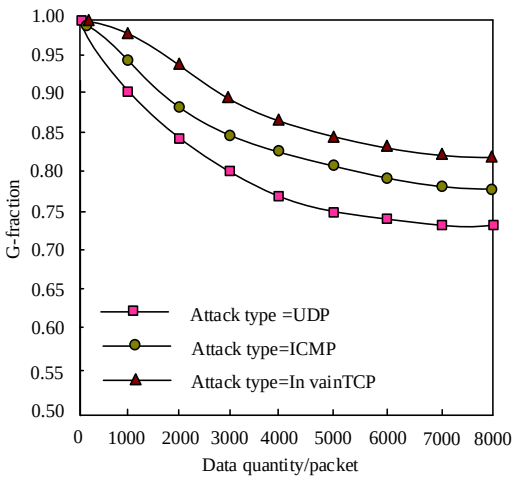


Fig. 7. Analysis of G score test results.

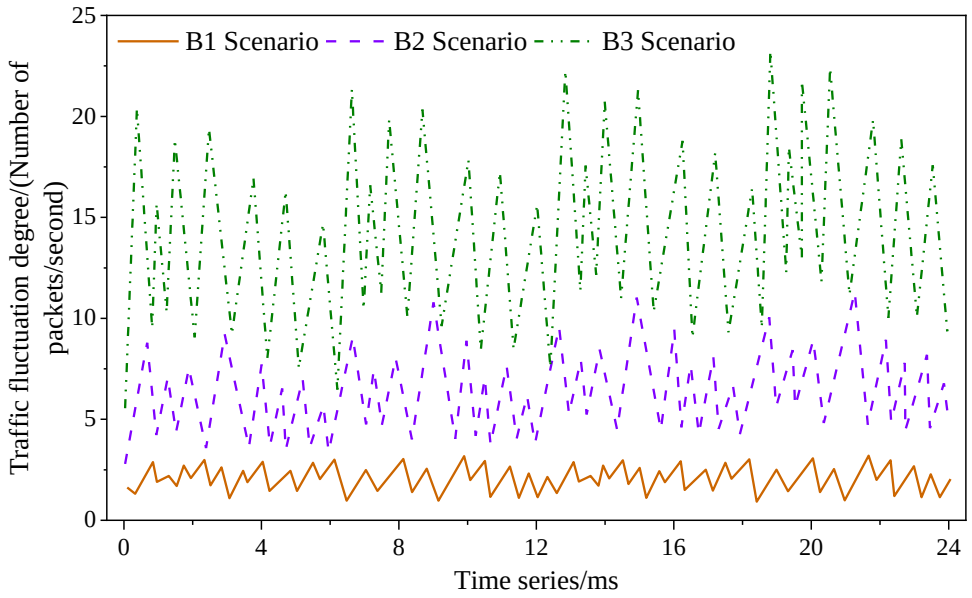
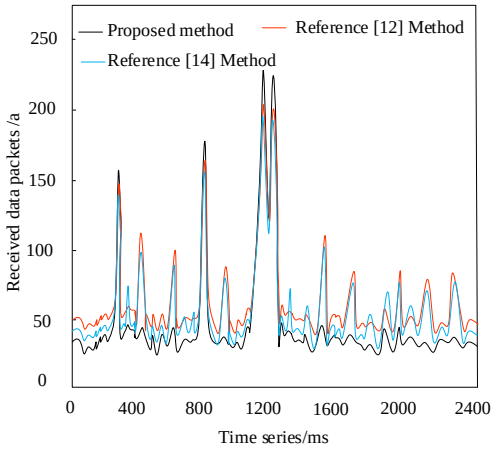


Fig. 8. Analysis of traffic fluctuation degree.

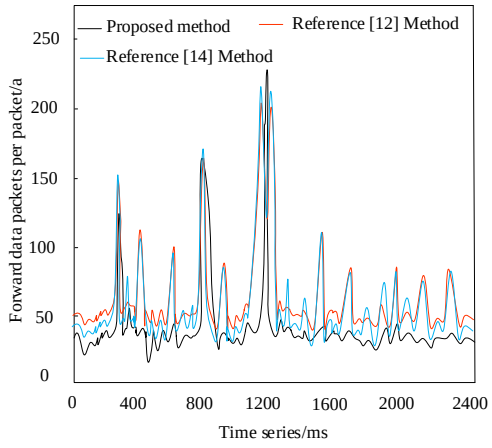
According to Fig. 8, when there is no attack in the network or the attack has no direct impact on TCP data traffic, the traffic always fluctuates below five packets/second, with a small fluctuation degree. When attacks in the network directly impact TCP data traffic, the fluctuation of network traffic increases. When serious attacks exist in the network, the traffic amplitude exceeds 20 packets/second fluctuations. From the above analysis, it can be seen that this method can effectively detect flow fluctuation.

Injecting attack traffic at different times, namely 400s, 800s, and 1200s, analyze the changes in the number of received and forwarded packets detected by the method proposed in this paper, the method proposed in reference [12], and the method proposed in reference [14]. The analysis results are shown in Fig. 9.

As shown in Fig. 9, after applying the method proposed in this article, under normal circumstances, the number of received packets is approximately equal to the number of forwarded packets; After injecting an attack, it will restrict the forwarding of abnormal packets, resulting in a lower number of forwarded packets than received packets. At this point, the number of forwarded packets shows a significant decrease. In response to persistent attacks, the number of packets forwarded by the port gradually decreases until the normal number of forwards is restored. From this, it can be seen that the method proposed in this article has strong ability to detect abnormal traffic and can achieve network security detection. By comparing the methods in reference [12] and reference [14], it can be seen that although the overall trend of the two methods is similar to that of the method in this paper, both methods show abnormal increase or decrease, indicating that the two comparison methods are affected by attacks and have misidentification phenomena..



(a) Analysis of changes in the number of received data packets.



(b) Analysis of changes in the number of forwarded packets.

Fig. 9. Analysis of changes in the number of packets during injection attacks.

IV. CONCLUSION

This paper studies the network security detection method based on abnormal traffic detection, uses this method, and applies this method to the experimental detection process. Experiments show that this method has a good detection effect on the common abnormal traffic and attacks in the network. Given the shortcomings of the current research, the following aspects can be improved in the future research work:

1) Find or build appropriate data sets. The existing real data sets have some shortcomings, lacking real attack data. Most researchers use traditional network data sets for experiments. However, due to the network environment's limitations, the simulation data cannot fully reflect the real network conditions.

2) Accurately identify the types of network attacks and make reasonable solutions. At present, this method can only achieve the detection and early warning of abnormal traffic and cannot achieve the processing of abnormal traffic. In the future, some effective abnormal traffic processing methods can be designed to improve network security.

AVAILABILITY OF DATA AND MATERIALS

The datasets used in this paper are available from the corresponding author upon request.

CONFLICTS OF INTEREST

The authors declared that they have no conflicts of interest regarding this work.

AUTHORSHIP CONTRIBUTION STATEMENT

Tao Xiao: Writing-Original draft preparation

Conceptualization, Supervision, Project administration.

Yang Ke: Language review

Hu YiWen: Methodology

Wang HongYa: Software

REFERENCES

[1] E. Balamurugan, A. Mehbodniya, E. Kariri, K. Yadav, A. Kumar, and M. A. Haq, "Network optimization using defender system in cloud computing security-based intrusion detection system withgame theory deep neural network (IDSGT-DNN)," Pattern Recognit Lett, vol. 156, pp. 142–151, 2022.

[2] C. Ding, Y. Chen, Z. Liu, A. M. Alshehri, and T. Liu, "Fractal characteristics of network traffic and its correlation with network security," Fractals, vol. 30, no. 02, p. 2240067, 2022.

[3] M. A. Ferrag, L. Shu, O. Friha, and X. Yang, "Cyber security intrusion detection for agriculture 4.0: Machine learning-based solutions, datasets, and future directions," IEEE/CAA Journal of Automatica Sinica, vol. 9, no. 3, pp. 407–436, 2021.

[4] S. Pande, A. Khamparia, and D. Gupta, "An intrusion detection system for health-care system using machine and deep learning," World Journal of Engineering, vol. 19, no. 2, pp. 166–174, 2022.

[5] A. K. Bediya and R. Kumar, "A novel intrusion detection system for internet of things network security," in Research Anthology on Convergence of Blockchain, Internet of Things, and Security, IGI Global, 2023, pp. 330–348.

[6] J. Chen and Y. Miao, "Study on network security intrusion target detection method in big data environment," International journal of internet protocol technology, vol. 14, no. 4, pp. 240–247, 2021.

[7] Y. Wang et al., "Exhaustive research on the application of intrusion detection technology in computer network security in sensor networks," J Sens, vol. 2021, pp. 1–11, 2021.

[8] Q. Ding and J. Li, "AnoGLA: An efficient scheme to improve network anomaly detection," Journal of Information Security and Applications, vol. 66, p. 103149, 2022.

[9] C. H. Nwokoye, I. I. Umeh, N. N. Mbeledogu, and V. O. S. Okeke, "Scan-Based Worms: The Impact of IPV4 Address Space on Epidemic Computer Network Models.," Engineering Letters, vol. 29, no. 2, 2021.

[10] C. Do Xuan, "Detecting APT attacks based on network traffic using machine learning," Journal of Web Engineering, pp. 171–190, 2021.

[11] L. Duan, J. Zhou, Y. Wu, and W. Xu, "A novel and highly efficient botnet detection algorithm based on network traffic analysis of smart systems," Int J Distrib Sens Netw, vol. 18, no. 3, p. 15501477211049910, 2022.

[12] M. Woźniak, J. Siłka, M. Wieczorek, and M. Alrashoud, "Recurrent neural network model for IoT and networking malware threat detection," IEEE Trans Industr Inform, vol. 17, no. 8, pp. 5583–5594, 2020.

[13] P. Steno, A. Alsadoon, P. W. C. Prasad, T. Al-Dala'in, and O. H. Alsadoon, "A novel enhanced region proposal network and modified loss function: threat object detection in secure screening using deep learning," J Supercomput, vol. 77, pp. 3840–3869, 2021.

[14] T. Gaber, A. El-Ghamry, and A. E. Hassanien, "Injection attack detection using machine learning for smart IoT applications," Physical Communication, vol. 52, p. 101685, 2022.

[15] K. Lin, X. Xu, and F. Xiao, "MFFusion: A multi-level features fusion model for malicious traffic detection based on deep learning," Computer Networks, vol. 202, p. 108658, 2022.

[16] G. Xie, Q. Li, and Y. Jiang, "Self-attentive deep learning method for online traffic classification and its interpretability," Computer Networks, vol. 196, p. 108267, 2021.

[17] A. Lung-Yut-Fong, C. Lévy-Leduc, and O. Cappé, "Distributed detection/localization of change-points in high-dimensional network traffic data," Stat Comput, vol. 22, no. 2, pp. 485–496, 2012.

[18] V. Mic and P. Zezula, "Data-dependent metric filtering," Inf Syst, vol. 108, p. 101980, 2022.

[19] B. H. Meyer, A. T. R. Pozo, and W. M. N. Zola, "Improving barnes-hut t-sne algorithm in modern gpu architectures with random forest knn and simulated wide-warp," ACM Journal on Emerging Technologies in Computing Systems (JETC), vol. 17, no. 4, pp. 1–26, 2021.

[20] L. Guo, S. Wang, J. Yin, Y. Wang, J. Yang, and G. Gui, "Federated user activity analysis via network traffic and deep neural network in mobile wireless networks," Physical Communication, vol. 48, p. 101438, 2021.

[21] S. Yang, L. Ning, X. Cai, and M. Liu, "Dynamic spatiotemporal causality analysis for network traffic flow based on transfer entropy and sliding window approach," J Adv Transp, vol. 2021, pp. 1–17, 2021.

[22] M. Li, Y. Liu, Q. Zheng, W. Qin, and X. Ren, "Stable feature selection based on brain storm optimisation for high-dimensional data," Electron Lett, vol. 58, no. 1, pp. 10–12, 2022.

[23] L. Luo et al., "Adaptive data dimensionality reduction for chemical process modeling based on the information criterion related to data association and redundancy," Ind Eng Chem Res, vol. 61, no. 2, pp. 1148–1166, 2022.

[24] J. Ren, H. Wang, K. Luo, and J. Fan, "A Priori Modeling of NO Formation with Principal Component Analysis and the Convolutional Neural Network in the Context of Large Eddy Simulation," Energy & Fuels, vol. 35, no. 24, pp. 20272–20283, 2021.

[25] W. Chang et al., "Prediction of hypertension outcomes based on gain sequence forward tabu search feature selection and xgboost," Diagnostics, vol. 11, no. 5, p. 792, 2021.

[26] Z. Zhang, H. Tang, and Z. Xu, "Fatigue database of complex metallic alloys," Sci Data, vol. 10, no. 1, p. 447, 2023.