IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 June 2024, accepted 8 July 2024, date of publication 19 August 2024, date of current version 30 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3445996

RESEARCH ARTICLE

Construction of Internet Traffic Monitoring Model Based on Improved Transformer Algorithm

JIEJING LIU[1], XU LIU[2], YANHAI WANG[3], AND HUA FU[4]
[1]College of Mathematics and Computer Science, Hengshui University, Hengshui 053010, China
[2]Office of Academic Affairs, Hebei University of Engineering, Handan 056038, China
[3]Modern Education Technology Center, Hebei University of Engineering, Handan 056038, China
[4]China Telecom Corporation Ltd., Handan Branch, Handan 056008, China

Corresponding author: Xu Liu (runninglx@163.com)

ABSTRACT With the popularization of the Internet, it is very important to effectively identify abnormal behaviors in network traffic. This study focuses on the construction of an internet traffic monitoring model, aiming to improve the accurate recognition rate of abnormal behavior and reduce information loss during small block segmentation. To this end, a internet traffic monitoring algorithm based on the improved Transformer is optimized. This model adopts a block segmentation algorithm that preserves important information during the segmentation process, thereby enhancing the segmentation quality and accuracy of the model. By effectively interacting with multiple receptive field information, the model reduces information loss and improves accuracy and efficiency. After experimental verification, the model performed well on CICIDS sample data, with an F1 value of 93% for normal internet traffic. The F1 value of internet attack traffic was 91%. Compared with the original Transformer model, it increased by 5% and 2.4%, respectively. On the NSLKDD sample, the improved algorithm proposed in the study had an area under the curve value of 0.90, which outperformed other models. This proves that it has significant advantages in the dual classification task of internet traffic anomaly monitoring. This study provides an effective deep learning algorithm for internet traffic anomaly monitoring, which is expected to provide strong support for network security assurance in practical application scenarios.

INDEX TERMS Deep learning, transformer, internet, flow rate, block segmentation, monitoring, Trans-M.

I. INTRODUCTION

The openness of the internet provides conditions for illegal actors. Conducting illegal activities through the internet not only endangers personal information security and property, but also poses a serious threat to social stability. In recent years, network security incidents have occurred frequently, causing serious damage to the network environment and order. This also prompts experts and scholars to search for more effective network traffic monitoring methods [1], [2], [3].

Internet traffic monitoring is the basis of digital services, such as online transactions, short-term video applications, online reservation providers and electronic ordering systems, and plays a vital role in ensuring the security and stable operation of these services [4], [5], [6]. Existing Internet traffic monitoring technologies face several limitations. Firstly, with the rapid growth of network traffic, traditional traffic monitoring systems may lack efficiency and accuracy in processing large-scale data. In addition, existing monitoring methods may suffer from undefined spots and high false alarm rates in detecting complex and diverse network attack methods. Therefore, researchers urgently need to develop more efficient, accurate, and robust traffic monitoring technologies to cope with the constantly evolving network threat environment. In this context, as a deep learning model in the field of computing, Transformer has good feature capture and adaptability, which can effectively handle large-scale parallel computing tasks. Therefore, it has applicability in the traffic monitoring. However, despite its significant advantages

The associate editor coordinating the review of this manuscript and approving it for publication was Walid Al-Hussaibi.

in capturing data features, Transformers still need further optimization to address high noise levels and complex attack behaviors.

Therefore, in order to provide a more efficient, accurate and reliable Internet traffic anomaly monitoring scheme to cope with the evolving network security challenges, this research proposes an improved Transformer-based Internet traffic monitoring algorithm Transformer Multi-Receive Field Fusion (Trans-M). This model is based on the optimized Transformer algorithm, which is specially designed for Internet traffic anomaly monitoring. Considering the shortcomings of traditional Transformer model in extracting local feature information, it is combined with extended convolutional units to better capture global and local information. In addition, by applying expansion convolutions with different unfolding rates, the model can perceive multiple receptive field regions, further enhancing the information perception ability. When dealing with block segmentation, a block segmentation algorithm is adopted, which can calculate and allocate appropriate weights to the information in the receptive field area. To reduce information loss during block segmentation, a decoder-based Multi-Receptive Field Fusion (MRFF) algorithm is also introduced and combined with self attention mechanism to reduce information loss.

The main contribution of the research is that it has brought significant positive impacts to the field of Internet traffic monitoring. By addressing the efficiency and accuracy issues of existing technologies in large-scale data processing and complex network attack detection, the research aims to significantly enhance the security and stability of the system. Furthermore, improved technology is expected to reduce false alarm rates and enhance robustness against high noise and complex attack behaviors, effectively protecting personal information and property security. The research results aim to improve the overall effect of Internet traffic monitoring, provide a solid guarantee for the safe operation of digital services such as online transactions, short-term video applications, online bookings, and promote the healthy and stable development of the network environment.

The overall framework of the study can be divided into four parts. The first part summarizes the achievements and shortcomings of research on Transformer and Internet traffic both domestically and internationally. The second part introduces Transformer and block segmentation models. Based on this, relevant improvements have been made to the Trans-M model. The third part conducts hyper-parameter and dual classification experimental analysis. The fourth part summarizes the research findings and points out the directions for further research.

II. RELATED WORKS

A. RESEARCH STATUS OF INTERNET TRAFFIC MONITORING TECHNOLOGY

Internet traffic monitoring provides critical support for network security and performance management. With the continuous progress of network technology and machine learning, many scientists and scholars have conducted research on the Internet traffic monitoring [7]. Duan et al. proposed an innovative method to efficiently monitor zombie networks by combining auto-encoder neural networks with Decision Trees Model (DT). This method used deep current monitoring and statistical analysis for feature selection, accurately characterizing the communication behavior between nodes. Self encoder was used for feature filtering and optimizing model construction. By generating a few samples and enhancing the DT with improved gradients, class balancing was achieved and botnet data was accurately monitored. The experimental results showed that this method performed superior in network traffic [8]. Zhang and Wang proposed that Internet traffic classification was crucial for multiple network activities. Due to the inability of traditional methods to adapt to the increasing demand for encryption, machine learning methods are gradually increasing. Although net flow is widely used by network operators, the application in network traffic classification is still immature. Combining net flow data with deep neural networks, an effective internet traffic classification module was proposed. The performance was verified on two real datasets [9]. Ponnusamy et al. believed that the lack of initial defense at the network or node level allowed attackers to launch attacks. The lack of readily available benchmark data for internet traffic added challenges. The study explored the characteristics of existing datasets and their applicability in traffic, and analyzed the characteristics of wireless network packets. The dynamic weight allocation improved threat classification. Combining domain heuristic methods and early classification results, 19 high information gain wireless network specific fields were identified as ML features [10]. Li et al. proposed that back-pressure control was initially applied to communication networks with packet routing. After multiple modifications, it was adapted to flow control and achieved satisfactory results. Most BP variants were based on the assumption that they could fully understand the network flow and traffic conditions, especially the queue length. However, it was actually difficult to obtain accurate queue length information. Compared with the original BP and other controllers, the experimental results showed that even with only 10% of the traffic, the average delay, throughput, and maximum parking queue length performed well in high demand scenarios [11].

Zanma et al. used a discrete homogeneous Markov chain to represent the multi-time varying network traffic state. Based on lost historical data, the probability matrix of Markov chains was used to estimate the network traffic state online. Then, based on the estimated network traffic status, the appropriate controllers were selected to optimize control performance. This method was validated through simulation and experiments, demonstrating the effectiveness in real-time networked control systems [12]. Zanma et al. proposed a progressive network traffic collective anomaly detection framework called CCAD, which was based on clustering methods. CCAD helped analysts effectively identify

collective anomalies in network traffic. Its working method was different from other anomaly monitoring methods, mainly by analyzing the impact of collective anomalies on the clustering results of network traffic data. Experiments showed that CCAD performed well in collective anomaly exploration. The monitoring rate was significantly improved compared with other existing methods [13]. Dong and Xia explored the impact of packet sampling on recognition accuracy. The research involved feature selection, behavior measurement correlation analysis, and traffic recognition algorithms. Through experiments, under grouped sampling, the importance of behavioral feature decreased. As long as the traffic is sufficient, feature selection is independent of sampling ratio. High sampling rate leads to a decrease in accuracy. To improve the accuracy, a deep belief network recognition method was introduced, demonstrating better performance than other methods [14]. Vidhya and Nagarajan proposed a progressive intrusion detection method based on machine learning technology, aiming to intelligently and effectively identify wireless network traffic intrusions without the need for additional hardware. The function of Intrusion Detection System (IDS) was to detect and prevent network intrusions, guarantee the security of user connections, and ensure the confidentiality and integrity of the network. Compared with existing methods, this method exhibited higher intrusion detection accuracy [15].

B. CURRENT STATUS OF ATTENTION MECHANISMS AND TRANSFER LEARNING RESEARCH

Many researchers have also contributed to attention mechanisms and transfer learning. Peng et al. proposed a method aimed at solving the graph neural network text classification models being unable to capture distant node information and reflect various scale features of text. This method introduced attention mechanism in Dense Connected Graph Convolutional Neural Network (DC-GCN). DC-GCN utilized dense connections to collect information from distant nodes, achieving small-scale feature multiplexing and generating features with different scales. Finally, by combining attention mechanisms with features, the relative importance was determined. The experimental results showed that the method performed excellently on four benchmark datasets [16]. Scholars such as Zulqarnain proposed a Bimodal GRU (TS-GRU) method based on feature attention mechanism, aimed at solving the emotion polarity recognition and classification in sentiment analysis. This method integrated pre-feature attention mechanism, combined sentence order modeling and word feature capture to model complex relationships between words, and utilized attention layers to capture the emotional polarity of keywords. The experimental results showed that the proposed method was effective in IMDB. Emotional analysis accuracy of 90.85%, 80.72%, and 86.51% were obtained on the MR and SST datasets, respectively [17]. Lauren et al. proposed a method based on deep transfer learning aimed at addressing the demand for large amounts of manually annotated data in supervised machine learning. This method utilized pre-trained language models to accumulate prior knowledge. After fine-tuning the model, it was transferred to specific political text tasks. The experimental results showed that for eight tasks, the model using transfer learning improved performance by 10.7% to 18.3% compared with the classical model [18]. F. Ullah and other researchers proposed a Transformer-based imbalanced network traffic transfer learning intrusion detection system, aimed at solving the feature complexity and data imbalance in network intrusion detection. This method utilized transformer-based transfer learning to learn network feature representation and feature interaction, and combined synthetic minority over-sampling techniques to balance abnormal traffic. The model exhibited good performance, providing an effective intrusion detection solution for the network security [19]. Fateh et al. proposed a comprehensive solution for multi-lingual handwritten digit recognition based on attention mechanism and transfer learning. This method utilized transfer learning to reduce computational costs while maintaining image quality and recognition accuracy, and introduced the innovative MRA module for feature extraction. Experimental results showed that this module significantly improved image quality and handwritten digit recognition accuracy, resulting in a nearly 2% increase in recognition accuracy for specific languages [20].

Table 1 shows a list of the relevant work proposed.

To sum up, the existing research has made significant progress in improving the accuracy and efficiency of Internet traffic monitoring. Among them, methods such as combining auto-encoder neural networks with decision trees, efficient classification and detection, etc. have demonstrated good performance. However, these methods still have limitations in dealing with large-scale data, high noise levels, and complex attack behaviors, with high false alarm rates and detection undefined spots. In this regard, the study proposes the Trans-M, which optimizes feature extraction methods and combines them with self attention mechanisms to enhance the ability to capture local and global information, effectively addressing the aforementioned limitations. The contribution of the research is that it provides an efficient, accurate and robust solution to improve the performance of Internet traffic anomaly monitoring, further expanding the research and application scope of Internet traffic monitoring.

III. INTERNET TRAFFIC MONITORING MODEL BASED ON IMPROVED TRANSFORMER

This study proposes an Internet traffic monitoring model based on an improved Transformer. The main goal is to strengthen the monitoring capacity of Internet traffic to better adapt to practical applications. Firstly, data blocks are segmented using the Transformer and block segmentation techniques. A multi-perception domain fusion algorithm is introduced to address the potential structural losses that may occur during block cutting. Finally, the MRFF algorithm is

TABLE 1. Related worksheet.

References	Main research content	The gap with the method in this article
L. Duan [8]	Efficient monitoring of zombie networks, combined with auto-encoder neural networks and decision trees for deep flow monitoring and statistical analysis	There are shortcomings in local feature extraction, and unable to handle high noise levels
L. Zhang [9]	Internet traffic classification using net flow data and deep neural network	Insufficient efficiency in large-scale data processing and ability to respond to complex attack behaviors
V. Ponnusamy [10]	The improvement of threat classification using existing datasets and dynamic weight allocation	The efficiency of processing large-scale data needs to be improved, and there is a lack of ability to respond to diverse attack behaviors
L. Li [11]	Improving flow control by modifying the back pressure control method and comparing experimental results of different controllers	Inadequate adaptability to complex traffic patterns
T. Zanma [12]	Using Markov chains to estimate network traffic status and optimize control performance	Unresolved information loss problem, lack of effective capture of local and global features
C. Wang [13]	Proposing a CCAD network traffic collective anomaly detection framework to improve monitoring rate	There is no optimization for high noise data, and the overall robustness needs to be enhanced
S. Dong [14]	Exploring the impact of packet sampling on the accuracy of traffic recognition and introduce deep belief networks for traffic recognition	There are issues with detection blind spots and high false alarm rates when dealing with complex and diverse attack behaviors
G. S. Vidhya [15]	Proposing a progressive intrusion detection method based on machine learning technology to intelligently identify wireless network traffic intrusions	Lack of strong robustness against multiple attack behaviors and high noise levels
Y. Peng [16]	Introducing attention mechanism to solve the problem of graph neural network text classification not being able to capture information from distant nodes	The application field is limited and has not been verified in the field of Internet traffic monitoring
M. Zulqarnain [17]	A Bimodal GRU Method Based on Feature Attention Mechanism to Solve the Problem of Emotional Polarity Recognition and Classification	Limited to emotion analysis, not verified on Internet traffic data
M. Laurer [18]	Proposing a deep transfer learning method that utilizes pre-trained language models for task specific transfer	The method is aimed at language data and lacks applicability and verification in Internet traffic monitoring
F. Ullah [19]	Transformer-based imbalanced network traffic transfer learning intrusion detection system, solving the problems of feature complexity and data imbalance	Insufficient ability to extract local feature information and cope with high noise levels
A. Fateh [20]	A multi-lingual handwritten digit recognition method based on attention mechanism and transfer learning, reducing computational costs and maintaining high recognition accuracy	The research focuses on image processing and recognition, but not on the empirical analysis of Internet traffic monitoring

combined with Transformer encoder to analyze and model Internet traffic data.

A. TRANSFORMER MODEL AND BLOCK SEGMENTATION CONSTRUCTION

With the rapid development of information technology, Internet traffic monitoring is crucial to ensure network security and performance management. Although existing technologies have real-time monitoring and anomaly detection capabilities, they still have limitations in handling large-scale data and complex attack behaviors. This section mainly introduces block segmentation technology into the Transformer model to reduce the information loss, retain more useful information, and enhance the quality and robustness of features. The Transformer model relies entirely on attention mechanisms to perform parallel computing. Previously, recurrent neural networks are mainly used for natural language processing and other sequential tasks. Processing sequence data is the key to deep learning. Usually, it is necessary to focus on the key parts of the input sequence. The attention mechanism allows the network to automatically focus on the parts related to the current position when processing different positions in the sequence. Adjusting weight can improve flexibility and adaptability [21], [22]. By assigning greater weights to important features, the attention mechanism enhances the network expression and generalization capabilities, as shown in Figure 1.

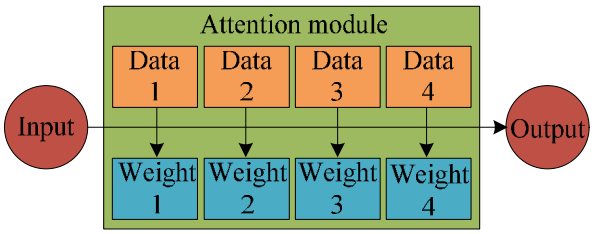

FIGURE 1. Attention mechanism diagram.

The Transformer model consists of an encoder and a decoder. Each part is composed of multiple layers of heap. Each layer includes a multi-head self attention sub-layer and a fully connected feed-forward network sub-layer. The decryption layer has an additional mask multi-head attention sub-layer compared with the encryption layer. The specific structure is shown in Figure 2.

The Transformer's multi-head attention mechanism parallelly processes each position of the input sequence. Multiple headers are used to learn different queries, keys, and value mappings to capture different semantic information. Masked multi-head attention blocks future location information during computation to prevent information leakage. The main difference between it and conventional long headed attention lies in the way it processes future information. The feed-forward neural network is the core component of the Transformer, which performs nonlinear transformation on the hidden layer representation of each position. The output of multi-head attention is mapped to another dimension space. It enhances the model's expressive power and predictive accuracy. Residual connection adds input directly

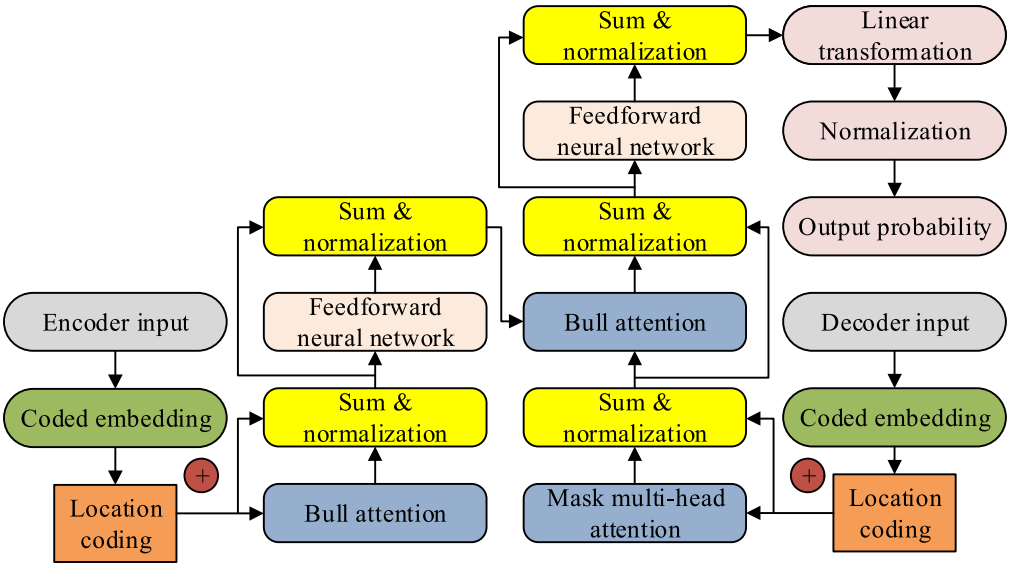

FIGURE 2. Diagram of encoder and decoder for transformer model.

to output, helping the model learn residual [23], [24], [25]. In Transformer, each sub-layer has residual connections. This design helps to avoid gradient vanishing and gradient explosion problems in deep networks. Layer normalization is used to standardize the input of each layer, which can accelerate training and improve the generalization ability and performance, as shown in equation (1).

$$\begin{cases} \text{LayerNorm}(X) = a \otimes \dfrac{X-\mu}{\sigma} + \beta \\ \mu = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i \\ \sigma = \sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2 + \varepsilon} \end{cases} \tag{1}$$

In equation (1), a and β are trainable parameter vectors. The dimension matches the final dimension of X. $\otimes$ represents element level multiplication. ε is a small constant used to ensure numerical stability. Compared with traditional deep learning models such as Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM), Transformer is suitable for handling internet traffic anomaly monitoring tasks due to the parallel computing advantages [26], [27]. This study draws on successful experiences in the natural language processing. The improved Transformer model is used to Internet traffic monitoring. Among them, the formation of convolutional attention module significantly improves the performance of the model by using attention mechanisms in image channels and feature dimensions. Transformer performs well in handling long-term dependencies of sequences, but the local feature information processing is limited. Therefore, it is improved. Based on the block segmentation technology, the information loss is reduced. More useful information is retained to enhance the quality and robustness of features. Figure 3 shows the model structure.

The model consists of four modules, including preprocessing and sequence image conversion, expansion convolution unit, block embedding unit, and encoding unit. The pre-processing stage organizes and standardizes the gaps and outliers in Internet traffic data. The sequence image conversion module converts the data sequence into a rectangular image and transmits it to an expansion convolution unit, which performs expansion convolution based on the expansion rate to extract feature information. The block embedding unit is composed of block segmentation and full connectivity. It is segmented using a block segmentation algorithm to increase the number of channels. The encoding unit integrates MRFF, multi-head attention, feed-forward network, and multi-receptive field information to process global key features, and map the results. In the pre-processing stage, the gaps and outliers in the data set are properly handled to ensure the accuracy of the model. The outliers refer to the Gaussian distribution theory, as shown in equation (2).

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{2}$$

In equation (2), σ represents the standard deviation of the data set. μ is the average of the data set. X refers to a specific data point. The data points of 99% to 100% are located in the $(\mu - 3\sigma, \mu + 3\sigma)$. The distance between the data points and the average is measured. If this distance exceeds three times the standard deviation, it can be confirmed that the data point is abnormal. The strategies for handling null values include direct discarding, interpolation, and the K-Nearest Neighbor (KNN) algorithm. The direct discarding method may result in missing information. The interpolation method calculates missing values based on neighboring values. The KNN rule

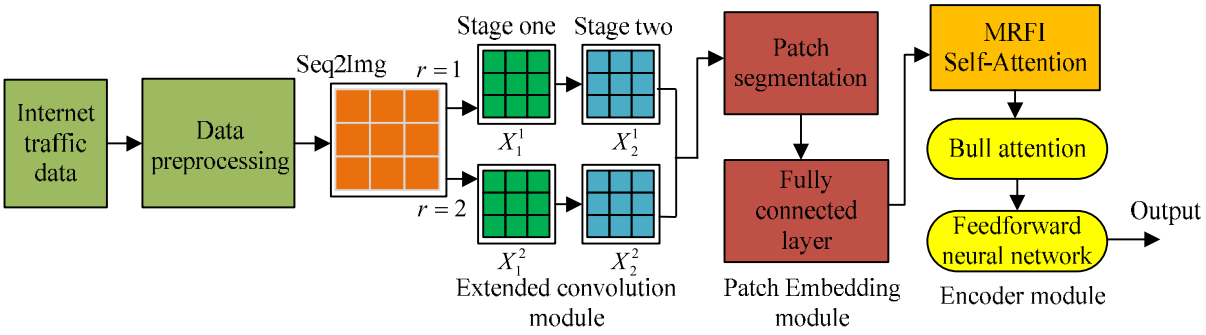

FIGURE 3. Trans-M structure diagram.

searches for K data points that are most similar to missing values for filling. The choice of method depends on the nature of the data. Both CICIDS and NSLKDD Internet traffic sample sets in the study have missing values. The CICIDS data set has fewer missing values. It has a large amount of data, approximately 2.84 million pieces. Therefore, data containing missing values is directly discarded. Relatively speaking, the NSLKDD data set has fewer missing values. However, due to the small amount of data, about 150000 pieces, each data point is extremely important. Therefore, the mean filling method is used to handle missing values. To accelerate model training, the data needs to undergo range reduction processing. The normalization method is used in the study, as shown in equation (3).

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3}$$

In equation (3), max and min represent the maximum and minimum values of the sample data, respectively. The input phase organizes the raw data into rectangles. Figure 4 is a schematic diagram of the sequence image conversion structure. This Figure shows the input process of Internet traffic data $\{x_1, x_2, \cdots, x_t\}$. It is converted into a rectangle with a width of W and a height of H. The H and W are calculated based on the given data sequence length t. If the length t of the data sequence is n times the width W, the data can be perfectly converted into a rectangle. If it is not an integer multiple, the conversion is achieved by adding zero at the end of the sequence. This transition from sequence to matrix adjusts the data structure to fit the input requirements of the model.

The pre-processed data is converted through a sequence image conversion module. The dimension becomes $W \times H \times C$. To achieve information complementarity between different receptive fields, the converted data is input into an expanding convolution unit for expanding convolution operations. The output data of stage one is denoted as $X_1^1 X_1^2 \in R^{W \times H \times C}$, while the other is denoted as $X_2^1 X_2^2 \in R^{W \times H \times C}$. X_m^r represents stage m. The structure of the extended convolutional unit is shown in Figure 5.

In Figure 5, 1×1 Conv and 3×3 Conv refer to the size of the convolutional kernel, respectively. H-swish and Softmax are activation functions. The data is first processed

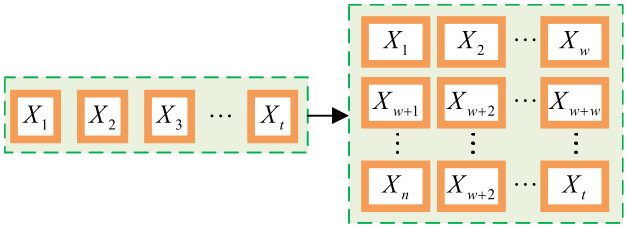

FIGURE 4. Schematic diagram of sequence image conversion structure.

by 1×1 Conv, then activated by H-swish activation function, and followed by 3×3 Conv and H-swish activation again. Finally, it activated by 1×1 Conv and Softmax. The convolutional module also integrates a residual network to prevent gradient vanishing. The specific forms of H-swish and Softmax activation functions are shown in equation (4).

$$\begin{cases} \mathrm{H-swish} = x \dfrac{\min(\max(x+3, 0), 6)}{6} \\ s_t = \mathrm{Soft\,max}(\sigma_t) = \dfrac{e^{-\sigma_t}}{\sum_{t=1}^{w} e^{-\sigma_t}} = [s_1, s_2, \cdots, s_w]_{w \times 1} \end{cases} \tag{4}$$

In equation (4), σ represents the input sequence. w is the length of the marked sequence. st is the calculated value obtained. The Softmax function allocates spatial attention weights based on the size of each element in the input sequence and ensures the sum of weights. The advantage of Softmax lies in the computational simplicity in the gradient descent optimization process.

B. CONSTRUCTION OF TRANS-M ALGORITHM BASED ON BOCK SEGMENTATION AND ENCODER

The Transformer algorithm divides data into several blocks through block segmentation and generates block vectors to enhance feature representation. Then, the block vectors are merged through a fully connected layer to fuse multi-domain information. On this basis, multi-head attention mechanism and layer normalization technology are further introduced into the encoder module to improve the modeling ability for long-distance dependencies. Finally, the MRFF method

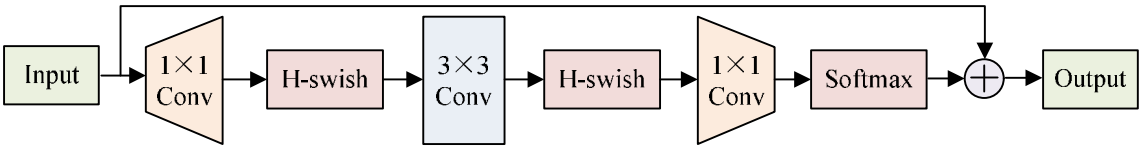

FIGURE 5. Structure diagram of the extended convolution unit.

and the multi-head attention mechanism are combined to accurately monitor Internet traffic data.

In the block segmentation step, block segmentation is used to convert the divided blocks into block vectors. In the subsequent fully connected stage, these block vectors undergo linear operations for feature merging to form the final block embedding. When performing block segmentation, data X_1^1 and X_2^1 are divided into several non-overlapping blocks. Each block size is S. Subsequently, block segmentation is used to calculate the vector representation of each block. The specific calculation process is shown in equation (5).

$$\begin{cases} p_{c,i}^r = \sum_{s=1}^{s^2} \frac{\exp(a_{c,i,s}^r) a_{c,i,s}^r}{\sum_{s'=1}^{S^2} \exp(a_{c,i,s'}^r)} \\ p_i^r = \left[p_{1,i}^r, p_{2,i}^r, \cdots, p_{2C,i}^r\right]. \end{cases} \tag{5}$$

In equation (5), $p_{c,i}^r \in R^{2C}$ describes the block segmentation calculation results of the i-th block in the channel numbered c at the expansion rate r. c is an integer from 1 to C. $a_{c,i,s}^r$ refers to the s-th position of the i-th block in the channel numbered c at the expansion rate r. p_i^r represents the value of the i-th block at the expansion rate r. Next, the block vector inputs a fully connected layer network with an output dimension of $4C$. $FC(4C)$ represents passing through the fully connected layer. The number of channels becomes $4C$. The output of $4C$ after passing through the fully connected layer is shown in equation (6).

$$P^r = \mathrm{FC}_{(4C)}(p^r) \tag{6}$$

In equation (6), the vector under the expansion rate r represents the characteristics of the block. Based on flexible pooling techniques, block segmentation algorithms can flexibly process different receptive domain information, more effectively preserve channel features, and improve the accuracy of block segmentation. Although the Transformer algorithm can effectively model long-distance dependencies of feature information, the block segmentation process may lose some organizational structure information. Integrating blocks with different receptive domains can alleviate this problem. Therefore, it is crucial to find effective methods to integrate block feature information from different receptive domains. To address this issue, a multi-sensory domain feature fusion algorithm based on the encoder is proposed. Multi-sensory domain feature fusion can interact with information from multiple sensory domains while modeling long-distance dependencies of input information, thereby reducing the Internet traffic information loss. The encoder module includes multi head attention mechanism, feed-forward network, and uncertainty modeling method. The study selected Bayesian statistics as the uncertainty modeling method, which combines prior knowledge and observation data, and uses Bayesian theorem to update the posterior distribution of model parameters. Bayesian statistical methods provide quantification of parameter uncertainty through a posterior distribution, which helps to consider model uncertainty in the decision-making process. In these two parts, residual networks and layer normalization techniques are introduced. The multi-head attention mechanism is an enhanced version of the self attention mechanism. It can calculate contextual information for each position in the input sequence and includes multiple heads that can calculate attention in parallel. Each head learns different queries, keys, and value mappings to capture different feature information [28], [29]. Figure 6 shows the structure of the self attention mechanism for multi-sensory domain feature fusion.

The comprehensive correlation between blocks i and j is generated by the synthesis of corresponding position information within each receptive field, as shown in equation (7).

$$\omega_{ij} = \frac{1}{\sqrt{d_{\mathrm{mod}\ el}}}((P_i^1 W^Q)(P_j^1 W^K)^T + (P_i^2 U^Q)(P_i^2 U^K)^T) \tag{7}$$

In equation (7), $d_{\mathrm{mod}\ el}$ represents the dimensions in the hidden layer of the model. When the expansion ratio is 1, the dot product output of P_i is shown in the first term of equation (8). When the expansion ratio is 2, the dot product output of P_i is shown in the second item of equation (8).

$$\begin{cases} \beta_i^1 = \sum_{j=1}^{n1} \frac{\exp(\omega_{ij})}{\sum_{j'=1}^{n1} \exp(\omega_{ij'})} (P_i^1 U^V) \\ \beta_i^2 = \sum_{j=1}^{n2} \frac{\exp(\omega_{ij})}{\sum_{j'=1}^{n2} \exp(\omega_{ij'})} (P_i^2 U^V). \end{cases} \tag{8}$$

In equation (8), $n1$ and $n2$ respectively indicate the number of blocks within the two receptive domains. $n1$ is equal to $n2$ [30]. The self attention mechanism is clearly presented by equation (9).

$$\begin{aligned} A(Q, K, V) &= \mathrm{Soft\,max}(\frac{Q_1 K_1^T + Q_2 K_2^T}{\sqrt{d_{\mathrm{mod}\ el}}}) V \\ &= [\beta_1^1, \ldots, \beta_{n1}^1; \beta_1^2, \ldots, \beta_{n2}^2] \end{aligned} \tag{9}$$

In equation (9), $Q = \mathrm{concat}[Q1; Q2]$, $K = \mathrm{concat}[K1; K2]$, and $V = \mathrm{concat}[V1; V2]$. These are concatenated by query sets, key sets, and value sets of two

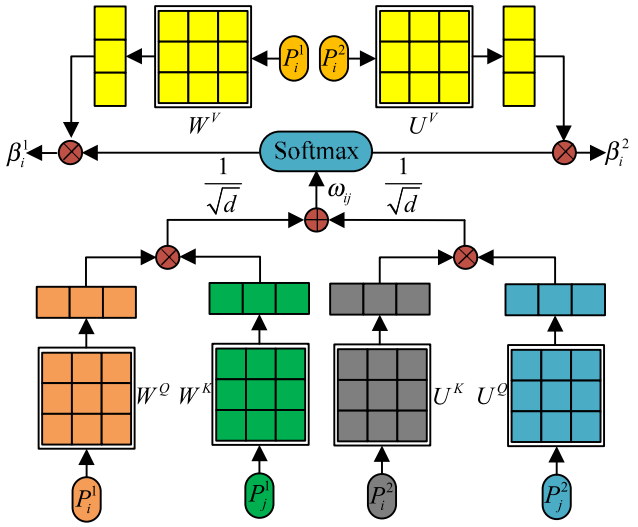

FIGURE 6. The structure of multiple receptive domain features fused from attention mechanism.

receptive fields. By performing multiplication operations between Q and K, the correlation between specific location features and other features can be evaluated. A high score means a strong correlation [31], [32]. Afterwards, the correlation score will be divided by the square root of the hidden layer dimension to maintain gradient stability. Then, the Softmax function is used to process the scores. The relative importance between each feature is determined. Finally, the output of the Softmax is multiplied by the V matrix to highlight important features while diluting the influence of irrelevant information. The calculation of each attention unit is shown in equation (10).

$$\text{head}_j = A(QW_j^Q, KW_j^K, VW_j^V) \tag{10}$$

In equation (10), W_j^Q, W_j^K, and W_j^V are linear transformation matrices for queries, keys, and values, respectively. The operational process of combining multiple self attention units to form multiple attention units is shown in equation (11).

$$\begin{aligned}&\text{MultiHead}(Q, K, V)\\ &\quad = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_n)W^o\end{aligned} \tag{11}$$

In equation (11), Concat is the matrix connection function. W^o is the additional weight matrix. By connecting n self attention mechanisms and performing matrix calculations with W^o, self attention from different sub-spaces can be compressed into a matrix to extract key features more accurately. After introducing the residual network into the multi-head attention mechanism, the output results are fed into the feed-forward network for calculation. The feed-forward network consists of two linear transformations, between which there is an activation function. The operation process of the feed-forward network is shown in equation (12).

$$\text{FeedForward}(x) = f(xW_1 + b_1)W_2 + b_2 \tag{12}$$

In equation (12), $f(\cdot)$ is the activation function. W_1 and W_2 represent the weight parameters, respectively. b is the bias term. The Trans-M model is developed to accurately monitor Internet traffic anomalies. The Trans-M applies a new block segmentation technique. The MRFF method of the decoder is introduced to address the structural information loss caused by segmentation [33], [34]. By combining the decoding part of MRFF and Transformer, this model can deeply analyze and model Internet traffic data. It is validated on standard datasets. The process includes obtaining and pre-processing data, dividing it into training and testing sets. Model training includes extracting information using extended convolutional modules with different expansion rates, and then processing blocks using block segmentation techniques and fully connected networks. Through MRFF and multi-head attention mechanism, the model can extract key features. After inputting training data into Trans-M and undergoing iterative training and parameter optimization, the model is saved when the loss decreases to an acceptable range or reaches the number of iterations. Finally, the model is evaluated on the testing set to calculate the accuracy of anomaly monitoring. The training process of the model is shown in Figure 7.

In order to comprehensively evaluate the performance of the classification model, multiple indicators such as accuracy, recall, and F1 score are usually used. Each indicator has its specific definition and calculation equation. The accuracy describes the proportion of correctly classified samples in the model, as shown in equation (13).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

In equation (13), TP is a true positive example. TN is a true negative example. FP is a false positive example. FN is a false negative example. Recall, also known as sensitivity, describes the proportion of the model correctly classified in all actual positive examples, calculated in equation (14).

$$\text{Re}\,call = \frac{TP}{TP + FN} \tag{14}$$

F1 score is the harmonic average of Precision and Recall, used to measure the overall performance of the classification performance, especially when the categories are imbalanced. The calculation is shown in equation (15).

$$F1 - score = 2 \times \frac{\text{Pr}\,ecision \times \text{Re}\,call}{\text{Pr}\,ecision + \text{Re}\,call} \tag{15}$$

Among them, Precision refers to the proportion of actual positive examples among the samples predicted by the model as positive examples, as shown in equation (16).

$$\text{Pr}\,ecision = \frac{TP}{TP + FP} \tag{16}$$

The above indicators comprehensively evaluate the performance of the model in different aspects, such as correctness, full coverage, and result balance, thus comprehensively characterizing the actual effectiveness of the model.

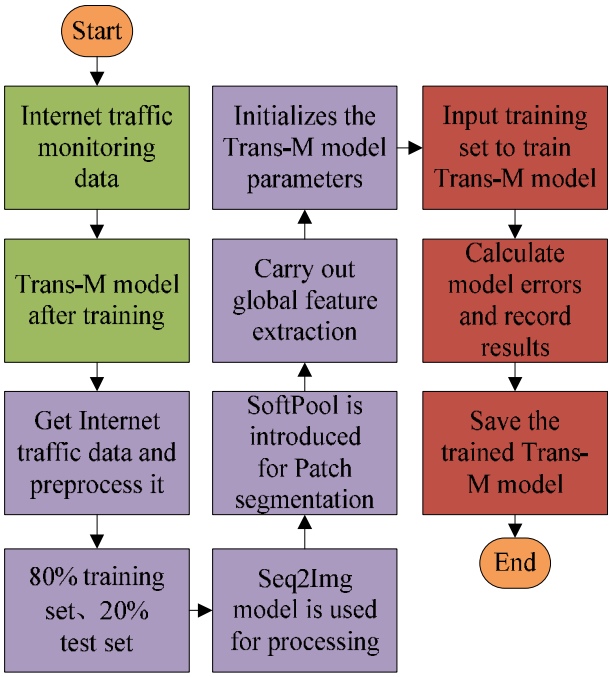

FIGURE 7. Training process of the model.

IV. EXPERIMENTAL RESULTS AND ANALYSIS OF INTERNET TRAFFIC MONITORING BASED ON TRANS-M

To further investigate the execution ability of the Trans-M model, this study explores it on different data sets. Firstly, the origin and pre-processing steps of the sample data set used are outlined. Subsequently, the evaluation criteria widely used in the Internet traffic anomaly recognition are deeply explored. The selected hyper-parameters are analyzed. Finally, a comparative experiment is conducted between the Trans-M and other different models.

A. HYPERPARAMETERS RESULTS AND ANALYSIS BASED ON TRANS-M

The CICIDS and NSLKDD data sets are selected as the experimental cores for this study. CICIDS contains approximately 2.84 million data points, involving normal Internet traffic and diverse abnormal traffic. Before the experiment, traffic data triggered by different attacks are classified and statistically analyzed. The 10 categories of Internet attacks and normal traffic are selected, totaling 11 categories, for in-depth experimental research. NSLKDD includes five categories, BENIGN, Dos, Probe, R2L, and U2R, with each data having 43 features. The training and testing data of NSLKDD are statistically analyzed and sampled by category, resulting in approximately 150000 pieces of data. In the binary classification experiment, BENIGN is used as normal Internet traffic, while the other four types are used as abnormal traffic. 4999 samples are collected for each type of abnormal traffic. In the multi-classification experiment, 4999 samples are taken for each category. All insufficient parts are sampled.

To evaluate the effectiveness of various improvement strategies in Trans-M, the study first conducts ablation experiments and uses True Positive Rate (TPR), False Positive Rate (FRP), Accuracy, Recall, specificity, and Frames Processed per Second (FPS) as evaluation indicators. The results of the ablation experiment are shown in Table 2. In Table 2, in the ablation experiment, each improvement strategy had a significant impact on performance. Firstly, the expansion convolutional unit enhances the model's ability to capture features, thereby improving TPR and Accuracy. Furthermore, the block segmentation algorithm helps to better understand the local structure of Internet traffic, thus reducing FPR. Most importantly, the MRFF self attention mechanism enables the model to better focus on important information, further improving TPR and accuracy. Compared with traditional Transformers, Trans-M improved TPR and accuracy by 5% and 4.5%, respectively, proving the effectiveness of the improved strategy. Meanwhile, the FPR also decreased from 0.12 to 0.07, further demonstrating the effectiveness of the improvement strategy. Therefore, the Trans-M model, which combines the expansion convolution unit, block segmentation algorithm and MRFF self attention mechanism, shows better performance in Internet traffic classification tasks, thanks to the optimization of model structure and attention mechanism.

To ensure the reasonable selection of hyper-parameters, the variable control method is used to select the hyper-parameters of the Trans-M model. The rationality of the selected parameters is confirmed by comparing the experimental results of different models. The three models involved include RNN, Bidirectional Long Short-Term Memory Network (Bi-LSTM), and Trans-M. The experimental results are shown in Figure 8. While maintaining the number of hidden layer neurons at 255 and the learning rate unchanged, the number of iterations is gradually adjusted to 999. From the experimental results, in the initial stage, the accuracy of RNN and Trans-M was similar, while Bi-LSTM had the highest accuracy. As the number of iterations gradually increased, the accuracy of each model improved. Although the Trans-M model was not always in the optimal state throughout the entire process, it ultimately achieved excellent performance. When the number of iterations reached 999, the accuracy of the Trans-M model reached the peak and showed good performance, thus determining the number of iterations as 999.

Next, the hyper-parameter value of the number of neurons in the hidden layer is adjusted. The fixed number of iterations is 999 and the learning rate is 0.001, while continuously adjusting the number of neurons in the hidden layer. Figure 9 displays the comparative experimental diagram of the scale and accuracy of the hidden layer. When adjusting the number of hidden layer neurons, the accuracy of Trans-M gradually improves. When the number of neurons in the hidden layer was 128, the accuracy of RNN and Bi-LSTM models was at the optimal state. However, when the number of hidden layer neurons reached 255, the Trans-M model exhibited the best performance and the accuracy surpassed the other two models. Therefore, the number of hidden layer neurons is 255.

TABLE 2. Results of Trans-M ablation experiment.

Ablation experiment	TPR	FPR	Accuracy	Recall	Specificity	FPS
Transformer	0.88	0.12	89%	88%	88%	220s
Transformer+expansion convolutional unit	0.90	0.10	90.5%	90%	90%	215s
Transformer+expansion convolutional units+block segmentation algorithm	0.92	0.08	92%	92%	92%	210s
Transformer+expansion convolutional units+block segmentation algorithm+MRFF self attention mechanism (Trans-M)	0.93	0.07	93.5%	93%	93%	205s

Finally, the learning rate is selected for hyper-parameter numerical experiments. During this training process, the number of iterations and the number of hidden layer neurons remain unchanged, at 999 and 255. Figure 10 shows the accuracy comparison of the Trans-M model at different learning rates. From Figure 10, during the experimental stage of learning rate, the accuracy of all three showed first increasing and then decreasing. When the learning rate was 0.001, the accuracy of the Trans-M model remained high. Therefore, the learning rate is 0.001. Based on the above experimental results, the model performs best when the number of iterations is 999, the number of hidden layer neurons is 255, and the learning rate is 0.001.

B. EXPERIMENTAL ANALYSIS OF DUAL CLASSIFICATION INTERNET TRAFFIC BASED ON TRANS-M

This study mainly focuses on the performance of Internet traffic classification tasks. The extension comparison is made between normal internet traffic and attack traffic. The parameter configuration used in the dual classification experiment includes a forgetting factor of 0.5, which means that each neural unit has a 38% to 40% probability of being omitted. Through experimental screening, the number of iterations, number of hidden layer neural units, and learning rate are determined to be 999, 225, and 0.001. Five decoder layers are stacked to form the encoder part of the model. Eight multi-head can synthesize the output of eight self attentions. To confirm the effectiveness of the Trans-M model proposed in this study, a comparative analysis is conducted between Trans-M and the other five models. They are DT, RNN, Bi-LSTM, the original Transformer model and Convolutional Neural Network-Residual Network (CNN-RN). In the dual classification experiment, the accuracy of DT, RN, Bi-LSTM, Transformer, CNN-RN, and Trans-M were 79%, 83%, 84.5%, 88%, 90%, and 92%, respectively. The accuracy gradually improved. Trans-M had the highest accuracy, with an improvement of approximately 2% compared with CNN ResNet. Compared with the DT, it was improved by approximately 18%. These data clearly indicate that Trans-M performs well in dual classification tasks.

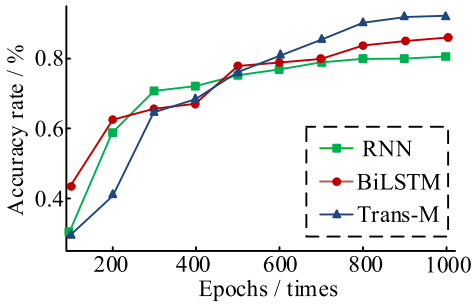

FIGURE 8. Results of Trans-M model iterations and accuracy.

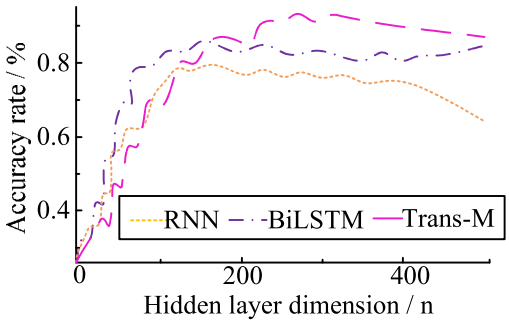

FIGURE 9. The number and accuracy of hidden layer neurons in Trans-M model.

To comprehensively evaluate the performance of Trans-M in dual classification tasks, accuracy, recall rate, and F1 score are compared. The comparison results of normal Internet traffic in the CICIDS dual classification diagram are shown in Figure 11.

In order to more comprehensively evaluate the performance of Trans-M in the dual classification task, the study compares the accuracy, recall rate and F1 score. Figure 11 shows the comparison results of the CICIDS data set under normal Internet traffic. In order to more comprehensively evaluate the performance of Trans-M in the dual classification task, the study compares the accuracy, recall rate and F1 score. Figure 11 shows the comparison results of the CICIDS data set under normal Internet traffic. In Figure 11 (a), The accuracy of the Trans-M model is significantly higher than other models. The accuracy of DT is the lowest, only 79%, while the accuracy of RN and Bi LST are 83% and 84.5%, respectively. In contrast, The accuracy of the Transformer model is 88%, The combination model of CNN ResNet achieved an accuracy of 90%. However, The accuracy of Trans-M reached 92%, ranking first among all models, with an improvement of about 2% compared to CNN ResNet and about 18% compared to decision tree models. This indicates that, The Trans-M model has strong learning ability and generalization performance in classification tasks. In Figure 11 (b), The F1 score of Trans-M is as high as 93%, which is nearly 5 percentage points higher than the F1 score of the initial Transformer scheme. This significant improvement validates the effectiveness of the implemented

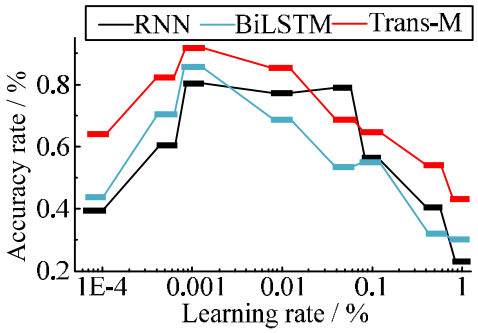

FIGURE 10. Trans-M model learning rate and accuracy results.

optimization strategy in enhancing the overall performance of the model. In other models, the F1 score of DT, RN, Bi LSTM, and Transformer is 79%, 83%, 84.5%, and 88%, respectively, while the F1 score of CNN ResNet is 90%. It can be seen from this that although there are differences among the models on F1 score, Trans-M undoubtedly performs the best. This not only emphasizes the advantages of Trans-M in balancing accuracy and recall, but also demonstrates its robust performance in complex traffic environments. In Figure 11 (c), the Area Under the Curve (AUC) of Trans-M reached 0.96, outperforming all other comparative models. In comparison, the AUC of Transformer and Bi-LSTM were 0.93 and 0.91, respectively, while DT had the lowest AUC of 0.85. This indicates that Trans-M performs better in balancing accuracy and recall, which verifies the excellent performance of Trans-M in Internet traffic classification tasks.

Figure 12 shows the comparison data of Internet attack traffic against CICIDS data set. From Figure 12, the improved Trans-M model also exhibited excellent performance in identifying internet attack traffic. The F1 score achieved an accuracy of 91%. Compared with the initial Transformer model, there was a 2.4% improvement. From Figures 11 and 12, the combination of convolutional neural networks and residual networks, as well as Trans-M and the original Transformer, perform relatively well in identifying normal and Internet attack traffic. The Bi-LSTM performs outstandingly in monitoring normal Internet traffic, but it is slightly inferior to other models in monitoring Internet attack traffic. The performance of DT and RNN is relatively weak. Especially in the task of identifying Internet attack traffic, the accuracy and recall are relatively low. Although Trans-M does not achieve the best performance in accuracy in Internet attack traffic, it has the best recall and F1 score for internet attack traffic. In summary, the Trans-M model proposed in the study exhibits better performance compared with other models.

Figure 13 shows the Receiver Operator Characteristic curve (ROC) for the dual classification task in the CICIDS data set. To further confirm the utility of the proposed Trans-M model and comprehensively evaluate the universality performance, the KSLKDD sample data is selected. A dual classification test experiment is conducted.

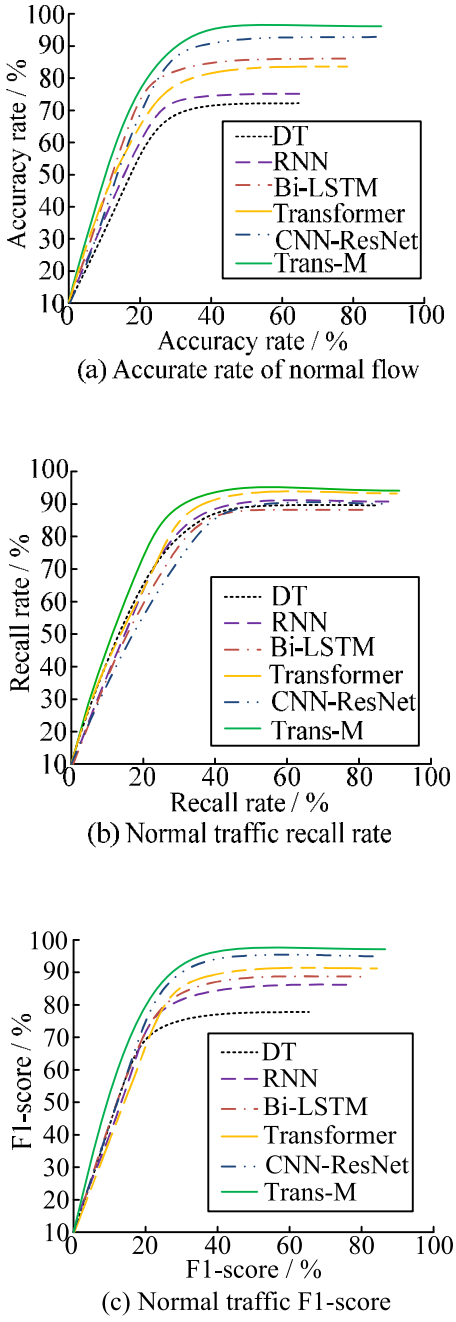

FIGURE 11. Comparison of indicators of normal traffic dual classification.

In Figure 14 (a), The Trans-M model performed the best in accuracy, reaching 93%, with an improvement of approximately 3.5% compared with the Transformer model's 89.5%. The accuracy of Bi-LSTM and CNN-RN were 88% and 85%, respectively, while DT had the lowest accuracy, only 79%. This result demonstrates the significant advantage of Trans-M in accuracy, especially in complex traffic classification tasks. The Trans-M model had a recall of 94%, while the Transformer, Bi-LSTM, and CNN-RN models were 92%, 90%, and 87%, respectively. DT still performed the worst at 82%. The F1 scores of Trans-M, Transformer, Bi-LSTM,

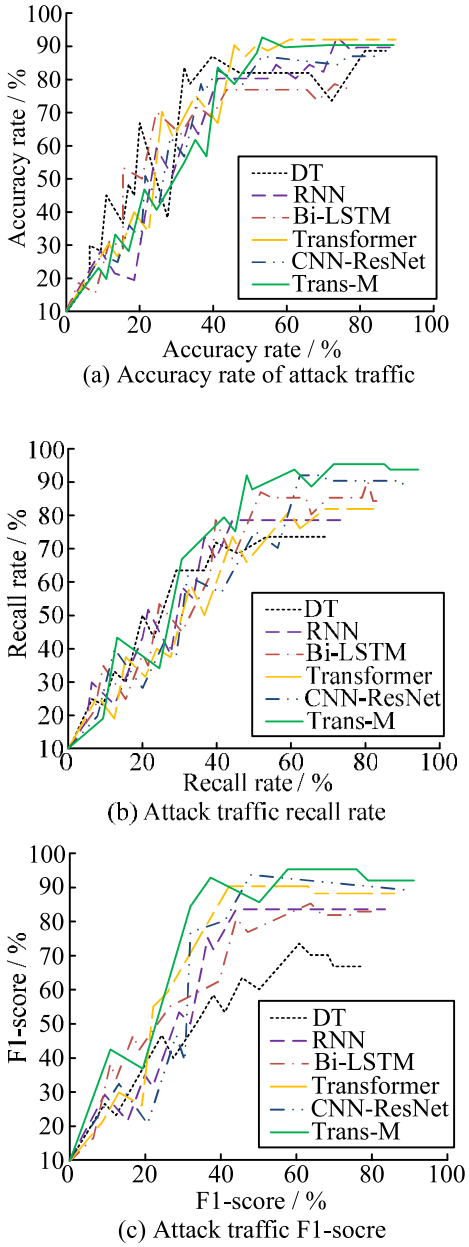

FIGURE 12. Comparison of indicators of both attack traffic categories.

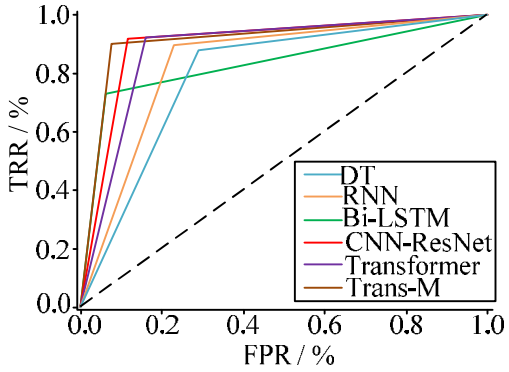

FIGURE 13. Comparison of working characteristic curves of subjects with double classification.

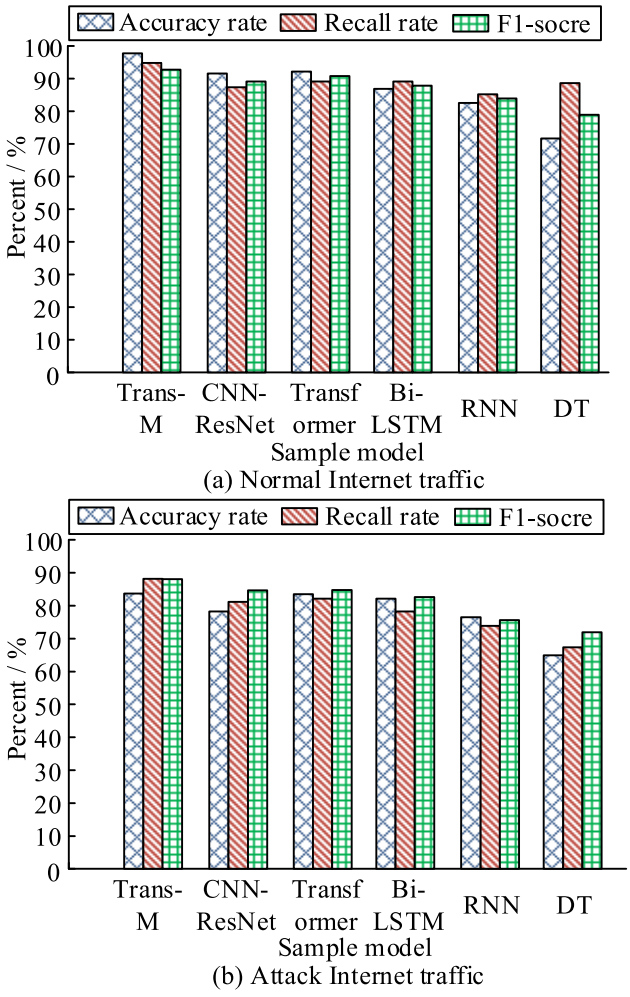

FIGURE 14. NSLKDD binary classification six model index comparison.

CNN-RN, and DT models were 93.5%, 91.1%, 89%, 88%, and 80%, respectively, demonstrating the superiority of the Trans-M model in balancing accuracy and recall. In Figure 14 (b), on the accuracy of Internet attack traffic, the Trans-M model was about 0.10% higher than the Transformer model. From the perspective of recall, compared with the Transformer model, Trans-M was improved about 0.20% in terms of normal Internet traffic. In terms of Internet attack traffic, it was about 4% higher than the Transformer model and 36.3% higher than the worst DT. In terms of F1 score, Compared with the Transformer model, the Trans-M model had a 1.9% increase in normal Internet traffic and a 2.3% increase in Internet attack traffic. Therefore, the Trans-M model is superior to other models in terms of accuracy, recall and F1 core, fully verifying its excellent performance in Internet traffic binary classification tasks.

Figure 15 presents the working characteristic curves of the subjects in the binary classification experiment conducted on the NSLKDD data set. A detailed comparison

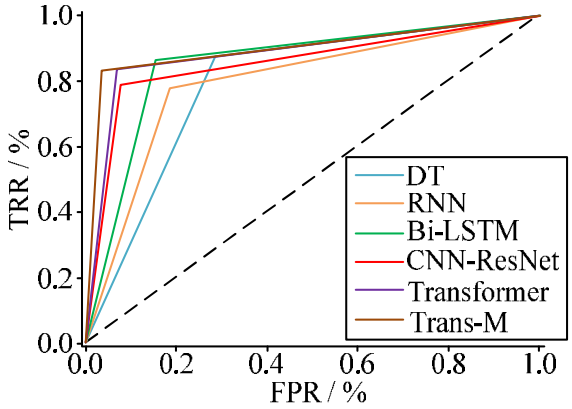

FIGURE 15. NSLKDD double classification subjects working characteristic curve comparison.

is made between normal Internet traffic and attack traffic. The Trans-M model exhibited the best performance among the dual classification responsibilities, with an ROC of 0.91. Next was the Transformer model, with an ROC of 0.89. The combination model of Bi-LSTM and CNN-RN also performs well, with an ROC of 0.85 and 0.87, respectively. The performance of the DT and the RNN was slightly inferior, with an ROC of 0.80 and 0.81. This study focuses on the binary classification task of Internet traffic anomaly detection. The performance effects using deep learning models for this task are explored. The superior performance of the Trans-M model in binary classification tasks is verified through the CICIDS and NSLKDD data sets. Compared with other comparative models, the Trans-M model showed significant advantages on both data sets, verifying the excellent performance in the binary classification task of Internet traffic anomaly detection. This conclusion indicates that the Trans-M model is more effective in handling complex Internet traffic data. Therefore, the study designs a powerful deep learning strategy for the binary classification task of Internet traffic anomaly detection, providing reliable support for Internet security assurance in practical applications.

Table 3 shows the accuracy, Recall and F1 score of the Trans-M model and five other models at different noise levels. From the table, the Trans-M model exhibited excellent performance at different noise levels, particularly in terms of recall and F1 score. Under noiseless conditions, the accuracy of Trans-M was 93%, significantly higher than Transformer's 89.5% and Bi-LSTM's 88%. Even when the noise level increased to 50%, the accuracy of Trans-M remained at 85%, while Transformer and Bi-LSTM decreased to 81% and 79%, respectively. In addition, the recall and F1 score of Trans-M were also higher than other models at various noise levels, further indicating its robustness and adaptability in processing noisy data. In contrast, DT performed worst under all noise conditions, highlighting its limitations in Internet traffic classification tasks. Therefore, the stability and superiority of the Trans-M model under various noise conditions have verified its potential application in complex data environments.

TABLE 3. Robustness results of Trans-M model at different noise levels.

Noise level	Indicator	Trans-M	Transforme [15]	Bi-LSTM [16]	CNN-RN [17]	RNN [18]	DT [19]
No noise	Accuracy	93.11%	89.59%	88.22%	85.12%	83.85%	79.11%
	Recall	94.12%	92.34%	90.23%	87.23%	85.67%	82.23%
	F1-score	93.51%	91.11%	89.23%	88.45%	84.45%	80.34%
10% noise	Accuracy	92.11%	88.45%	86.21%	83.46%	81.21%	77.54%
	Recall	93.22%	90.23%	88.65%	85.78%	83.23%	80.26%
	F1-score	92.51%	89.07%	87.65%	86.56%	82.54%	78.67%
20% noise	Accuracy	91.34%	87.32%	85.45%	82.32%	80.14%	75.26%
	Recall	92.23%	89.95%	87.67%	84.12%	82.23%	78.64%
	F1-score	91.56%	88.09%	86.89%	85.34%	81.54%	76.18%
30% noise	Accuracy	89.78%	85.34%	83.23%	80.67%	78.58%	73.19%
	Recall	90.56%	87.43%	85.45%	82.78%	80.54%	76.15%
	F1-score	89.52%	86.43%	84.13%	83.11%	79.26%	74.53%
40% noise	Accuracy	87.13%	83.23%	81.34%	78.23%	76.29%	71.14%
	Recall	88.23%	85.11%	83.45%	80.56%	78.28%	73.27%
	F1-score	87.55%	84.12%	82.48%	81.89%	77.27%	72.18%
50% noise	Accuracy	85.67%	81.09%	79.58%	75.23%	74.74%	69.47%
	Recall	86.67%	83.10%	81.92%	78.45%	76.65%	71.52%
	F1-score	85.51%	82.11%	80.11%	79.67%	75.78%	70.17%

V. CONCLUSION

In order to accurately identify abnormal behaviors of complex Internet traffic and avoid information loss in the block segmentation phase, a Trans-M model based on improved Transformer algorithm was proposed. This model adopted a block segmentation algorithm to ensure that a large amount of useful information could be retained during the segmentation stage, thereby improving the segmentation quality of the model. By interacting with multiple receptive fields of information, the Trans-M model effectively reduced information loss and improved model accuracy and efficiency. After multiple rounds of testing during the experimental stage, the optimal hyper-parameter combination with 999 iterations, 255 hidden layer neurons, and a learning rate of 0.001 was determined to achieve the best performance of the Trans-M model. On the CICIDS sample data, the Trans-M model achieved 93% F1 score in normal Internet traffic and 91% F1 score in Internet attack traffic, which was 5% and 2.4% higher than the original Transformer model, respectively. In addition, in the binary classification experiment of the NSLKDD data set, the AUC of Trans-M model was 0.90, which was superior to other models. The Bi-LSTM has an optimal recall rate of 89.8% for normal Internet traffic, while the DT had the best accuracy for Internet attack traffic. The Trans-M model performs well in all indicators, demonstrating excellent performance in the binary classification task of internet traffic anomaly monitoring.

Therefore, the main findings of the study are summarized as follows:

1. Accurately identify abnormal behaviors in complex Internet traffic, and effectively prevent information loss in the block segmentation phase.

2. Improve the Transformer algorithm by constructing a Trans-M model, retaining a large amount of useful information through block segmentation algorithm, and improving segmentation quality and model performance.

3. On the CICIDS data set, the normal Internet traffic F1 core of the Trans-M model reached 93%. Internet attack traffic F1 core was 91%, which was 5% and 2.4% higher than the original Transformer.

4. In the NSLKDD data set dual classification experiment, the AUC of the Trans-M model reached 0.90, which was superior to other models.

The limitation of this study is the lack of diversity of data sets. In the future, more different types of Internet traffic data should be considered to verify the generalization ability of the Trans-M model. The future research directions can be expanded from the following two aspects:

1. Explore more improvement strategies and techniques to enhance the Trans-M model performance.

2. Combine other advanced in-depth learning technologies and algorithms to improve the accuracy and real-time of the model and provide more powerful technical support for Internet traffic security monitoring.

REFERENCES

[1] N. Hubballi and P. Khandait, "*KeyClass*: Efficient keyword matching for network traffic classification," *Comput. Commun.*, vol. 185, pp. 79–91, Mar. 2022, doi: 10.1016/j.comcom.2021.12.021.

[2] W. Wei, H. Gu, W. Deng, Z. Xiao, and X. Ren, "ABL-TC: A lightweight design for network traffic classification empowered by deep learning," *Neurocomputing*, vol. 489, no. 7, pp. 333–344, Jun. 2022, doi: 10.1016/j.neucom.2022.03.007.

[3] H. Han, Z. Yan, X. Jing, and W. Pedrycz, "Applications of sketches in network traffic measurement: A survey," *Inf. Fusion*, vol. 82, pp. 58–85, Jun. 2022, doi: 10.1016/j.inffus.2021.12.007.

[4] C. Do, D. Duong, and D. Hoang, "A multi-layer approach for advanced persistent threat detection using machine learning based on network traffic," *J. Intell. Fuzzy Syst.*, vol. 40, no. 6, pp. 11311–11329, Jan. 2021, doi: 10.3233/JIFS-202465.

[5] P. A. Mitroshin, Y. Y. Shitova, Y. A. Shitov, A. A. Mitroshin, and D. N. Vlasov, "GIS-monitoring of regional transport network traffic as a method to study commuting: Moscow region case," *J. Phys., Conf. Ser.*, vol. 1828, no. 1, Feb. 2021, Art. no. 012073, doi: 10.1088/1742-6596/1828/1/012073.

[6] C. D. Xuan, H. Thanh, and N. T. Lam, "Optimization of network traffic anomaly detection using machine learning," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 3, pp. 2360–2370, Jun. 2021, doi: 10.11591/ijece.v11i3.pp2360-2370.

[7] C. D. Xuan, "Detecting APT attacks based on network traffic using machine learning," *J. Web Eng.*, vol. 20, no. 1, pp. 171–190, Jan. 2021, doi: 10.13052/jwe1540-9589.2019.

[8] 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. Sensor Netw.*, vol. 18, no. 3, pp. 182459–182476, Mar. 2022, doi: 10.1177/15501477211049910.

[9] Z. Long and W. Jinsong, "Network traffic classification based on a deep learning approach using NetFlow data," *Comput. J.*, vol. 66, no. 8, pp. 1882–1892, Aug. 2023, doi: 10.1093/comjnl/bxac049.

[10] V. Ponnusamy, A. Yichiet, N. Jhanjhi, M. Humayun, and M. F. Almufareh, "IoT wireless intrusion detection and network traffic analysis," *Comput. Syst. Sci. Eng.*, vol. 40, no. 3, pp. 865–879, Mar. 2022, doi: 10.32604/csse.2022.018801.

[11] L. Li, V. Okoth, and S. E. Jabari, "Backpressure control with estimated queue lengths for urban network traffic," *IET Intell. Transp. Syst.*, vol. 15, no. 2, pp. 320–330, Feb. 2021, doi: 10.1049/itr2.12027.

[12] T. Zanma, D. Hashimoto, K. Koiwa, and K. Liu, "Estimation of network traffic status and switching control of networked control systems with data dropout," *IET Cyber-Phys. Syst., Theory Appl.*, vol. 7, no. 2, pp. 69–80, Nov. 2021, doi: 10.1049/cps2.12024.

[13] C. Wang, H. Zhou, Z. Hao, S. Hu, J. Li, X. Zhang, B. Jiang, and X. Chen, "Network traffic analysis over clustering-based collective anomaly detection," *Comput. Netw.*, vol. 205, Mar. 2022, Art. no. 108760, doi: 10.1016/j.comnet.2022.108760.

[14] S. Dong and Y. Xia, "Network traffic identification in packet sampling environment," *Digit. Commun. Netw.*, vol. 9, no. 4, pp. 957–970, Aug. 2023, doi: 10.1016/j.dcan.2022.02.003.

[15] G. Sri Vidhya and R. Nagarajan, "Performance analysis of network traffic intrusion detection system using machine learning technique," *Int. J. Commun. Antenna Propag.*, vol. 12, no. 2, p. 111, Apr. 2022, doi: 10.15866/irecap.v12i2.21724.

[16] Y. Peng, W. Wu, J. Ren, and X. Yu, "Novel GCN model using dense connection and attention mechanism for text classification," *Neural Process. Lett.*, vol. 56, no. 2, pp. 1–17, Apr. 2024, doi: 10.1007/s11063-024-11599-9.

[17] M. Zulqarnain, R. Ghazali, M. Aamir, and Y. M. M. Hassim, "An efficient two-state GRU based on feature attention mechanism for sentiment analysis," *Multimedia Tools Appl.*, vol. 83, no. 1, pp. 3085–3110, Jan. 2024, doi: 10.1007/s11042-022-13339-4.

[18] M. Laurer, W. van Atteveldt, A. Casas, and K. Welbers, "Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI," *Political Anal.*, vol. 32, no. 1, pp. 84–100, Jan. 2024, doi: 10.1017/pan.2023.20.

[19] F. Ullah, S. Ullah, G. Srivastava, and J. C.-W. Lin, "IDS-INT: Intrusion detection system using transformer-based transfer learning for imbalanced network traffic," *Digit. Commun. Netw.*, vol. 10, no. 1, pp. 190–204, Feb. 2024, doi: 10.1016/j.dcan.2023.03.008.

[20] A. Fateh, R. T. Birgani, M. Fateh, and V. Abolghasemi, "Advancing multilingual handwritten numeral recognition with attention-driven transfer learning," *IEEE Access*, vol. 12, pp. 41381–41395, 2024, doi: 10.1109/ACCESS.2024.3378598.

[21] M. Hasanvand, "Machine learning methodology for identifying vehicles using image processing," *Artif. Intell. Appl.*, vol. 1, no. 3, pp. 170–178, Jul. 2023, doi: 10.47852/bonviewaia3202833.

[22] I. Hidayat, M. Z. Ali, and A. Arshad, "Machine learning-based intrusion detection system: An experimental comparison," *J. Comput. Cognit. Eng.*, vol. 2, no. 2, pp. 88–97, Jul. 2022, doi: 10.47852/bonviewjcce2202270.

[23] Y. Lei, "Research on microvideo character perception and recognition based on target detection technology," *J. Comput. Cognit. Eng.*, vol. 1, no. 2, pp. 83–87, Jan. 2022, doi: 10.47852/bonviewjcce19522514.

[24] S. Pal, A. Roy, P. Shivakumara, and U. Pal, "Adapting a Swin Transformer for license plate number and text detection in drone images," *Artif. Intell. Appl.*, vol. 1, no. 3, pp. 145–154, Apr. 2023, doi: 10.47852/bonviewaia3202549.

[25] B. Rajendran and S. Venkataraman, "Detection of malicious network traffic using enhanced neural network algorithm in big data," *Int. J. Adv. Intell. Paradigms*, vol. 19, nos. 3–4, pp. 370–379, Jan. 2021, doi: 10.1504/IJAIP.2021.116366.

[26] A. M. Vulfin, V. I. Vasilyev, V. E. Gvozdev, K. V. Mironov, and O. E. Churkin, "Network traffic analysis based on machine learning methods," *J. Phys., Conf. Ser.*, vol. 2001, no. 1, Aug. 2021, Art. no. 012017, doi: 10.1088/1742-6596/2001/1/012017.

[27] O. M. A. Alssaheli, Z. Z. Abidin, N. A. Zakaria, and Z. A. Abas, "Implementation of network traffic monitoring using software defined networking ryu controller," *WSEAS Trans. Syst. CONTROL*, vol. 16, pp. 270–277, May 2021, doi: 10.37394/23203.2021.16.23.

[28] Ł. Wawrowski, M. Michalak, A. Białas, R. Kurianowicz, M. Sikora, M. Uchroński, and A. Kajzer, "Detecting anomalies and attacks in network traffic monitoring with classification methods and XAI-based explainability," *Proc. Comput. Sci.*, vol. 192, pp. 2259–2268, Aug. 2021, doi: 10.1016/j.procs.2021.08.239.

[29] M. P. Karpowicz, "Adaptive tuning of network traffic policing mechanisms for DDoS attack mitigation systems," *Eur. J. Control*, vol. 61, pp. 101–118, Sep. 2021, doi: 10.1016/j.ejcon.2021.07.001.

[30] Z. Ming, L. Zhang, Y. Xu, and M. Bakshi, "An algorithm for matrix recovery of high-loss-rate network traffic data," *Appl. Math. Model.*, vol. 96, pp. 645–656, Aug. 2021, doi: 10.1016/j.apm.2021.03.036.
[31] A. R. Abbasi and D. Baleanu, "Recent developments of energy management strategies in microgrids: An updated and comprehensive review and classification," *Energy Convers. Manage.*, vol. 297, Dec. 2023, Art. no. 117723, doi: 10.1016/j.enconman.2023.117723.
[32] A. R. Abbasi and M. Mohammadi, "Probabilistic load flow in distribution networks: An updated and comprehensive review with a new classification proposal," *Electr. Power Syst. Res.*, vol. 222, Sep. 2023, Art. no. 109497, doi: 10.1016/j.epsr.2023.109497.
[33] X. Zheng, B. Li, Q. Wang, D. Wang, and Y. Li, "Emerging low-nuclearity supported metal catalysts with atomic level precision for efficient heterogeneous catalysis," *Nano Res.*, vol. 15, no. 9, pp. 7806–7839, Sep. 2022, doi: 10.1007/s12274-022-4429-9.
[34] A. Abbasi and A. Seifi, "Fast and perfect damping circuit for ferroresonance phenomena in coupling capacitor voltage transformers," *Electr. Power Compon. Syst.*, vol. 37, no. 4, pp. 393–402, Mar. 2009, doi: 10.1080/15325000802548780.

JIEJING LIU was born in Hebei, in 1981. He received the bachelor's and master's degrees from Hebei University, in 2003 and 2012, respectively. He was a Teacher with the Department of Electronic Information Engineering, Hengshui University, from 2003 to 2015; the Deputy Director of the Office of Academic Affairs, Hengshui University, from 2015 to 2021; and the Dean of the Department of Mathematics and Computer Science, Hengshui University, from 2021 to 2024. He has published a total of three monographs and more than 30 papers.

XU LIU was born in Hebei, in 1980. He received the bachelor's and master's degrees from Hebei University, in 2003 and 2013, respectively. He was a Clerk and the Deputy Director of the Office of Academic Affairs, Hebei University of Engineering, from 2003 to 2019 and from 2020 to 2024, respectively. He has published a total of eight papers.

YANHAI WANG was born in Hebei, in 1976. He received the bachelor's degree from North China University of Water Resources and Electric Power, in 1999. From 2003 to 2024, he was a Clerk of the Education Technology Center, Hebei University of Engineering. He has published a total of six papers.

HUA FU was born in Hebei, in 1981. She received the bachelor's degree from Nanjing University of Posts and Telecommunications, in 2004, and the master's degree from Hebei University of Engineering, in 2019. From 2004 to 2024, she was a Staff Member of China Telecom Corporation Ltd., Handan Branch. She has published a total of three papers.