electronics

MDPI

Article

MFF: A Multimodal Feature Fusion Approach for Encrypted Traffic Classification

Hong Huang [1,2], Yinghang Zhou [1,*], Feng Jiang [3], Xiaolin Zhou [1] and Qingping Jiang [1]

[1] School of Computer Science and Engineering, Sichuan University of Science and Engineering, Yibin 644000, China; huanghong@suse.edu.cn (H.H.); 324085406139@stu.suse.edu.cn (X.Z.); 324085406111@stu.suse.edu.cn (Q.J.)
[2] Key Laboratory of Enterprise Informatization and IoT Measurement and Control Technology for Universities in Sichuan Province, Zigong 643000, China
[3] School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China; fjiang@hit.edu.cn
* Correspondence: 323085406233@stu.suse.edu.cn

Abstract

With the widespread adoption of encryption technologies, encrypted traffic classification has become essential for maintaining network security awareness and optimizing service quality. However, existing deep learning-based methods often rely on fixed-length truncation during preprocessing, which can lead to the loss of critical information and degraded classification performance. To address this issue, we propose a Multi-Feature Fusion (MFF) model that learns robust representations of encrypted traffic through a dual-path feature extraction architecture. The temporal modeling branch incorporates a Squeeze-and-Excitation (SE) attention mechanism into ResNet18 to dynamically emphasize salient temporal patterns. Meanwhile, the global statistical feature branch uses an autoencoder for the nonlinear dimensionality reduction and semantic reconstruction of 52-dimensional statistical features, effectively preserving high-level semantic information of traffic interactions. MFF integrates both feature types to achieve feature enhancement and construct a more robust representation, thereby improving classification accuracy and generalization. In addition, SHAP-based interpretability analysis further validates the model's decision-making process and reliability. Experimental results show that MFF achieves classification accuracies of 99.61% and 99.99% on the ISCX VPN-nonVPN and USTC-TFC datasets, respectively, outperforming mainstream baselines.

Keywords: encrypted traffic classification; feature fusion; ResNet; SE attention; SHAP

Academic Editor: Juan-Carlos Cano

Received: 13 May 2025
Revised: 24 June 2025
Accepted: 25 June 2025
Published: 26 June 2025

Citation: Huang, H.; Zhou, Y.; Jiang, F.; Zhou, X.; Jiang, Q. MFF: A Multimodal Feature Fusion Approach for Encrypted Traffic Classification. *Electronics* **2025**, *14*, 2584. https://doi.org/10.3390/electronics14132584

1. Introduction

With the widespread adoption of encryption technologies and the growing demand for data privacy protection, encrypted traffic on the Internet has experienced explosive growth. According to Google's latest Transparency Report released in January 2025 [1], all of the world's top 100 websites now fully support the HTTPS protocol. On Windows and Mac platforms, the proportion of webpages loaded via HTTPS has reached 94% and 98%, respectively. While encryption significantly enhances the security of data transmission and the protection of user privacy, it also introduces substantial challenges to traffic classification tasks due to its obfuscation of content. This growing opacity complicates network management, content regulation, and security defense mechanisms. As a result, encrypted

traffic classification has emerged as a critical and urgent research challenge in the field of cybersecurity [2,3].

Currently, encrypted traffic classification methods can be broadly categorized into three main types [4] (see Figure 1): (a) rule-based approaches relying on plaintext features; (b) machine learning methods based on statistical features; and (c) deep learning approaches utilizing raw traffic features. Rule-based methods typically depend on unencrypted information such as port numbers, Deep Packet Inspection (DPI), and certificate data exposed during TLS/SSL handshakes to classify traffic [5,6]. However, as the degree of encryption continues to increase, the effectiveness of such methods has significantly declined.

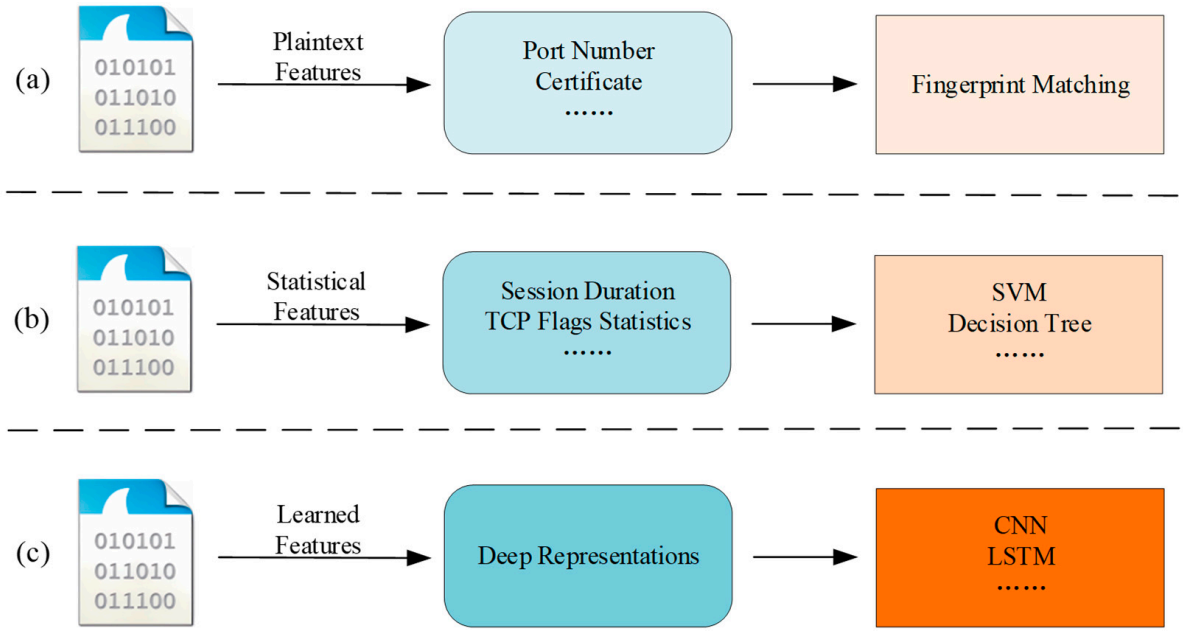


Figure 1. Encrypted traffic classification methods.

To address this limitation, researchers have explored machine learning techniques based on statistical features, which enable classification without decryption by extracting traffic-level metrics such as packet length and inter-arrival time [7]. For example, Draper-Gil et al. [8] proposed a classification approach based on time-related statistical features that effectively distinguishes VPN from non-VPN traffic. Nevertheless, these methods often suffer from limited generalization across different network environments and datasets.

In recent years, deep learning methods have gained popularity in encrypted traffic classification due to their end-to-end feature learning capabilities [9–11]. For instance, Wang et al. [9] introduced a CNN-based traffic classification model that can automatically extract latent features from raw data, overcoming the limitations of handcrafted feature engineering. However, these methods usually require input data to be truncated to a fixed length, which may lead to the loss of critical information and subsequently impair classification performance.

To address the problem of feature loss caused by data truncation, this paper proposes an MFF model for encrypted traffic classification. MFF aims to enhance the robustness of traffic representation by jointly modeling temporal and statistical features, thereby improving classification accuracy and generalization capability. The MFF model comprises two parallel feature extraction branches: on one hand, a ResNet18-based architecture is augmented with a Squeeze-and-Excitation (SE) attention mechanism to enhance the model's ability to adaptively capture key temporal features; on the other hand, an autoencoder

is employed to construct high-dimensional representations from 52 statistical features, serving as global descriptors of traffic flows. These two types of features are subsequently fused to form a comprehensive representation, which is passed through a multi-layer fully connected network for adaptive integration and final classification.

Extensive experiments conducted on two public datasets, ISCX VPN-nonVPN [8] and USTC-TFC [12], across seven evaluation groups demonstrate that MFF significantly outperforms existing methods in both accuracy and robustness. In addition, the interpretability of the model is further validated through SHAP-based analysis, which confirms the effectiveness of the proposed feature fusion strategy.

The main contributions of this paper are as follows:

- A multimodal encrypted traffic classification model named MFF is proposed, which integrates both temporal and statistical features. This model effectively addresses the feature loss problem caused by traffic truncation, thereby enhancing classification accuracy and generalization performance.
- A dual-path feature extraction and fusion mechanism is designed, where the ResNet18 architecture is enhanced with a Squeeze-and-Excitation (SE) attention mechanism for temporal feature extraction, and a deep autoencoder is employed to perform nonlinear dimensionality reduction on statistical features. By fusing these two types of features, the model achieves feature enhancement and improved overall performance.
- Extensive experiments on the ISCX VPN-nonVPN and USTC-TFC datasets across seven evaluation groups demonstrate the superiority of the proposed method. Furthermore, SHAP-based interpretability analysis is conducted to enhance the transparency and trustworthiness of the model.

The remainder of this paper is organized as follows. Section 2 reviews classical approaches and recent advances in the field of encrypted traffic classification. Section 3 describes the datasets used and provides a detailed explanation of the proposed MFF model. Section 4 presents the experimental results and offers an in-depth analysis of the model's performance and interpretability. Finally, Section 5 concludes the paper and outlines potential directions for future research.

2. Related Work

2.1. Rule-Based Methods Leveraging Plaintext Features

Traffic classification methods based on plaintext features identify encrypted traffic by analyzing information exposed during the TLS/SSL handshake process, such as certificate data, traffic fingerprints, and keyword patterns, from which classification rules are derived. Common approaches include port-based identification and Deep Packet Inspection (DPI) techniques [4]. However, with the increasing prevalence of port obfuscation and advanced encryption technologies, traditional port-based identification methods are becoming less effective in complex network environments. Nevertheless, due to their simplicity and low computational overhead, these methods remain widely used in practice as auxiliary tools for traffic recognition.

In the realm of DPI-based research, recent efforts have proposed more sophisticated rule-matching strategies to enhance classification performance. For instance, the FlowPrint method [13] generates application fingerprints by analyzing temporal dependencies and destination host features within encrypted flows, enabling automated recognition of previously unseen applications through clustering and correlation graph construction. However, its performance is susceptible to noise from third-party shared traffic (e.g., advertisement libraries), which can significantly degrade identification accuracy. Similarly, Hayes et al. [14] proposed the k-Fingerprinting method, which constructs traffic fingerprints based on features such as packet count, sequence, and transmission timing, demonstrating

strong performance in open network environments. Yet, in anonymous communication scenarios such as VPNs, these features can be easily manipulated, leading to a marked decline in classification accuracy [15].

In contrast, the proposed MFF model in this work eliminates reliance on plaintext features, offering improved generalizability and robustness across diverse network conditions.

2.2. Machine Learning Methods Based on Statistical Features

To overcome the limitations of rule-based approaches, researchers have increasingly adopted machine learning techniques [16], leveraging statistical features of encrypted traffic for classification. Depending on the granularity of feature extraction, these methods can be broadly categorized into packet-level and session-level approaches. For example, AppScanner [17] employs packet-level statistical features for classification, using a feature contribution thresholding method to select 40 effective features out of 54, and then trains a random forest classifier to achieve high classification performance. Cao et al. [18] proposed the SPP-SVM model, which integrates feature scaling, PCA-based dimensionality reduction, and an enhanced particle swarm optimization algorithm to automatically tune the parameters of the SVM classifier, thereby improving the accuracy and efficiency of encrypted traffic classification. Additionally, Dusi et al. [19] introduced a method for classifying SSH encrypted traffic based on flow-level statistical features, combining Gaussian Mixture Models (GMM) with Support Vector Machines (SVM), which demonstrated strong performance in SSH traffic classification tasks.

Despite their promising results, machine learning methods based on statistical features suffer from a critical limitation: their performance heavily depends on expert knowledge and manually engineered feature sets [20]. The process of feature design is not only time-consuming and labor-intensive but also constrained by the scope of expert understanding, which undermines the generalizability of these models in dynamic and complex network environments. In contrast, the proposed MFF model in this study does not rely solely on handcrafted statistical features. Instead, it integrates both statistical and deep learning–based self-learned features, significantly enhancing the model's robustness and stability.

2.3. Deep Learning Methods Based on Self-Learned Features

In recent years, deep learning has garnered considerable attention in encrypted traffic classification due to its ability to automatically learn high-dimensional data representations. Wang et al. [9] were among the first to propose transforming raw traffic byte sequences into grayscale images for input into a Convolutional Neural Network (CNN), thereby circumventing the limitations of traditional handcrafted feature engineering and enabling direct feature extraction from raw traffic data. Building on this idea, Mohammad et al. [21] introduced the Deep Packet framework, which combines CNNs with stacked autoencoders (SAEs) to construct an end-to-end encrypted traffic classification model, achieving improved recall performance. Lin et al. [22] proposed the TSCRNN model, which integrates CNNs with Recurrent Neural Networks (RNNs) to effectively capture temporal dependencies between packets, thereby enhancing classification accuracy. Zhang et al. [15] developed the ICLSTM model, employing an Inception module alongside Long Short-Term Memory (LSTM) networks to jointly model local spatial and temporal features within packet sequences. More recently, Lin et al. [23] introduced ET-BERT, a transformer-based model that incorporates a dual-byte encoding strategy to further improve the fidelity of traffic representation. Additionally, Zhang et al. [24] proposed a graph-based deep traffic classification method that integrates adaptive data augmentation with Graph Neural Networks (GNNs), significantly improving model generalization under complex network conditions.

While deep learning–based methods have markedly advanced the performance of encrypted traffic classification [25,26], most existing approaches are constrained by the need to truncate traffic data to fit fixed input dimensions. This truncation can result in the loss of critical information from the original traffic, thereby adversely affecting classification accuracy. A summary of mainstream methods is provided in Table 1. To address this issue, the proposed MFF model incorporates statistical features and jointly models them with temporal features automatically extracted by deep learning, effectively achieving feature enhancement. This fusion mitigates the potential loss of critical information caused by data truncation, thereby improving the robustness and generalization ability of the classification model.

Table 1. Summary of mainstream methods.

Reference	Method	Advantages	Limitations
[13]	FlowPrint	Identifies applications without prior knowledge of features	Susceptible to interference from third-party shared traffic
[14]	k-Fingerprinting	Fast training and inference	Vulnerable to feature tampering
[17]	AppScanner	High degree of automation	Model stability affected by version updates
[18]	SVM	Optimizes classification parameters, reducing computational complexity	Sensitive to data scaling and feature dimensionality
[21]	CNN	Integrates feature extraction and classification	Relatively low accuracy
[21]	CNN + SAE	Automatic feature extraction	High dependency on specific datasets
[15]	Inception-LSTM	Effectively handles class imbalance	Complex parameter tuning, prone to overfitting
[22]	CNN + RNN	Efficient processing of encrypted traffic volumes	Complexity in handling long flows
[23]	ET-BERT	Strong representation capability	Requires significant computational resources for training/inference
[24]	TFE-GNN	High accuracy	High model complexity and computational cost

3. Methodology

In recent years, multimodal feature fusion has become a prominent research focus in the task of encrypted traffic classification [4,20,26,27]. For example, Wei et al. [28] proposed the FE-MTDM method, which performs feature grouping and generation using the K-means algorithm and then feeds both original and generated features into an ensemble classifier consisting of a shallow neural network and a random forest. Miao et al. [29] introduced the FERNN-AC method, which enhances classification performance by incorporating a feature enhancement module and an angle constraint mechanism, combined with a Recurrent Neural Network (RNN) for temporal feature extraction.

However, these approaches typically rely on direct concatenation or shallow feature combinations and lack effective modeling of interactions between different feature modalities or structured fusion design. In contrast, our proposed MFF model introduces several key improvements in feature fusion strategy: (1) for temporal modeling, we enhance the ResNet18 architecture with a Squeeze-and-Excitation (SE) attention mechanism to improve the model's focus on critical dynamic features at the channel level; (2) for statistical feature processing, we use an autoencoder to perform nonlinear dimensionality reduction and semantic reconstruction of raw statistical information, providing greater abstraction and compact representation compared to simple concatenation; (3) at the fusion stage, we design a unified mapping mechanism for dual-path outputs, enabling effective integration of

heterogeneous features through a multi-layer fully connected structure; and (4) in terms of interpretability, we apply SHAP analysis to uncover the model's decision rationale, thereby improving its transparency and reliability.

The remainder of this section introduces the proposed MFF model for encrypted traffic representation learning. Figure 2 illustrates the overall architecture. Section 3.1 describes the datasets used in the experiments. Section 3.2 details the data preprocessing steps. Section 3.3 provides an in-depth explanation of the MFF model architecture and the design of each component module.

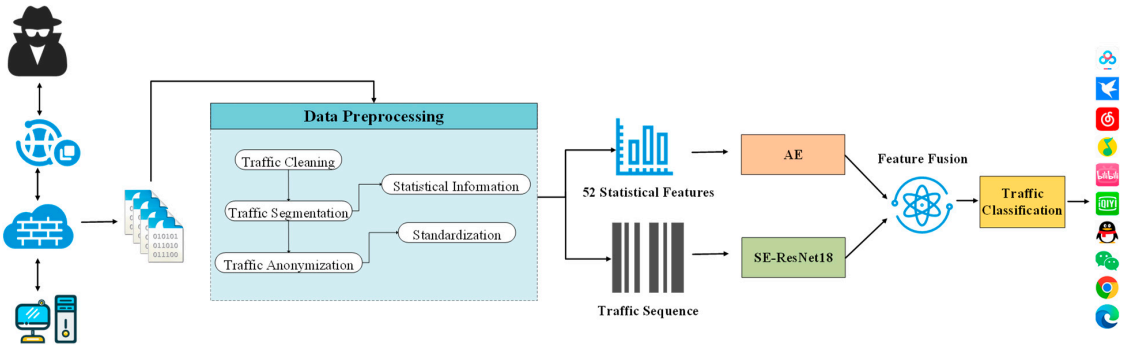


Figure 2. Overall architecture of the proposed MFF model.

3.1. Datasets

This study employs two benchmark datasets for encrypted traffic classification: ISCX VPN-nonVPN [8] and USTC-TFC [12], as detailed below.

The ISCX VPN-nonVPN dataset is a widely used standard benchmark in the field of encrypted traffic analysis. It comprises 14 traffic categories, including seven types of regular encrypted traffic and seven types of encrypted traffic transmitted through VPN tunnels, with a total data volume of 26.2 GB. For experimental purposes, we select six clearly labeled classes from each category (i.e., regular encrypted and VPN-encrypted traffic). The specific traffic types are listed in Table 2.

Table 2. Description of the ISCX VPN-nonVPN dataset.

Type	Traffic Name
Regular encrypted traffic	Chat, Email, File Transfer, P2P, Streaming, VoIP
VPN encrypted traffic	VPN-Chat, VPN-Email, VPN-File Transfer, VPN-P2P, VPN-Streaming, VPN-VoIP

The USTC-TFC dataset consists of 20 categories in total—10 malicious and 10 benign traffic classes—with an overall size of 3.71 GB. Each traffic type is annotated with detailed labels, facilitating effective model training and performance evaluation. The full list of traffic categories is provided in Table 3.

Table 3. Description of the USTC-TFC dataset.

Type	Traffic Name
Benign	BitTorrent, Facetime, FTP Gmail, MySQL, Outlook, Skype, SMB, Weibo, WorldOfWarcraft,
Malware	Cridex, Geodo, Htbot, Miuref, Neris, Nsis-ay, Shifu, Tinba, Virut, Zeus

3.2. Data Preprocessing

Data preprocessing is a critical step in encrypted traffic classification, as it directly affects the effectiveness of subsequent feature extraction and the overall performance of the classification model. In the proposed MFF model, the preprocessing pipeline consists of four main stages: traffic segmentation, statistical feature extraction, traffic anonymization, and flow length normalization. The specific procedures and implementation details for each stage are described as follows.

3.2.1. Session-Based Traffic Segmentation

In real-world network environments, the collected traffic is typically a sequential mixture of packets from multiple applications [30]. For example, traffic captured at a gateway may contain data transmitted by different hosts across different applications. To isolate the traffic belonging to individual applications or sessions, it is necessary to segment the raw traffic stream. In this context, a flow is defined as a sequence of packets sharing the same 5-tuple information: (source IP, source port, destination IP, destination port, transport protocol).

The raw traffic stream, denoted as P_{raw}, consists of a sequence of packets with varying sizes. Each packet is denoted as p_i and the complete sequence is represented as follows:

$$\begin{gathered} P_{raw} = \{p_1,\ p_2,\ p_3,\ \cdots,\ p_n\} \\ p_i = (x_i, l_i, t_i) \\ i = 1, 2, \cdots, n,\ l_i \in \mathbb{N},\ t_i \in [0, \infty) \end{gathered} \tag{1}$$

Here, n denotes the total number of packets, x_i represents the 5-tuple information, l_i is the length of the i-th packet in bytes, and t_i denotes the timestamp at which the packet was captured.

Existing traffic segmentation methods are typically divided into two categories: flow-based segmentation and session-based segmentation. A flow, denoted as P_{flow}, refers to a collection of packets that share the same 5-tuple information:

$$P_{flow} = \{p_1 = (x, l_1, t_1), \cdots, p_n = (x, l_n, t_n),\ \text{where } t_1 < t_2 < \cdots < t_n \tag{2}$$

In contrast, a session is defined as a group of flows in which the source and destination addresses in the 5-tuple are interchangeable. Compared to flow-based segmentation, session-based segmentation provides notable advantages for encrypted traffic classification. First, it avoids traffic fragmentation caused by splitting data into short flows, which may disrupt the semantic continuity of traffic sequences and degrade model performance. Second, it preserves essential temporal patterns—such as session duration and inter-packet intervals—that are critical for capturing the behavioral characteristics of encrypted traffic.

Therefore, this study adopts session-based segmentation to maintain the completeness of encrypted traffic sequences, thereby enhancing classification accuracy and improving model robustness.

3.2.2. Statistical Feature Extraction

After the raw PCAP data are segmented by session, a traffic feature extraction tool [31] is employed to extract 52 statistical features for each session. The design of these features draws upon several representative studies in the field of encrypted traffic classification [3,4,8,20,25] and leverages existing tool support for standardized extraction. These features span multiple dimensions, including protocol flags, temporal characteristics, length distribution, information entropy, burst patterns, and TLS-specific fields, aiming to

comprehensively characterize the behavioral patterns of network sessions from a global perspective. Such feature dimensions have been widely applied in tasks such as VPN detection, malicious communication identification, and behavioral modeling, demonstrating strong generality and discriminative power. Detailed descriptions of the features are provided in Table 4.

Table 4. Description of the 52 extracted statistical features.

No.	Feature Description	No.	Feature Description
1	Avg. TCP SYN flag count per session	27	Burst duration
2	Avg. TCP URG flag count	28	Avg. burst interval
3	Avg. TCP FIN flag count	29	Byte transmission rate (B/s)
4	Avg. TCP ACK flag count	30	Burst packet count
5	Avg. TCP PSH flag count	31	Uplink/downlink byte count
6	Avg. TCP RST flag count	32	Uplink/downlink packet count
7	Proportion of DNS packets in session	33	Packet inter-arrival entropy
8	Proportion of TCP packets in session	34	Packet length entropy
9	Proportion of UDP packets in session	35	Packet inter-arrival peak
10	Proportion of ICMP packets in session	36	Packet interval entropy
11	Session duration (s)	37	TLS JA3 fingerprint entropy
12	Mean time gap between adjacent packets	38	Packet length peak
13	Min. time gap between adjacent packets	39	Packet length variance
14	Max. time gap between adjacent packets	40	Median packet interval
15	Std. deviation of inter-packet intervals	41	Median packet length
16	Avg. packet length	42	25th percentile of packet length
17	Min. packet length	43	75th percentile of packet length
18	Max. packet length	44	Proportion of small packets (<32 B)
19	Std. deviation of packet length	45	Packet rate (pkt/s)
20	Proportion of small packets (<32 B) in session	46	TCP duplicate packet ratio
21	Avg. TCP payload size	47	TLS record count
22	Max. TCP payload size	48	Avg. TLS record length
23	Min. TCP payload size	49	Avg. TCP window size
24	Std. deviation of TCP payload size	50	Std. deviation of TCP window size
25	DNS-to-TCP packet ratio	51	Empty packet count
26	Total number of packets in session	52	Number of out-of-order TCP packets

To eliminate scale disparities among different features, we apply Z-score normalization to each statistical feature x_i, performing feature standardization prior to model input. The normalized value is denoted as x_i' and is computed as shown in Equation (3):

$$x_i' = \frac{x_i - \mu_i}{\sigma_i + \varepsilon} \tag{3}$$

Here, μ_i and σ_i represent the mean and standard deviation of the i-th statistical feature, respectively, and ε is a small constant (set to 0.001 in this study) added to avoid division-by-zero errors.

3.2.3. Traffic Anonymization

To prevent the model from overly relying on specific sensitive information, traffic data are anonymized. Specifically, the IP addresses and MAC addresses in the session data are replaced with a placeholder value of 0×00 to eliminate the strong correlation between traffic classes and specific address information. In the original dataset, different traffic categories often carry unique IP addresses. Without anonymization, the model might use these distinctive features for classification, leading to overfitting [32].

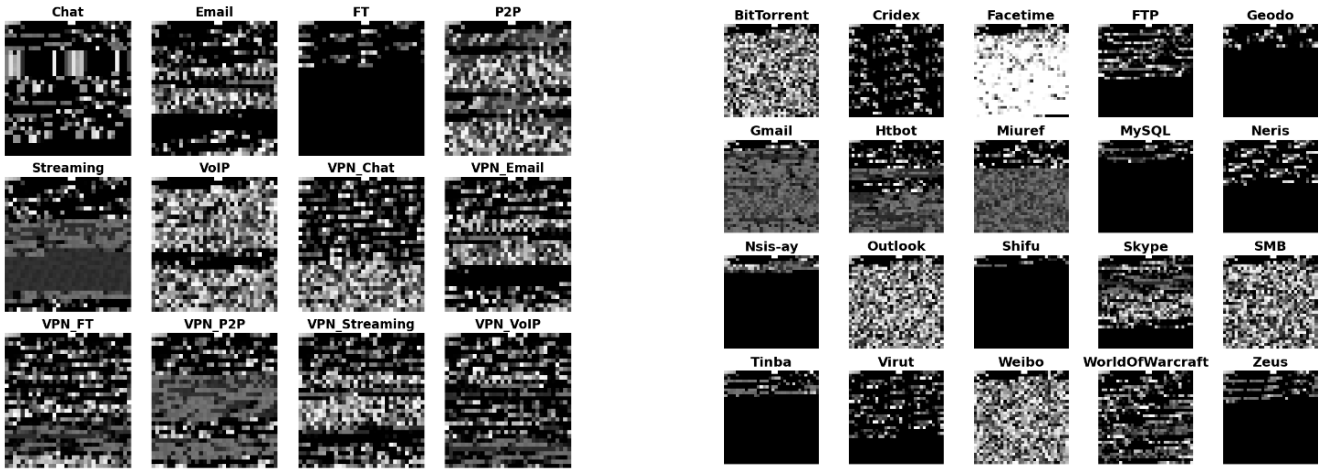
3.2.4. Standardized Traffic Length

To meet the input dimension requirements of the SE-ResNet18 model, a standardized processing workflow for traffic data was designed. Considering the sensitivity of deep learning models to input size, a truncation-padding strategy is applied to standardize the original traffic to a fixed length of 1024 bytes. Specifically, for traffic samples longer than 1024 bytes, only the first 1024 bytes are retained to capture the essential handshake information; for samples shorter than 1024 bytes, the remaining bytes are padded with 0×00. This standardization process not only ensures consistency in model input but also preserves the temporal features of the traffic data to the greatest extent.

3.2.5. Visualization-Based Analysis of Preprocessed Data

After completing the aforementioned preprocessing steps, we perform a visualization-based analysis of the traffic data to further explore underlying patterns and assess the effectiveness of the procedure. Specifically, each 1024-byte traffic sequence is reshaped into a 32×32 grayscale image matrix, in which each byte is directly mapped to a pixel intensity value ranging from 0 to 255.

Figure 3 illustrates the visualization results for both datasets: (a) 12 traffic categories from the ISCX VPN-nonVPN dataset and (b) 20 traffic categories from the USTC-TFC dataset. The visualizations reveal clear spatial pattern differences among traffic categories, which provide strong evidence for the validity of the proposed preprocessing approach. Upon closer inspection, each traffic class exhibits distinct texture patterns, resulting in visually discernible inter-class separability. For instance, in the USTC-TFC dataset, Cridex traffic presents a regular block-like texture, whereas Outlook traffic shows a dispersed dot-like pattern. These pronounced visual differences reflect strong inter-class separability and form a meaningful foundation for temporal feature extraction by the SE-ResNet18 model.



(**a**) Visualization of ISCX VPN-nonVPN Traffic (**b**) Visualization of USTC-TFC Traffic

Figure 3. Visualization of encrypted traffic samples.

3.3. Architecture Design

To mitigate the information loss caused by fixed-length truncation, the MFF model jointly models both temporal and statistical features of network traffic. The temporal feature branch employs an SE-ResNet18 architecture to model the normalized raw byte sequences, effectively capturing local dynamic patterns in traffic transmission, such as burst behavior and packet interaction rhythms. In parallel, the statistical feature branch uses an autoencoder to perform dimensionality reduction and reconstruction on session-level features, preserving global distributional information such as packet length distribution, inter-arrival

time, and TCP flag statistics. Through this "local + global" dual-path design, MFF effectively alleviates the contextual information loss caused by data truncation, achieves feature enhancement, and ultimately improves the model's ability to perceive and discriminate complex traffic behaviors. Prior studies have also shown that statistical features provide valuable supplementary information for long-term context modeling [3,20,25], while temporal features are more suitable for capturing transient behavioral variations [2,15,22,30].

The proposed MFF model aims to improve feature expressiveness and classification robustness in encrypted traffic classification tasks. The core components of the model consist of four major parts: data preprocessing, dual-branch feature extraction modules, feature fusion, and classification. The model takes as input a 1024-byte standardized traffic byte sequence and a 52-dimensional vector of statistical features, which are fed into the temporal and statistical branches, respectively. The output features from both branches are then integrated in the fusion and classification module, where a multilayer perceptron is used to perform feature fusion and produce the final prediction.

This section provides a detailed explanation of each module, including temporal feature extraction based on SE-ResNet18 (Section 3.3.1), statistical feature extraction based on autoencoder (AE) (Section 3.3.2), and feature fusion and classification (Section 3.3.3).

3.3.1. Temporal Feature Extraction Branch Based on SE-ResNet18

For temporal feature modeling, ResNet18-1D is adopted as the backbone network to extract features from encrypted traffic sequences. With its 18-layer architecture, ResNet18 provides a good balance between network depth and feature extraction capacity. Given that encrypted traffic is inherently sequential, prior studies [3,7,33] have shown that 1D convolution is better suited than 2D convolution for capturing fine-grained temporal dependencies. Moreover, the residual connections introduced in ResNet [34] help alleviate gradient degradation in deep networks, thereby improving model stability and enhancing feature propagation. The modular design of the ResNet family also facilitates extension to deeper architectures such as ResNet-34 and ResNet-50, offering strong scalability.

To enhance the model's sensitivity to informative features across channels, a Squeeze-and-Excitation (SE) attention mechanism [35,36] is integrated after each residual block. The SE module adaptively learns the importance of each channel, enabling dynamic feature recalibration and improving the model's ability to focus on key patterns within traffic sequences. Structurally, the SE block consists of two stages: squeeze and excitation, as illustrated in Figure 4. In the squeeze stage, global average pooling is applied to the 1D feature map $X \in \mathbb{R}^{L \times C}$, resulting in a single scalar per channel, defined as the following:

$$z_c = \frac{1}{L} \sum_{i=1}^{L} X_{i,c} \tag{4}$$

In the excitation stage, the aggregated descriptors are passed through two fully connected layers with nonlinear activations to generate channel-wise weights:

$$s = \sigma(W_2 \cdot \delta(W_1 \cdot z)) \tag{5}$$

where W_1 and W_2 are learnable parameters, $\delta(\cdot)$ is the ReLU activation function, and $\sigma(\cdot)$ denotes the sigmoid function. Finally, the original feature map is rescaled by these learned weights through channel-wise multiplication, enhancing the model's ability to emphasize critical temporal features.

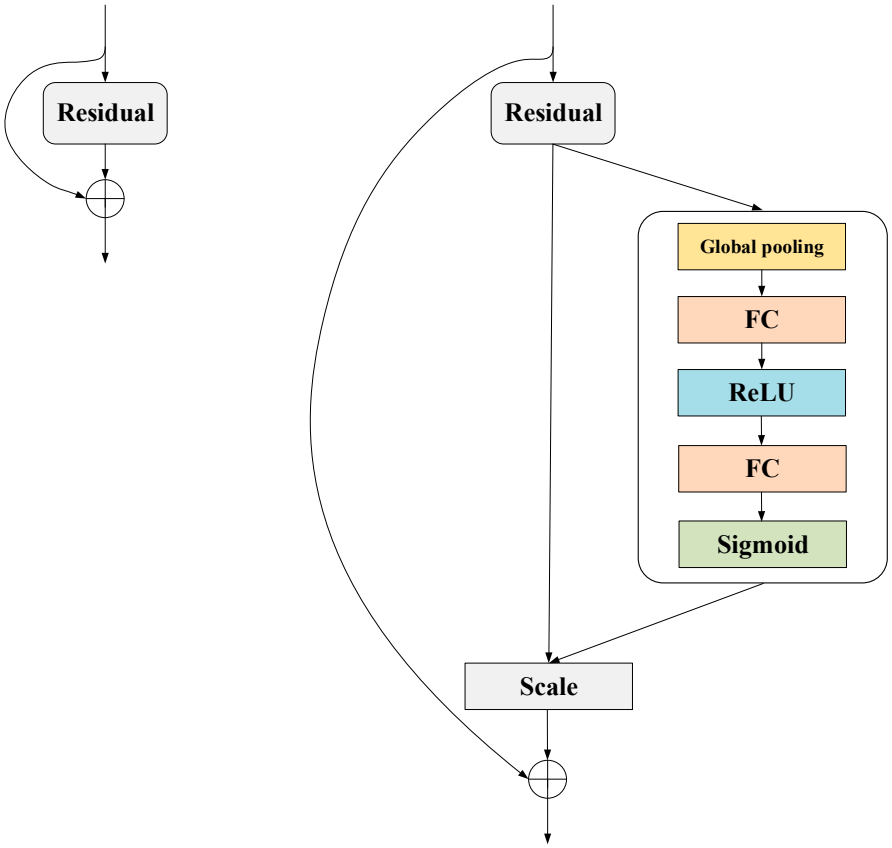


Figure 4. Structure of the Squeeze-and-Excitation (SE) module.

3.3.2. Statistical Feature Extraction Branch Based on AE

To address the potential information loss caused by feature sparsity, MFF incorporates a statistical feature enhancement branch. This branch processes the 52-dimensional statistical feature vector extracted from each session—covering traffic volume, timing, packet count, and other key metrics—to ensure the semantic completeness of non-sequential features.

To further improve the expressive power of statistical features, we design a reconstruction-based autoencoder (AE) that performs nonlinear compression and reconstruction of high-dimensional input. The encoder-decoder structure allows the model to learn latent representations that capture the intrinsic relationships among statistical features, thereby generating more compact and informative feature embeddings. The AE output is used as the enhanced representation of the statistical input. The detailed network architecture of the AE is illustrated in Figure 5.

Both the encoder and decoder are implemented as fully connected neural networks. The encoder is denoted as a function $g(\cdot)$, which maps the input x_i into a latent representation h_i, as shown in Equation (6):

$$h_i = g(x_i) \tag{6}$$

The decoder is modeled as another function $f(\cdot)$, which reconstructs the input from the latent space:

$$\widetilde{x}_i = f(h_i) = f(g(x_i)) \tag{7}$$

The training objective of the autoencoder is to jointly optimize f and g by minimizing the reconstruction error between the input and its reconstruction. We use Mean Absolute Error (MAE) as the loss metric, defined in Equation (8):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - f(g(x_i))| \tag{8}$$

To prevent overfitting and improve generalization, we also incorporate L_2 regularization on the model parameters. The complete objective function is given in Equation (9):

$$\arg\min_{f,g}\left(\frac{1}{N}\sum_{i=1}^{N}|x_i - f(g(x_i))| + \lambda\sum_{j}\theta_j^2\right) \tag{9}$$

Through this encoding-decoding process, the model generates 56-dimensional abstract representations of the statistical input, effectively capturing the hidden complexity within traffic behavior.

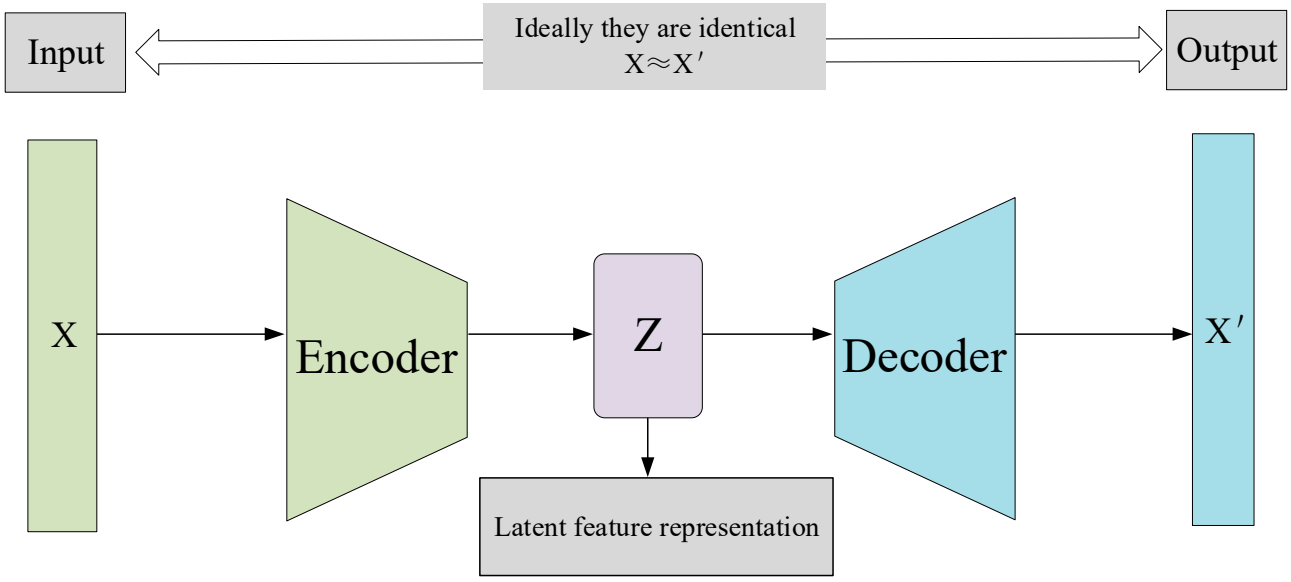


Figure 5. Architecture of the autoencoder (AE) module.

3.3.3. Feature Fusion and Classification

In the feature fusion and classification module, MFF integrates the outputs from the two feature extraction branches to form a unified representation of the traffic flow. Specifically, the temporal features F_{seq}, extracted via SE-ResNet18, capture local temporal dependencies and sequential patterns in encrypted traffic. The statistical features F_{stats}, derived from the autoencoder branch, encode the overall distributional characteristics of the flow. These two feature vectors are concatenated and passed to the fusion module to obtain the combined feature representation F_{concat}, as defined in Equation (10):

$$F_{\text{concat}} = f(F_{\text{seq}} \oplus F_{stats}) \tag{10}$$

After feature fusion, a two-layer fully connected neural network is designed to further explore the potential correlations between encoded features from different modalities. The first fully connected (FC) layer is followed by a ReLU activation function to introduce nonlinearity, thereby enhancing the model's representational capacity and yielding more expressive and discriminative feature representations. Subsequently, the second FC layer maps the fused features to the class space, which is then followed by a Softmax classifier to

produce the final traffic category label for classification. This process is formally defined in Equation (11). The structure of the core modules and the complete data processing workflow of MFF are illustrated in Figure 6, and the parameter configurations of the feature fusion and classification module are listed in Table 5.

$$y = \text{Softmax}(\text{W}^{(2)} \cdot \text{ReLU}(\text{W}^{(1)} \cdot F_{\text{concat}} + b^{(1)}) + b^{(2)}) \tag{11}$$

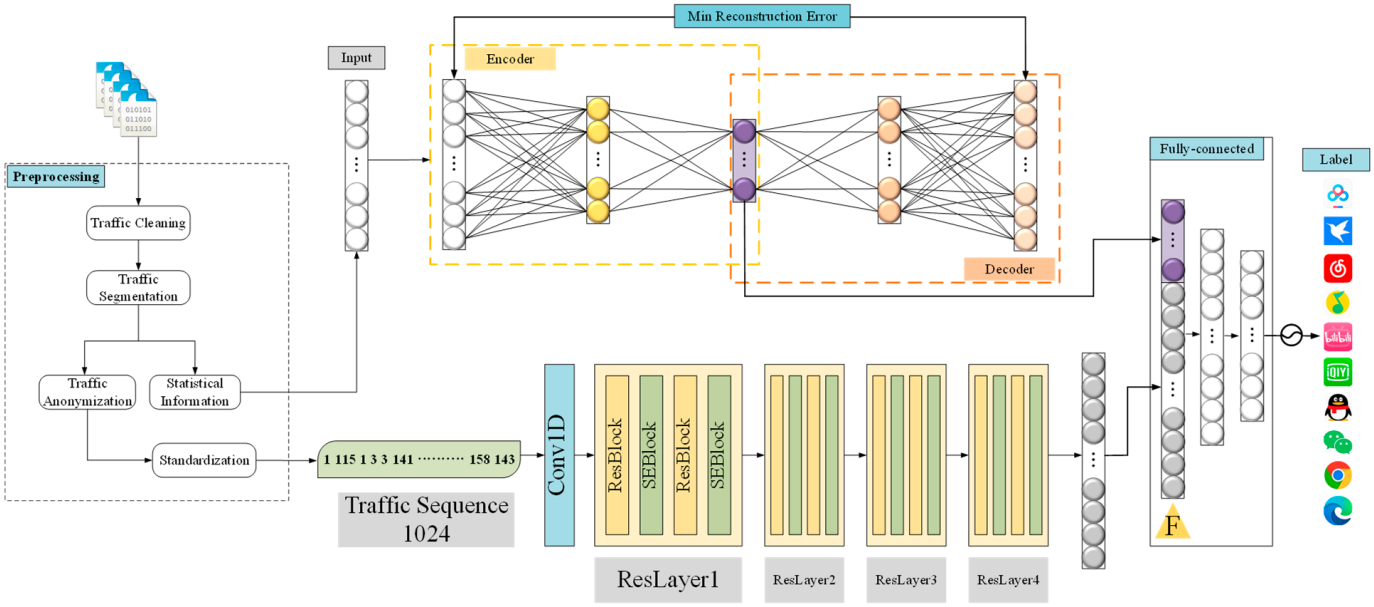


Figure 6. Detailed structure of the core modules in the MFF model.

Table 5. Detailed architecture and key parameters of the MFF model.

Module	Layer	Operation	Input	Filter	Output
SE-ResNet18	Conv-1	Conv1d	1×1024	1×9	32×1024
	ResLayer-1	ResBlock + SE	32×1024	1×3	32×1024
	ResLayer-2	ResBlock + SE	32×1024	1×3	64×512
	ResLayer-3	ResBlock + SE	64×512	1×3	128×256
	ResLayer-4	ResBlock + SE	128×256	1×3	256×128
	Avg Pooling	Avg Pooling	256×128	-	256×1
	Flatten	Flatten	256×1	-	256
AE	AutoEncoder-1	fully connected + ReLU	52	-	40
	AutoEncoder-2	fully connected	40	-	26
	AutoEncoder-3	fully connected + ReLU	26	-	40
	AutoEncoder-4	fully connected	40	-	52
Classification	Fully connected-1	fully connected + ReLU	256 + 26	-	100
	Fully connected-2	fully connected + Softmax	100	-	num classes

The ReLU activation function is defined as follows:

$$\text{ReLU}(x) = \begin{cases} x, \text{ if } x > 0 \\ 0, \text{ otherwise} \end{cases} \tag{12}$$

Through this fusion and classification pipeline, the model effectively maps the integrated features to the target traffic class and achieves accurate encrypted traffic recognition.

4. Experiment and Analysis

4.1. Experimental Setup and Configuration

All experiments in this study were conducted on a 64-bit Windows 11 Professional operating system. The hardware configuration includes an AMD Ryzen 7 7735H 3.20 GHz processor, 32 GB of RAM, and an NVIDIA GeForce RTX 4060 Laptop GPU with 8 GB of video memory. The experimental environment is based on Python 3.7.16, with the deep learning framework PyTorch 1.11.0 and CUDA version 11.3. Detailed specifications of the experimental environment are provided in Table 6.

Table 6. Experimental setup.

Category	Parameter
System	Windows 11 Professional
CPU	AMD Ryzen 7 7735H 3.20 GHz
Memory	32 GB
Graphics Card	NVIDIA GeForce RTX 4060 Laptop (8 GB)
Python Version	3.7.16
Deep Learning Backend	PyTorch 1.11.0
Cuda Version	11.3

During model training, reasonable hyperparameter configurations were determined through network searches to improve the convergence and performance of the model. The settings are as follows: the learning rate was set to 0.001, the Adam optimizer was used, and a weight decay coefficient of 0.1 was introduced to mitigate overfitting. The batch size was set to 128 to balance training efficiency and stability. The ResNet18 architecture used in the model has a residual block configuration of [2], corresponding to the stacking layers of four residual groups. Additionally, the number of training epochs was set to 120 to ensure sufficient convergence on all datasets.

The choice of 1024 bytes as the standardized input length in this study is grounded in a statistical analysis of the datasets. As shown in Figure 7, approximately 87% of session traffic in the ISCX VPN-nonVPN dataset falls within the range of 0 to 1024 bytes. Similarly, about 70% of sessions in the USTC-TFC dataset lie within this interval. These findings suggest that setting a fixed input length of 1024 bytes effectively captures the critical feature regions of most traffic samples, while mitigating the risk of information loss.

4.2. Experimental Settings

To comprehensively validate the proposed MFF model, experiments were conducted on two representative public datasets: ISCX VPN-nonVPN and USTC-TFC. These datasets cover various types of encrypted traffic, including VPN, non-VPN, malicious, and benign flows, providing good diversity and classification challenges. Each dataset was divided into training, validation, and test sets in a ratio of 8:1:1. Detailed descriptions of the datasets and preprocessing steps are provided in Section 3.

To simulate realistic deployment scenarios, seven classification tasks were designed, covering multiple levels including protocol encapsulation, service categories, and malware detection. The experimental configurations are summarized in Table 7, which systematically evaluates the classification performance of MFF across different settings.

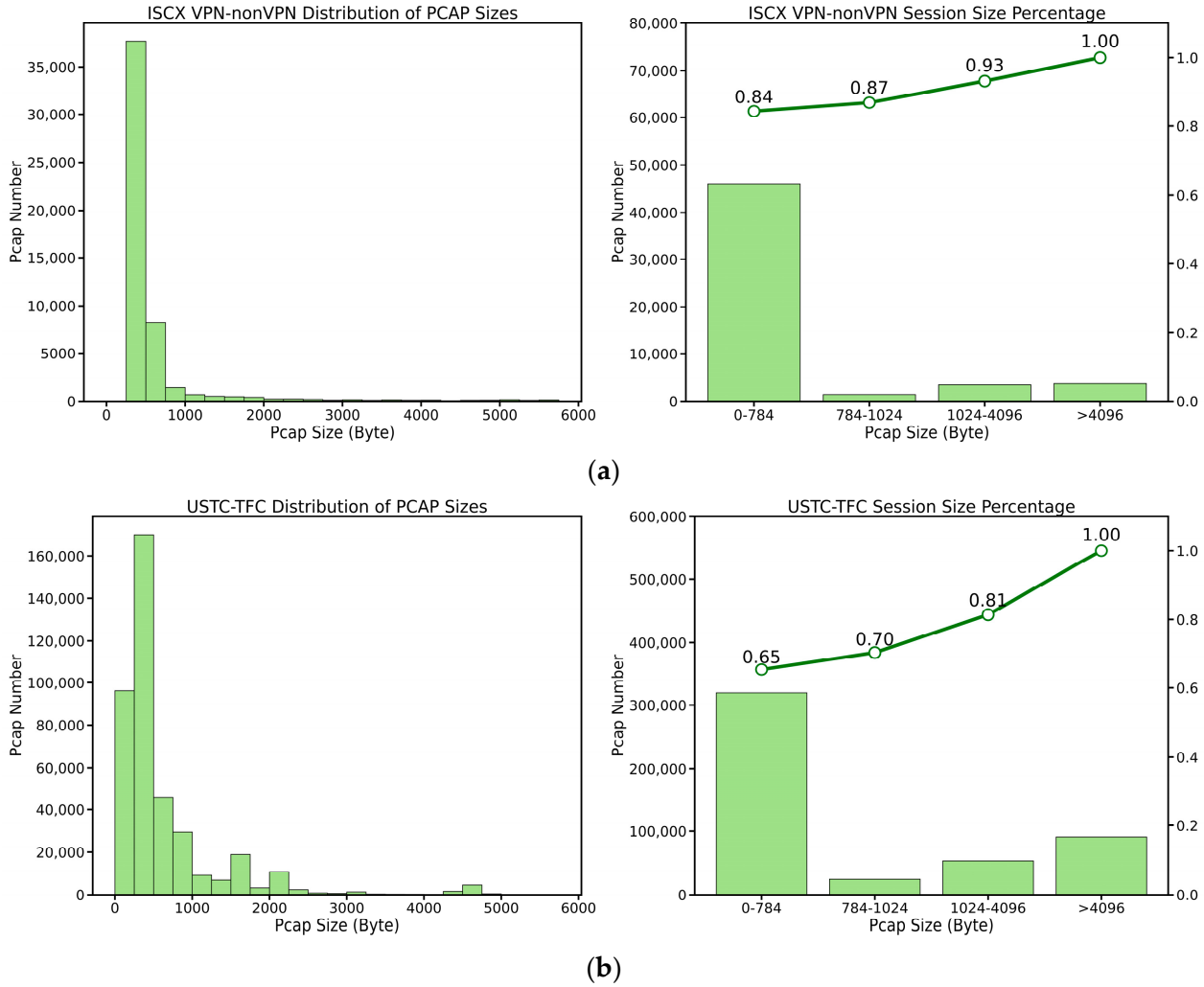


Figure 7. (**a**) Statistical distribution of session sizes in the ISCX VPN-nonVPN dataset. (**b**) Statistical distribution of session sizes in the USTC-TFC dataset.

Table 7. The experimental content.

Experiment	Dataset	Description	Classes
1	ISCX VPN-nonVPN	Classification based on encapsulation type	2
2		Non-VPN encrypted service classification	6
3		VPN encrypted service classification	6
4		Combined encrypted service classification	12
5	USTC-TFC	Classification of benign and malicious traffic	2
6		Fine-grained benign traffic classification	10
7		Malware family classification	20

4.3. Evaluation Metrics

To comprehensively evaluate the performance of the model, four widely used evaluation metrics are adopted in this study: accuracy (Equation (13)), precision (Equation (14)), recall (Equation (15)), and F1-score (Equation (16)). These metrics are widely recognized in both academic research and real-world applications and are instrumental in assessing the effectiveness and classification capability of the model from multiple perspectives. The specific formulas for each metric are defined as follows:

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

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{16}$$

4.4. Experimental Results and Analysis

4.4.1. Ablation Study

To thoroughly evaluate the individual contributions of each functional module in the MFF model to encrypted traffic classification, three ablation models are designed. These models aim to investigate the relationship between temporal features and global statistical features, as well as their respective impact on overall model performance. The experiments are conducted on the ISCX VPN-nonVPN dataset, with four distinct experimental scenarios to ensure diversity and comprehensiveness in the evaluation. To ensure fairness and comparability, all experiments use the same network architecture and hyperparameter settings except for the ablated components. The ablation designs are as follows:

(1) w/o SE and ResNet: Statistical feature path only

In this configuration, the entire temporal branch is removed, including the SE attention module and the ResNet18 network. Only the statistical feature branch (i.e., the autoencoder module) is retained, and the final classification is performed through fully connected layers. This setup forms a single-path structure using only statistical features, designed to evaluate the standalone modeling capability of statistical features in the absence of temporal information.

(2) w/o SE and AE: Basic temporal feature path only

This configuration removes both the statistical feature branch (AE module) and the SE module, retaining only the base ResNet18 network as the temporal feature extractor. It represents a single-path structure that uses only temporal features, aiming to assess the classification performance of basic temporal information without support from global statistical features, while also isolating the effect of attention-based enhancement from the SE module.

(3) w/o AE: Full temporal modeling without statistical features

In this setting, the SE-ResNet18 network is preserved as the temporal feature extraction branch, while the statistical feature path (AE module) is removed. This configuration is used to evaluate the combined effect of the SE attention mechanism and the residual structure on temporal feature modeling and to further explore the complementary value of statistical information in feature fusion.

(4) Full MFF structure: Complete multimodal design with temporal and statistical features

This configuration retains both the statistical feature branch (based on the autoencoder) and the temporal feature branch (SE-ResNet18) and employs a unified mapping structure to deeply fuse the two types of features. The model forms a complete multi-path architecture, aiming to jointly model local dynamic behaviors (e.g., burst traffic, packet interaction rhythms) and global statistical patterns (e.g., length distribution, protocol features), thereby achieving feature enhancement and decision optimization. By comparison with the three simplified or partial structures above, the effectiveness and necessity of the multimodal fusion strategy can be comprehensively evaluated.

The detailed results of the ablation study are presented in Table 8 and Figure 8, which show the classification accuracy under the four experimental configurations. Figure 9

further illustrates the confusion matrix of the MFF model on the ISCX VPN-nonVPN dataset, reflecting its ability to distinguish between different traffic categories. Figure 10 provides a detailed breakdown of precision, recall, and F1-score for each individual class.

Table 8. Results of the ablation study.

Model	AE	SE	ResNet	Acc (Exp 1)	Acc (Exp 2)	Acc (Exp 3)	Acc (Exp 4)
w/o SE and ResNet	✓	×	×	0.9956	0.7525	0.6639	0.7376
w/o SE and AE	×	×	✓	0.9989	0.9831	0.9877	0.9765
w/o AE	×	✓	✓	0.9991	0.9862	0.9918	0.9883
MFF	✓	✓	✓	0.9993	0.9946	0.9959	0.9961

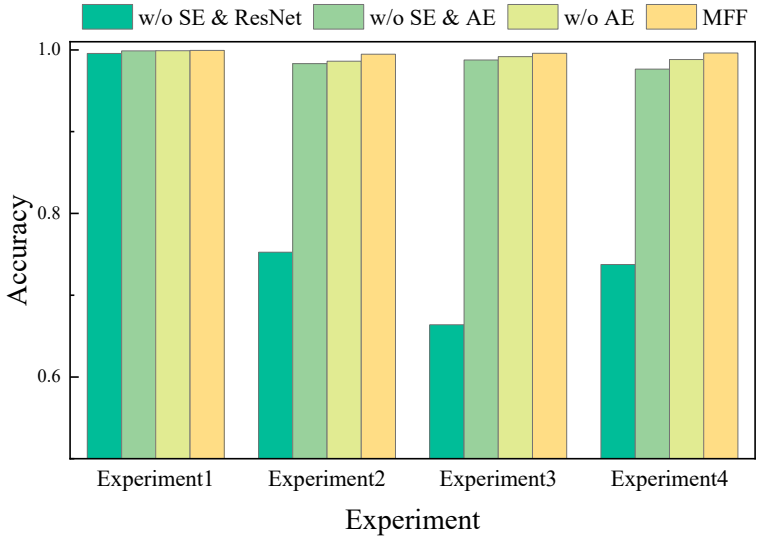


Figure 8. Accuracy comparison of ablation experiments.

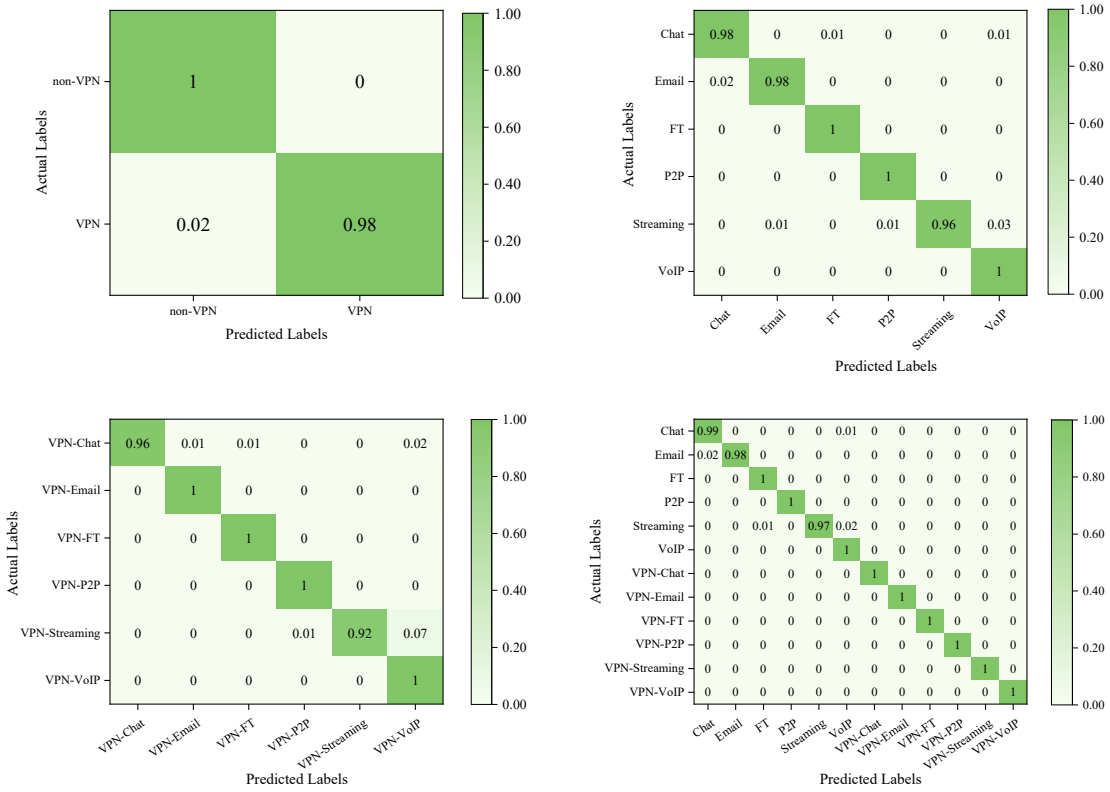


Figure 9. Confusion matrices of the MFF model across four experimental settings.

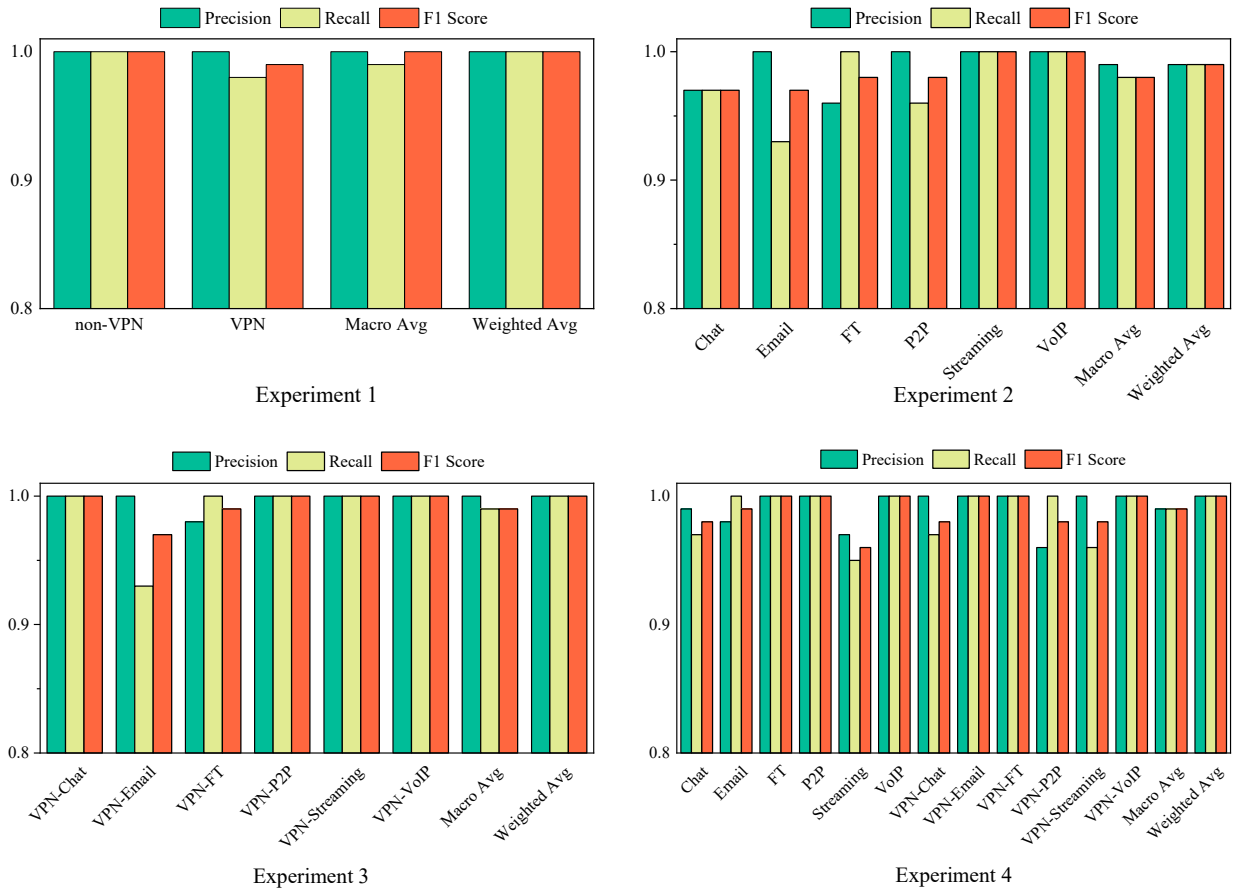


Figure 10. Detailed classification results for the four experimental settings.

The experimental results clearly show that when the SE (Squeeze-and-Excitation) module and ResNet module are removed, the model's classification performance significantly declines, especially in Experiment 2 and Experiment 4, where the accuracy drops to 75.25% and 73.76%, respectively. This phenomenon indicates that the temporal semantic features extracted by SE-ResNet play a key role in encrypted traffic classification, particularly in modeling local dynamic behaviors, such as burst patterns and packet interaction rhythms.

Furthermore, when only the AE (autoencoder) module is removed, while the complete temporal modeling structure is retained, the model's accuracy decreases across all four experimental setups, further confirming the complementary value of the AE module in overall feature representation. The global statistical features introduced by AE enhance the model's macro-level perception of traffic patterns, thereby improving the completeness and generalization ability of feature representation.

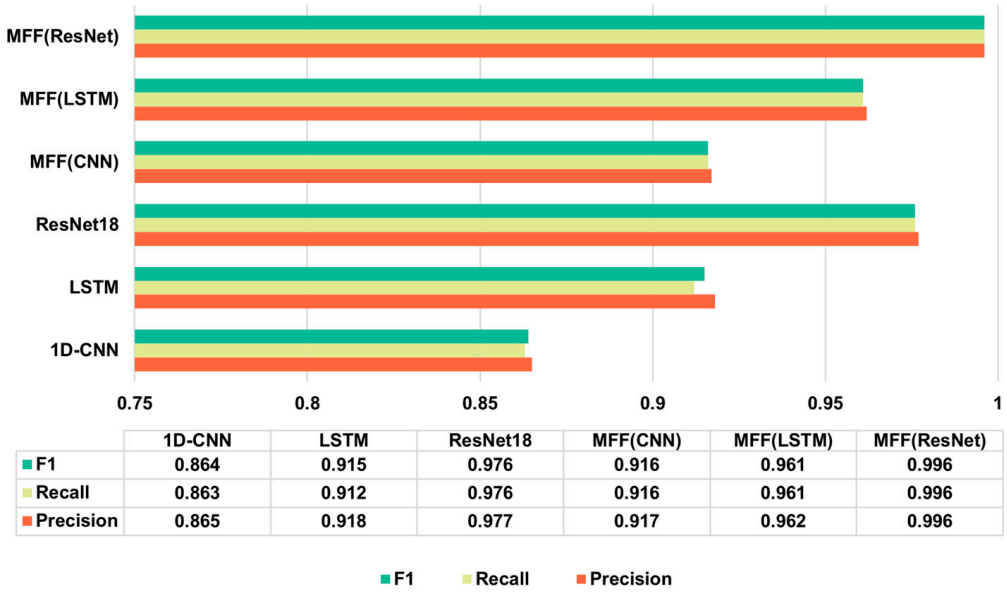
With the full MFF structure, classification performance reaches its optimal level in all experimental configurations. For example, in Experiment 1 and Experiment 4, the accuracy reaches 99.93% and 99.61%, respectively. This excellent performance is attributed to MFF's ability to synergistically fuse local temporal information with global statistical features. Although statistical features are used as auxiliary inputs, they effectively complement the dominant temporal features, significantly enhancing the model's feature coverage and discriminative robustness in complex encrypted traffic scenarios.

In summary, the ablation experiments clearly validate the specific roles and synergistic effects of each functional module within the MFF framework: SE-ResNet strengthens temporal modeling, thereby improving the model's discriminative ability, while the AE module introduces global information that enhances feature diversity and generalization.

Based on this, the multi-feature fusion strategy employed in MFF effectively boosts the overall classification performance in encrypted traffic analysis.

To evaluate the impact of different temporal feature extraction modules under the multi-feature fusion framework, this study constructs six temporal modeling structures—1D-CNN, LSTM, ResNet18, MFF(CNN), MFF(LSTM), and MFF(ResNet)—while keeping all other components and training parameters unchanged. A 12-class classification experiment was conducted on the ISCX VPN-nonVPN dataset.

As shown in Figure 11, the ResNet18 model outperforms both 1D-CNN and LSTM in terms of Precision, Recall, and F1-score, demonstrating superior capability in temporal feature representation. Specifically, the F1-score improves from 86.4% with 1D-CNN to 97.6% with ResNet18, indicating that the deep residual structure offers significant advantages in modeling complex encrypted traffic sequences.



	1D-CNN	LSTM	ResNet18	MFF(CNN)	MFF(LSTM)	MFF(ResNet)
F1	0.864	0.915	0.976	0.916	0.961	0.996
Recall	0.863	0.912	0.976	0.916	0.961	0.996
Precision	0.865	0.918	0.977	0.917	0.962	0.996

Figure 11. Classification performance comparison of different temporal feature extraction models (1D-CNN, LSTM, ResNet18) and their MFF-based variants on the ISCX VPN-nonVPN dataset.

By analyzing the results of MFF(CNN), MFF(LSTM), and MFF(ResNet), we observe that the inclusion of AE and SE consistently improves model performance. For instance, compared to their respective base models (1D-CNN, LSTM, and ResNet18), the F1-scores of MFF(CNN), MFF(LSTM), and MFF(ResNet) are increased by 5.2%, 4.6%, and 2%, respectively, validating the effectiveness of AE and SE attention mechanisms in enhancing multi-feature fusion and achieving feature enhancement. Furthermore, MFF(ResNet) outperforms MFF(CNN) and MFF(LSTM) with F1-score improvements of 8% and 3.5%, respectively, making it the best-performing multi-feature fusion model.

Based on these results, ResNet18 is selected as the temporal feature extractor in the final model to ensure stable performance and high accuracy of the overall architecture.

4.4.2. Comparative Experiments and Result Analysis with Other Models

The MFF model is compared with a range of state-of-the-art methods, covering traditional fingerprint-based approaches, statistical feature-based methods, deep learning models, and pre-trained architectures. Specifically, the following models are selected as baselines: FlowPrint [13], AppScanner [17], DeepPacket [21], PERT [37], ET-BERT [23], ICLSTM [15], TFE-GNN [24], CMTSNN [38], Flow-GNN [39], ATVITSC [40], NetMamba [41], and LAMBERT [42].

Experiments are conducted under two classification scenarios: a 12-class task on the ISCX VPN-nonVPN dataset and a 20-class task on the USTC-TFC dataset. The performance of each method is comprehensively evaluated using four metrics: accuracy, precision, recall, and F1-score. The detailed results are presented in Tables 9 and 10 and Figure 12.

Table 9. Comparison results on the ISCX VPN-nonVPN dataset.

Method	Accuracy	Precision	Recall	F1
AppScanner [17]	0.7182	0.7339	0.7225	0.7197
FlowPrint [13]	0.7962	0.8042	0.7812	0.7820
DeepPacket [21]	0.9329	0.9377	0.9306	0.9321
PERT [37]	0.9352	0.9400	0.9349	0.9368
ET-BERT [23]	0.9890	0.9891	0. 9890	0. 9890
ICLSTM [15]	0.981	0.98	0.98	0.981
TFE-GNN [24]	0.9591	0.9526	0.9593	0.9536
ATVITSC [40]	0.9789	0.9789	0.9788	0.9789
NetMamba [41]	0.9899	0.9899	0.9899	0.9899
LAMBERT [42]	0.9915	0.9917	0.9915	0.9915
MFF	0.9961	0.9961	0.9961	0.9961

Table 10. Comparison results on the USTC-TFC dataset.

Method	Accuracy	Precision	Recall	F1
AppScanner [17]	0.8954	0.8984	0.8968	0.8892
FlowPrint [13]	0.8146	0.6434	0.7002	0.6573
Deeppacket [21]	0.9640	0.9650	0.9631	0.9641
PERT [37]	0.9909	0.9911	0.9910	0.9911
ET-BERT [23]	0.9929	0.9930	0.9930	0.9930
LAMBERT [42]	0.9930	0.9931	0.9930	0.9930
CMTSNN [38]	0.9876	0.9884	0.9881	0.9855
Flow-GNN [39]	0.9970	0.9959	0.9961	0.9974
ATVITAC [40]	0.9966	0.9967	0.9967	0.9966
NetMamba [41]	0.9990	0.9991	0.9990	0.9990
MFF	0.9999	0.9999	0.9999	0.9999

Note: Some results of the baseline methods in Tables 9 and 10 are directly cited from the original papers. The reported F1-scores may slightly differ from those calculated using Equation (16), due to differences in the averaging strategy (e.g., weighted vs. macro average) adopted by the original authors.

Based on the experimental results, the following conclusions can be drawn:

On the ISCX VPN-nonVPN dataset, the proposed MFF model demonstrates outstanding classification performance. Compared to the classical method DeepPacket, MFF achieves improvements of 6.32% in accuracy and 6.40% in F1-score. Even when compared to larger deep learning models such as ET-BERT and LAMBERT, MFF outperforms them by 0.71% and 0.46% in F1-score, respectively. It is worth noting that MFF requires significantly fewer parameters and computational resources than these large-scale models yet still delivers superior performance in this task. This indicates that MFF achieves a favorable balance between feature extraction capability and model efficiency, enabling more effective differentiation among encrypted traffic types.

On the USTC-TFC dataset, MFF also exhibits robust performance, achieving 99.99% across all four evaluation metrics. Although ATVITAC and NetMamba also perform well on this task, MFF still holds a slight advantage in each metric, further demonstrating the adaptability and effectiveness of the proposed multimodal fusion strategy in handling complex traffic classification tasks.

In summary, the experimental findings validate the effectiveness of the multimodal feature fusion strategy employed in MFF. By jointly leveraging global statistical features and temporal semantic representations, the model enhances its capacity to express critical characteristics while preserving information completeness, thereby improving classification accuracy and delivering more reliable encrypted traffic identification.

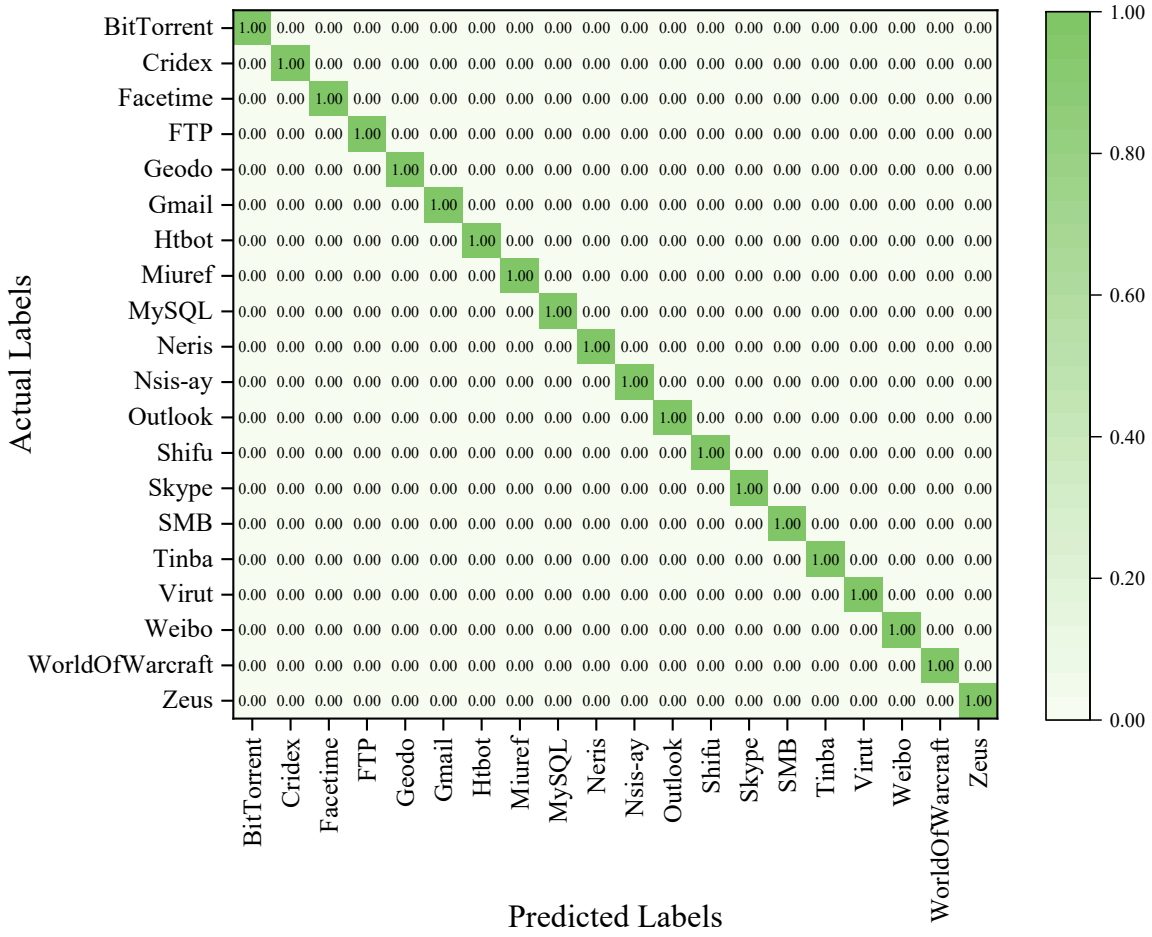


Figure 12. Confusion matrix results for experiment 7.

4.4.3. Interpretability Analysis

To enhance the interpretability of the proposed model, we conduct a SHAP-based analysis to visualize the decision rationale of MFF across different classification tasks [43,44]. The analysis is performed on two datasets: ISCX VPN-nonVPN (12-class classification) and USTC-TFC (binary classification of benign vs. malicious traffic), as illustrated in Figures 13 and 14.

The MFF model takes as input a combination of temporal features (feature indices 1–1024) and statistical features (indices 1025–1076). The former are extracted via a ResNet18 network integrated with a Squeeze-and-Excitation (SE) attention mechanism, while the latter are reconstructed from raw statistical indicators using an autoencoder.

The SHAP feature ranking results reveal that temporal features dominate in both classification tasks. For example, on the ISCX dataset, 18 out of the top 20 most important features are temporal features, with Seq_Feat26 and Seq_Feat27 being particularly influential. In the USTC-TFC binary task, all of the top 20 features are temporal, demonstrating the consistent discriminative power of temporal modeling across different traffic scenarios. While the dominant temporal features vary between tasks (e.g., Seq_Feat100 and

Seq_Feat25 play key roles in binary classification), this variability highlights the model's adaptability and task-level generalization capability.

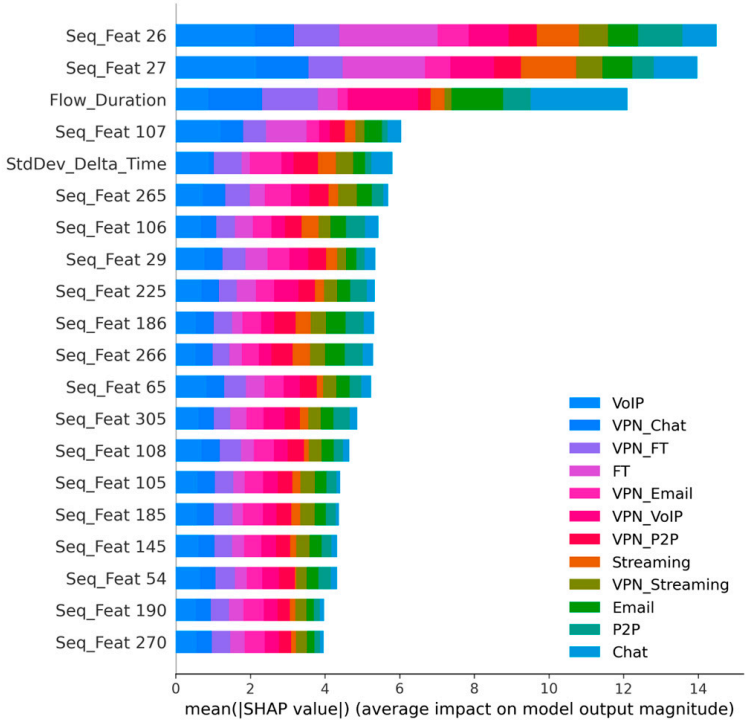


Figure 13. Feature importance visualization for the 12-class task on the ISCX VPN-nonVPN dataset.

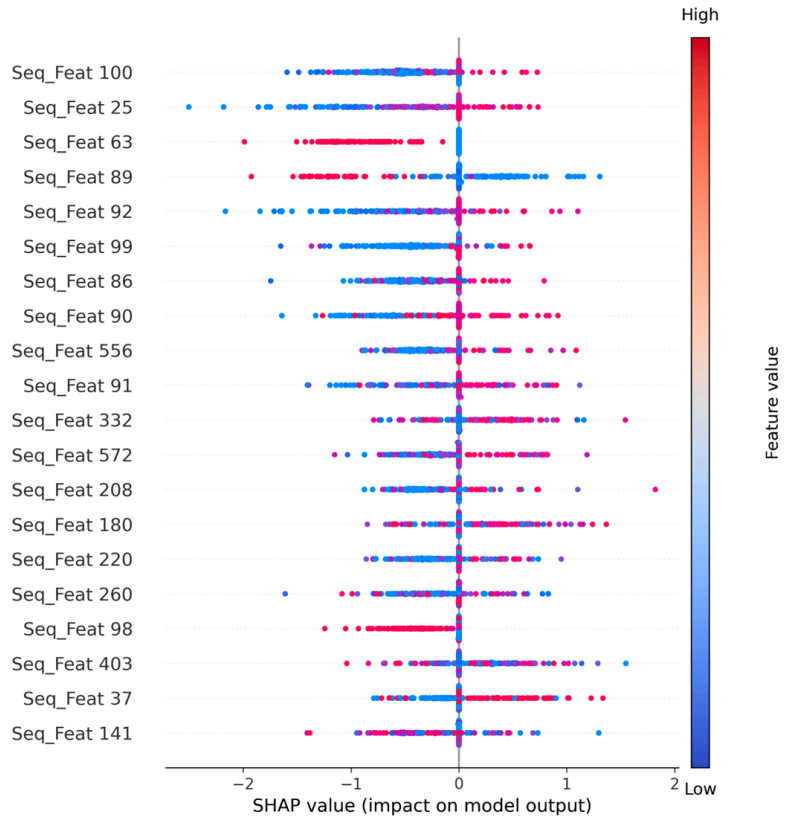


Figure 14. Feature importance visualization for the binary classification task on the USTC-TFC dataset.

Although statistical features are generally ranked lower in importance, they exhibit non-negligible contributions in certain class-level distinctions. In the 12-class task on the ISCX dataset, two statistical features—Flow_Duration and StdDev_Delta_Time—appear among the top five, helping to differentiate classes such as VPN-P2P and VoIP, which may share similar temporal patterns but differ in protocol behaviors. This suggests that statistical features act as global semantic compensators, enhancing the model's ability to identify fine-grained differences.

In summary, the SHAP analysis validates the effectiveness of MFF's multimodal feature fusion strategy. Temporal features serve as dominant predictors, while statistical features provide structural complements in specific semantic contexts. The synergy between these two modalities not only improves the model's discriminative power but also enhances interpretability and result stability.

5. Conclusions and Future Work

Encrypted traffic classification has become a crucial topic in network security. With the increasing adoption of deep learning, deep model-based traffic classification approaches have attracted growing attention due to their promising performance [3,45]. However, most existing methods rely on truncating input data during preprocessing, which often results in the loss of critical information and degrades classification accuracy. To address this issue, we propose a Multi-feature Fusion Framework (MFF) that jointly models temporal and statistical features to achieve feature enhancement and improve classification precision. Experimental results demonstrate that MFF offers both high effectiveness and robustness in encrypted traffic classification tasks.

The MFF model comprises three main stages: data preprocessing, feature extraction, and feature fusion. In the extraction stage, temporal features are learned via a ResNet18 architecture enhanced with Squeeze-and-Excitation (SE) attention to emphasize key dynamic patterns. Meanwhile, a deep autoencoder is employed to nonlinearly reduce and reconstruct 52-dimensional session-level statistical features, preserving macro-level semantic information. These two feature types are then integrated through a unified mapping and fed into a classifier. Under four evaluation settings on the ISCX VPN-nonVPN dataset, MFF consistently outperforms traditional machine learning and deep learning baselines in terms of accuracy and robustness, highlighting its architectural advantages. The performance improvement stems from two key factors: the effective modeling of temporal dependencies via 1D convolutions, and the fusion of local and global features to mitigate context loss caused by input truncation. Additional experiments on the USTC-TFC dataset further validate MFF's stability and generalizability across diverse scenarios, achieving 99.99% accuracy in a 20-class classification task.

For future work, we aim to enhance MFF in the following directions:

(1) Incorporate advanced temporal modeling techniques, such as transformer architectures or hybrid attention mechanisms, to better capture long-range dependencies in encrypted traffic;
(2) Investigate the impact of varying input segment lengths (e.g., 512, 1024, 2048 bytes) on model performance to improve adaptability and generalization;
(3) Explore model interpretability and lightweight design to improve deployability and efficiency, particularly in edge computing environments;
(4) Extend support for emerging encryption protocols such as TLS 1.3 and QUIC by adapting the model architecture and expanding the dataset using automated traffic collection tools. These enhancements will further improve MFF's practical relevance in real-world encrypted traffic analysis.

Author Contributions: Conceptualization, H.H., Y.Z. and F.J.; data curation, Y.Z., X.Z. and Q.J.; formal analysis, Y.Z. and Q.J.; funding acquisition, H.H. and F.J.; investigation, Y.Z., X.Z. and Q.J.; methodology, H.H., Y.Z. and F.J.; project administration, H.H. and F.J.; resources, H.H., F.J. and X.Z.; software, Y.Z.; supervision, H.H.; validation, H.H., Y.Z., X.Z. and Q.J.; visualization, Y.Z.; writing—original draft, Y.Z.; writing—review and editing, H.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Laboratory Project of Enterprise Informatization and IoT Measurement and Control Technology for Universities in Sichuan Province (NO: 2024WYJ06), Central Guidance for Local Science and Technology Development Fund Projects (NO: 2024ZYD0266), Tibet Science and Technology Program (NO: XZ202401YD0023).

Data Availability Statement: The original contributions presented in this study are included in the article. For further inquiries, please contact the corresponding author.

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

References

1. HTTPS Encryption on the Web. Available online: https://transparencyreport.google.com/https/overview?hl=en (accessed on 2 January 2025).
2. Dong, W.; Yu, J.; Lin, X.; Gou, G.; Xiong, G. Deep learning and pre-training technology for encrypted traffic classification: A comprehensive review. *Neurocomputing* **2024**, *617*, 128444. [CrossRef]
3. Feng, Y.; Li, J.; Mirkovic, J.; Wu, C.; Wang, C.; Ren, H.; Xu, J.; Liu, Y. Unmasking the Internet: A Survey of Fine-Grained Network Traffic Analysis. *IEEE Commun. Surv. Tutor.* **2025**, *early access*. [CrossRef]
4. Wang, Z.; Yang, Y.; Wang, Y. A survey of encrypted traffic classification: Datasets, representation, approaches and future thinking. In Proceedings of the 2024 IEEE/ACIS 24th International Conference on Computer and Information Science (ICIS), Shanghai, China, 20–22 September 2024; pp. 113–120.
5. Fernandes, S.; Antonello, R.; Lacerda, T.; Santos, A.; Sadok, D.; Westholm, T. Slimming down deep packet inspection systems. In Proceedings of the IEEE INFOCOM Workshops 2009, Rio de Janeiro, Brazil, 19–25 April 2009; pp. 1–6.
6. Wang, X.; Jiang, J.; Tang, Y.; Liu, B.; Wang, X. StriD^{2}FA: Scalable Regular Expression Matching for Deep Packet Inspection. In Proceedings of the 2011 IEEE International Conference on Communications (ICC), Kyoto, Japan, 5–9 June 2011; pp. 1–5.
7. Alwhbi, I.A.; Zou, C.C.; Alharbi, R.N. Encrypted network traffic analysis and classification utilizing machine learning. *Sensors* **2024**, *24*, 3509. [CrossRef] [PubMed]
8. Draper-Gil, G.; Lashkari, A.H.; Mamun, M.S.I.; Ghorbani, A.A. Characterization of encrypted and vpn traffic using time-related. In Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP), Rome, Italy, 19–21 February 2016; pp. 407–414.
9. Wang, W.; Zhu, M.; Wang, J.; Zeng, X.; Yang, Z. End-to-end encrypted traffic classification with one-dimensional convolution neural networks. In Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, China, 22–24 July 2017; pp. 43–48.
10. Alotaibi, F.M. Network Intrusion Detection Model Using Fused Machine Learning Technique. *Comput. Mater. Contin.* **2023**, *75*, 2479–2490.
11. Meng, X.; Lin, C.; Wang, Y.; Zhang, Y. Netgpt: Generative pretrained transformer for network traffic. *arXiv* **2023**, arXiv:2304.09513.
12. Wang, W.; Zhu, M.; Zeng, X.; Ye, X.; Sheng, Y. Malware traffic classification using convolutional neural network for representation learning. In Proceedings of the 2017 International Conference on Information Networking (ICOIN), Da Nang, Vietnam, 11–13 January 2017; pp. 712–717.
13. Van Ede, T.; Bortolameotti, R.; Continella, A.; Ren, J.; Dubois, D.J.; Lindorfer, M.; Choffnes, D.; van Steen, M.; Peter, A. Flowprint: Semi-supervised mobile-app fingerprinting on encrypted network traffic. In Proceedings of the Network and Distributed System Security Symposium (NDSS), San Diego, CA, USA, 23–26 February 2020; Volume 27.
14. Hayes, J.; Danezis, G. k-fingerprinting: A robust scalable website fingerprinting technique. In Proceedings of the 25th USENIX Security Symposium (USENIX Security 16), Vancouver, BC, Canada, 16–18 August 2017; pp. 1187–1203.
15. Lu, B.; Luktarhan, N.; Ding, C.; Zhang, W. ICLSTM: Encrypted traffic service identification based on inception-LSTM neural network. *Symmetry* **2021**, *13*, 1080. [CrossRef]
16. De Lucia, M.J.; Cotton, C. Detection of encrypted malicious network traffic using machine learning. In Proceedings of the MILCOM 2019-2019 IEEE Military Communications Conference (MILCOM), Norfolk, VA, USA, 12–14 November 2019; pp. 1–6.
17. Taylor, V.F.; Spolaor, R.; Conti, M.; Martinovic, I. Robust smartphone app identification via encrypted network traffic analysis. *IEEE Trans. Inf. Forensics Secur.* **2017**, *13*, 63–78. [CrossRef]

18. Cao, J.; Fang, Z.; Qu, G.; Sun, H.; Zhang, D. An accurate traffic classification model based on support vector machines. *Int. J. Netw. Manag.* **2017**, *27*, e1962. [CrossRef]
19. Dusi, M.; Este, A.; Gringoli, F.; Salgarelli, L. Using GMM and SVM-based techniques for the classification of SSH-encrypted traffic. In Proceedings of the 2009 IEEE International Conference on Communications, Dresden, Germany, 14–18 June 2009; pp. 1–6.
20. Shen, M.; Ye, K.; Liu, X.; Zhu, L.; Kang, J.; Yu, S.; Li, Q.; Xu, K. Machine learning-powered encrypted network traffic analysis: A comprehensive survey. *IEEE Commun. Surv. Tutor.* **2022**, *25*, 791–824. [CrossRef]
21. Lotfollahi, M.; Jafari Siavoshani, M.; Shirali Hossein Zade, R.; Saberian, M. Deep packet: A novel approach for encrypted traffic classification using deep learning. *Soft Comput.* **2020**, *24*, 1999–2012. [CrossRef]
22. Lin, K.; Xu, X.; Gao, H. TSCRNN: A novel classification scheme of encrypted traffic based on flow spatiotemporal features for efficient management of IIoT. *Comput. Netw.* **2021**, *190*, 107974. [CrossRef]
23. Lin, X.; Xiong, G.; Gou, G.; Li, Z.; Shi, J.; Yu, J. Et-bert: A contextualized datagram representation with pre-training transformers for encrypted traffic classification. In Proceedings of the ACM Web Conference 2022, Lyon, France, 25–29 April 2022; pp. 633–642.
24. Zhang, H.; Yu, L.; Xiao, X.; Li, Q.; Mercaldo, F.; Luo, X.; Liu, Q. Tfe-gnn: A temporal fusion encoder using graph neural networks for fine-grained encrypted traffic classification. In Proceedings of the ACM Web Conference 2023, Austin, TX, USA, 30 April–4 May 2023; pp. 2066–2075.
25. Zhao, J.; Jing, X.; Yan, Z.; Pedrycz, W. Network traffic classification for data fusion: A survey. *Inf. Fusion* **2021**, *72*, 22–47. [CrossRef]
26. Azab, A.; Khasawneh, M.; Alrabaee, S.; Choo, K.K.R.; Sarsour, M. Network traffic classification: Techniques, datasets, and challenges. *Digit. Commun. Netw.* **2024**, *10*, 676–692. [CrossRef]
27. Sharma, A.; Lashkari, A.H. A survey on encrypted network traffic: A comprehensive survey of identification/classification techniques, challenges, and future directions. *Comput. Netw.* **2024**, *257*, 110984. [CrossRef]
28. Wei, N.; Yin, L.; Zhou, X.; Ruan, C.; Wei, Y.; Luo, X.; Chang, Y.; Li, Z. A feature enhancement-based model for the malicious traffic detection with small-scale imbalanced dataset. *Inf. Sci.* **2023**, *647*, 119512. [CrossRef]
29. Miao, G.; Wu, G.; Zhang, Z.; Tong, Y.; Lu, B. Boosting Encrypted Traffic Classification Using Feature-Enhanced Recurrent Neural Network with Angle Constraint. *IEEE Trans. Big Data* **2024**. *preprints*. [CrossRef]
30. Huang, H.; Zhang, X.; Lu, Y.; Li, Z.; Zhou, S. BSTFNet: An Encrypted Malicious Traffic Classification Method Integrating Global Semantic and Spatiotemporal Features. *Comput. Mater. Contin.* **2024**, *78*, 3929–3951. [CrossRef]
31. Maonan, W.; Kangfeng, Z.; Ning, X.; Yanqing, Y.; Xiujuan, W. CENTIME: A direct comprehensive traffic features extraction for encrypted traffic classification. In Proceedings of the 2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS), Chengdu, China, 23–26 April 2021; pp. 490–498.
32. Shi, Z.; Luktarhan, N.; Song, Y.; Tian, G. BFCN: A novel classification method of encrypted traffic based on BERT and CNN. *Electronics* **2023**, *12*, 516. [CrossRef]
33. Wang, M.; Zheng, K.; Luo, D.; Yang, Y.; Wang, X. An encrypted traffic classification framework based on convolutional neural networks and stacked autoencoders. In Proceedings of the 2020 IEEE 6th International Conference on Computer and Communications (ICCC), Chengdu, China, 11–14 December 2020; pp. 634–641.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
35. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
36. Jin, X.; Xie, Y.; Wei, X.S.; Zhao, B.R.; Chen, Z.M.; Tan, X. Delving deep into spatial pooling for squeeze-and-excitation networks. *Pattern Recognit.* **2022**, *121*, 108159. [CrossRef]
37. He, H.Y.; Yang, Z.G.; Chen, X.N. PERT: Payload encoding representation from transformer for encrypted traffic classification. In Proceedings of the 2020 ITU Kaleidoscope: Industry-Driven Digital Transformation (ITU K), Ha Noi, Vietnam, 7–11 December 2020; pp. 1–8.
38. Zhu, S.; Xu, X.; Gao, H.; Xiao, F. CMTSNN: A deep learning model for multiclassification of abnormal and encrypted traffic of internet of things. *IEEE Internet Things J.* **2023**, *10*, 11773–11791. [CrossRef]
39. Huoh, T.L.; Luo, Y.; Li, P.; Zhang, T. Flow-based encrypted network traffic classification with graph neural networks. *IEEE Trans. Netw. Serv. Manag.* **2022**, *20*, 1224–1237. [CrossRef]
40. Liu, Y.; Wang, X.; Qu, B.; Zhao, F. ATVITSC: A novel encrypted traffic classification method based on deep learning. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 9374–9389. [CrossRef]
41. Wang, T.; Xie, X.; Wang, W.; Wang, C.; Zhao, Y.; Cui, Y. Netmamba: Efficient network traffic classification via pre-training unidirectional mamba. In Proceedings of the 2024 IEEE 32nd International Conference on Network Protocols (ICNP), Charleroi, Belgium, 28–31 October 2024; pp. 1–11.
42. Liu, T.; Ma, X.; Liu, L.; Liu, X.; Zhao, Y.; Hu, N.; Ghafoor, K.Z. LAMBERT: Leveraging Attention Mechanisms to Improve the BERT Fine-Tuning Model for Encrypted Traffic Classification. *Mathematics* **2024**, *12*, 1624. [CrossRef]

43. Mosca, E.; Szigeti, F.; Tragianni, S.; Gallagher, D.; Groh, G. SHAP-based explanation methods: A review for NLP interpretability. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 4593–4603.
44. Shanmugam, V.; Razavi-Far, R.; Hallaji, E. Addressing Class Imbalance in Intrusion Detection: A Comprehensive Evaluation of Machine Learning Approaches. *Electronics* **2024**, *14*, 69. [CrossRef]
45. Mosaiyebzadeh, F.; Pouriyeh, S.; Han, M.; Liu, L.; Xie, Y.; Zhao, L.; Batista, D.M. Privacy-Preserving Federated Learning-Based Intrusion Detection System for IoHT Devices. *Electronics* **2024**, *14*, 67. [CrossRef]