MEMTD: Encrypted Malware Traffic Detection Using Multimodal Deep Learning

Xiaotian Zhang, Jintian Lu, Jiakun Sun, Ruizhi Xiao, and Shuyuan Jin(✉)

School of Computer Science and Engineering, Sun Yat-sen University,
Guangzhou, China
{zhangxt73,lujd6,sunjk3,xiaorzh3}@mail2.sysu.edu.cn,
jinshuyuan@mail.sysu.edu.cn

Abstract. Malware that generates encrypted traffic presents a great threat to Internet security. The existing state-of-the-art malware traffic detection techniques based on deep learning (DL) ignore the heterogeneity of encrypted traffic, resulting in their inability to further improve detection performance. This paper applies multimodal DL to detect encrypted malware traffic, proposing a multimodal encrypted malware traffic detection (MEMTD) approach. MEMTD extracts features from three types of modal data—the transport layer security (TLS) handshake payload bytes (encryption behavior modal data), packet length sequence (spatial modal data), and packet arrival-time interval sequence (time modal data) of encrypted traffic. Moreover, an intermediate fusion mechanism is adopted in the MEMTD approach to mine the dependencies among modalities and fuse the discriminative traffic features, improving detection performance. The experimental results on datasets containing 8 malware families and normal traffic show that the MEMTD approach achieves 0.9996 macro-F1 and outperforms other single-modal DL detection methods.

Keywords: Malware traffic detection · Encrypted traffic · Multimodal deep learning · Intermediate fusion mechanism

1 Introduction

Encrypted malware traffic detection is a challenge in cyberspace security. Software what intentionally executes malicious payloads on victim computers is considered malware [1]. There are different types of malware families or malicious software, including viruses, botnets, Trojan horses, etc., which have caused great damage to the property and privacy of Internet users. Due to the wide usage of encryption techniques, malware adopts transport layer security (TLS) to hide its malicious attempts, which makes the detection of malware traffic more difficult.

To detect encrypted malware traffic, some researchers have utilized machine learning (ML) and deep learning (DL) algorithms. ML-based detection combines artificially designing statistical features (e.g., maximum packet length) and

T. Di Noia et al. (Eds.): ICWE 2022, LNCS 13362, pp. 357–372, 2022.
https://doi.org/10.1007/978-3-031-09917-5_24

selecting the ML model (e.g., logistic regression) to classify encrypted traffic into different malware families [2–5]. This kind of detection decomposes the encrypted traffic classification problem into two subproblems, i.e., feature engineering and model training driven by domain experts. The result of each subproblem directly affect final classification performance [6], resulting in that ML-based detection cannot guarantee the best classification performance. Moreover, ML-based detection requires expert experience. In recent years, DL has emerged, allowing for the combination of feature design and model training in an end-to-end model and the automatic learning of complex feature representations [6–12]. Generally, encrypted malware traffic detection methods based on DL employ only one modal type of encrypted traffic, such as packet payload bytes or header fields, to classify encrypted traffic. These single-modal detection methods based on DL have focused on designing complex deep neural networks instead of considering other modalities of encrypted traffic to improve detection performance.

The above researchers employing DL-based detection have ignored the heterogeneity of traffic data, which allows for the conversion of network traffic into multimodal data. As multimodal data, the payload of packets exchanged during the handshake phase, the packet length sequence (PLS), and the packet arrival-time interval sequence (PAIS) can contribute to the detection of encrypted malware traffic. First, TLS handshake payload bytes (HPBs) can be employed to extract traffic information about the encryption suite and security degree, i.e., encryption behavior modal information. Second, the PLS can be utilized to extract the information of the packet length changes in traffic flow, i.e., spatial modal information. Finally, the PAIS can represent the information of the arrival time changes in the flow, i.e., time modal information. In this context, the multimodal DL method is suitable for extracting features and classifying encrypted network traffic since can exploit all the available modal information of network traffic jointly to obtain a hierarchical representation.

This paper proposes a multimodal encrypted malware traffic detection approach, namely MEMTD, which jointly extracts features from the TLS HPBs, PLS, and PAIS of encrypted traffic. Moreover, the different modal feature extraction networks in MEMTD—ConvNet, GruNet-1, and GruNet-2—are designed to learn the representation of different modal information. While maintaining high-precision performance, MEMTD has better robustness because feature representation is enhanced by multiple modal inputs. MEMTD further adopts an intermediate fusion mechanism to avoid overfitting and any possible learning failures in representing the associations between modalities.

Our contributions can be briefly summarized as follows:

- We propose an MEMTD approach to detect encrypted malware traffic. This approach extracts three types of modal information—encryption behavior, spatial, and time modal information—from the encrypted traffic, improving detection performance with enhanced feature representation.
- We adopt the intermediate fusion mechanism to improve the utilization of multimodal features. The intermediate fusion mechanism has the capability of learning the dependencies between modalities, which employs two fusion networks to fuse multimodal features gradually.

- The experimental results demonstrate that the MEMTD approach achieves high performance with 99.94% accuracy (ACC) and 99.96% macro-F1 indicators.

2 Related Works

2.1 ML-Based Detection

With the advent of ML technology, researchers have tried to classify encrypted traffic without decryption. Anderson et al. proposed for the first time that unencrypted TLS headers can be used together with packet statistics as features of encrypted traffic to identify malware families, with l1-multinomial logistic regression [2]. Shekhawat et al. leveraged TLS handshake and statistical features to represent encrypted traffic and used random forest algorithms to design a two-layer detection framework for fast benign traffic filtering and malware traffic classification [3]. A feature set that included a modified NIST testing suite to represent the randomness of the content was proposed to identify the traffic protocol [4]. Moreover, [5] was based on random forest algorithms and designed the features of packet information, time, transmission control protocol (TCP) flag field, and application layer load information for identifying malicious encrypted traffic. The above ML-based method required the manual design of traffic features and could not solve the encrypted traffic classification in an end-to-end manner. However, various research views concerning the analysis of encrypted traffic have been proposed in these works, contributing to the understanding of the heterogeneity of encrypted traffic.

2.2 DL-Based Detection

To avoid the limitation of manually designing encrypted traffic features, the researchers used DL technology to design the end-to-end detection model for encrypted traffic classification. FS-Net with multilayer encoder-decoder structure built by Liu et al. can input the PLSs of raw flows to achieve excellent encrypted traffic classification performance [6]. In another work, Bi-gated recurrent unit (GRU) was also used to implement feature extraction, but unlike FS-Net with a reconstruction mechanism, the method proposed in [7] added an attention mechanism to learn the local flow information. Dong et al. proposed the CETAnalytics framework, which adopts packet payload bytes as input to classify encrypted traffic and implemented the framework through a 1-dimensional convolutional neural network (1D-CNN) with residual structure and the Bi-GRU with an attention mechanism [8]. A traffic classification method that integrates long short-term memory (LSTM) and a CNN to identify traffic via three packet payloads at any position of the encrypted flow was proposed by [9]. Moreover, [10,11] adopted a CNN to build an end-to-end encrypted traffic classification neural network. For training with fewer samples or speeding up traffic recognition, many improvements were made to the network architecture.

In the abovementioned single-modal detection based on DL, many complex neural network structures have been adopted by researchers, but the improvement of distinction performance is limited. An alternative idea is multimodal DL, which extracts complementary information from the modalities and yields a richer representation since this method can produce much-improved performance compared to using a single modality [13]. Although multimodal DL has been widely used in computer vision [14], emotion recognition [15], and other fields, it is difficult to find research based on multimodal DL in the field of encrypted malicious traffic detection. The only application of multimodal DL methods in the field of traffic classification is the MIMETIC approach proposed by Giuseppe et al. [16]. However, the MIMETIC framework is designed for mobile network traffic classification and adopts only two modalities from traffic bytes as input, which does not verify its effectiveness in terms of malicious traffic detection tasks. To detect encrypted malware traffic, this paper proposes a multimodal DL method that can use the three different modalities jointly to represent features and adopt the intermediate fusion mechanism to enhance the ability to learn the dependencies between modalities.

3 Multimodal Encrypted Malware Traffic Detection

3.1 Overview

As shown in Fig. 1, the MEMTD approach contains three steps, namely, multimodal data preprocessing, different modal feature extraction, and intermediate fusion and detection.

In the first step, MEMTD converts the raw traffic into multimodal data, TLS HPBs, PLS, and PAIS. In the second step, ConvNet, GruNet-1, and GruNet-2 are employed to extract the different modal feature vectors. In the last step, MEMTD adopts an intermediate fusion mechanism to gradually fuse the three feature vectors and classify the encrypted traffic. Moreover, to improve the different modal feature representation ability, the MEMTD approach adopts a pretraining and fine-tuning training process.

3.2 Multimodal Data Preprocessing

The first step of the MEMTD approach is the multimodal data preprocessing of the traffic flow. A flow is a set of packets sharing 5 tuples (i.e., the IP of the source, IP of the destination, port of the source, port of the destination, and transport-level protocol), taking no account of their sending directions. To analyze flow in multiple modalities, we extract TLS HPBs, PLS, and PAIS from a flow as MEMTD input.

For notational convenience, an input flow to MEMTD is defined as follows:

$$F = [p_1, p_2, \ldots, p_{|F|}] \tag{1}$$

$$p_i = \{t_i, l_i, b_i, d_i\}, i \in \{1, 2, \ldots, |F|\}, t_1 < t_2 < \ldots < t_{|F|} \tag{2}$$

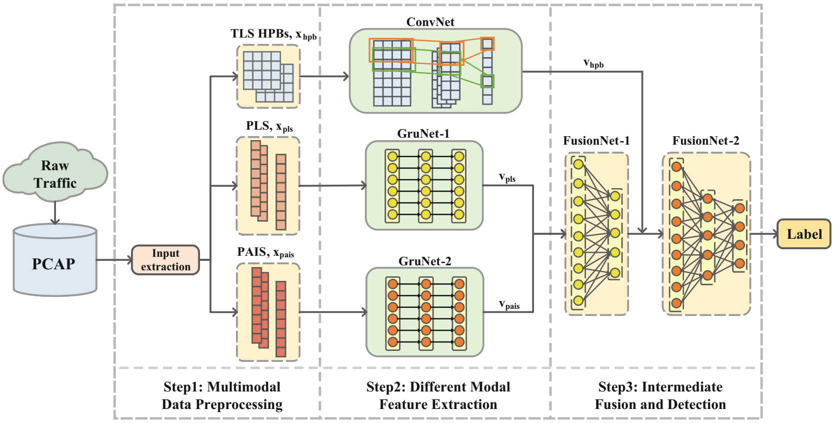


Fig. 1. Overview of the MEMTD approach.

where p_i stands for the *i-th* data packet in the flow, including the packet arrival-time $t_i \in [0, \infty]$ in seconds, the packet length $l_i \in \{1, 2, \ldots, 1500\}$ in bytes, the packet payload bytes $b_i = [b_i^1, b_i^2, \ldots, b_i^{l_i}], b_i^j \in \{0, 1, \ldots, 225\}, j = 1, 2, \ldots, l_i$, and the sending direction of the packet $d_i \in \{1, -1\}$ (1 and -1 indicate whether the sender is the client).

Unlike benign users, malware leverages TLS encryption technology for confusion rather than security, so malware developers tend to use older cipher suites than those utilized by enterprises when adopting TLS [2]. In addition, the unencrypted fields in the data packet sent during the TLS handshake phase, such as the protocol version of the TLS, the list of optional cipher suites provided by the client, and extension items, all can be considered discriminative information used to classify malware families. Therefore, the MEMTD approach introduces the HPBs, defined as $x_{hpb} = (b_{ch}, b_{sh})$, to obtain prominent features in the dimension of encryption behavior, where $b_{ch}, b_{sh} \in b_i, i \in [1, 2, \ldots, |F|]$ are the packets that carry client hello and server hello messages in the TLS handshake phase.

After the TLS handshake, the flow enters the data transmission stage, containing significant application layer information. However, the packet payload in the data transmission stage is encrypted and is thus not suitable to be directly input into the deep neural network. Although some researchers have employed encrypted packet payload bytes as the input of the CNN, it has been demonstrated that the network learns only the lengths of encrypted packets as features [12]. Therefore, PLS and PAIS are designed to record the flow information of all positions, including those in the data transmission stage. PLS is defined as a sequence with $|F|$ elements, $x_{pls} = [d_1 l_1, d_2 l_2, \ldots, d_{|F|} l_{|F|}]$ and $d_i, l_i \in p_i$. While x_{pls} includes the behavior information of a flow in the content dimension, PAIS describes a flow from the perspective of time, denoted as $x_{pais} = [a_1, a_2, \ldots, a_{|F|-1}]$, where the element a_i $(i = 1, 2, \ldots, |F| - 1)$ is calculated by the following:

$$a_i = d_{i+1}(t_{i+1} - t_i) \tag{3}$$

where $t_i \in p_i$ and $d_{i+1}, t_{i+1} \in p_{i+1}$. To format the inputs of the DL network, all the x_{hpb}, x_{pls}, and x_{pais} of traffic flows perform zero-padded and truncated operations in the input extraction step.

3.3 Different Modal Feature Extraction

The second step of MEMTD is different modal feature extraction, which includes three neural networks to develop the features of different modalities.

To capture the local information on the TLS HPBs, we design a ConvNet constructed with the translation-invariant 1D-CNN structure to accept x_{hpb}, as shown in Figure Fig. 2(a). Its first layer accepts a 256-dimensional vector, x_{hpb}, and embeds each element of x_{hpb} into a 60-dimensional vector to enrich the representation of information preserved in each byte. Then, the embedded vector is passed to the Conv Block composed of two 1D convolution layers with three convolution kernel lengths (i.e., 2, 4, and 6). In addition, the output of the first convolution operation is normalized and passed to a rectified linear unit (ReLU) function. Subsequently, after the processing of the maximum pooling layer connected to the Conv Block, the output vector with fewer dimensions becomes the feature vector of HPBs, v_{hpb}. Finally, v_{hpb} is sent to a classification layer to obtain the distribution over different malware families and benign users during pretraining, and the traffic category with the maximum probability is adopted as the output label. A multilayer perceptron with a softmax layer implements the classification layer. During fine-tuning, the v_{hpb} is sent to the FusionNet of MEMTD.

Different from extracting the field information in the packet payload by 1D-CNN, x_{pls} and x_{pais} are more appropriate for utilizing GRU to obtain the feature vector representing the entire flow. GRU and LSTM have long-term memory for the sequence, but GRU has fewer parameters than LSTM. Specifically, x_{pls} and x_{pais} utilize GruNet-1 and GruNet-2 to extract different modal features, respectively, and these two networks have the same structure, GruNet, but different parameters, as shown in Fig. 2(b). First, the embedding layer of GruNet accepts sequence data and selects x_{pls} or x_{pais} as input according to the different usages of traffic modalities. Moreover, three Bi-GRU layers are responsible for extracting the feature vector of sequence data, and their hidden state dimension is 80. The output of the embedding layer is defined as $s = [s_1, s_2, \ldots, s_L]$, where $L = 20$ represents the number of Bi-GRU units per layer. Each layer of the three-layer Bi-GRU has a forward $\overrightarrow{GRU}_i$ network and a backward $\overleftarrow{GRU}_i$ network, where $i \in \{1, 2, 3\}$ represents the network level. The calculation process of Bi-GRU relies on the hidden state to accumulate and transfer the information of sequence elements from two directions, as follows:

$$\overrightarrow{h}^{i}_{t} = \overrightarrow{GRU}_i(\overrightarrow{h}^{i}_{t-1}, (\overrightarrow{h}^{i-1}_{t}, \overleftarrow{h}^{i-1}_{t})), t \in \{1, 2, \ldots, L\}, i \in \{1, 2, 3\} \tag{4}$$

$$\overleftarrow{h}^{i}_{t} = \overleftarrow{GRU}_i(\overleftarrow{h}^{i}_{t-1}, (\overrightarrow{h}^{i-1}_{t}, \overleftarrow{h}^{i-1}_{t})), t \in \{1, 2, \ldots, L\}, i \in \{1, 2, 3\} \tag{5}$$

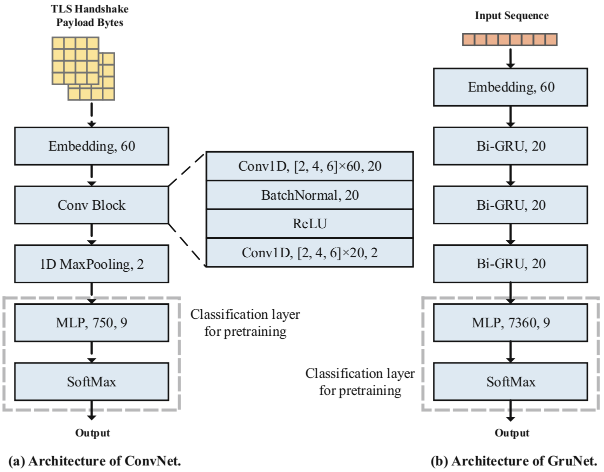


Fig. 2. Network structure used in the different modal feature extraction stage of MEMTD. Figure 2(a) shows the architecture of ConvNet, and Fig. 2(b) shows that of GruNet-1 and GruNet-2. The classification layer used for pretraining in the figure is removed in the fine-tuning stage of MEMTD.

where $\overrightarrow{h}_t^i$ and $\overleftarrow{h}_t^i$ represent the hidden states of $\overrightarrow{GRU}_i$ and $\overleftarrow{GRU}_i$ at time step t, respectively. States $\overrightarrow{h}_0^i, \overleftarrow{h}_{L+1}^i, i \in \{0,1,2\}$ are both zero vectors, and $\overrightarrow{h}_t^0 = s_t, \overleftarrow{h}_t^0 = s_t, t \in \{1,2,\ldots,L\}$.

The feature vector extracted by Bi-GRU, $v_{seq} = [\overrightarrow{h}_1^3, \overleftarrow{h}_1^3, \ldots, \overrightarrow{h}_L^3, \overleftarrow{h}_L^3, \overrightarrow{h}_L^1, \overleftarrow{h}_1^1, \overrightarrow{h}_L^2, \overleftarrow{h}_1^2]$, is designed as v_{pls} or v_{pais} according to the different instances, GruNet-1 or GruNet-2, respectively. Similar to ConvNet processing HPBs modal feature vectors, v_{pls} and v_{pais} are input to the multilayer perceptron with softmax during pretraining and then sent to the FusionNet of MEMTD during fine-tuning.

3.4 Intermediate Fusion and Detection

Through intramodal learning, ConvNet, GruNet-1, and GruNet-2 output the feature vectors of different modalities, v_{hpb}, v_{pls}, and v_{pais}, respectively, employed as the input of the third step of MEMTD. In this step, MEMTD processes feature fusion, and the shared presentation layers (i.e., FusionNet-1 and FushionNet-2) learn the dependencies of the modalities and classify the traffic as belonging to a benign or malware family.

MEMTD provides an intermediate fusion mechanism to combine modal feature vectors from different sources because simple splicing may lead to the insufficient usage of the extracted information [13]. First, the output vectors of GruNet-1 and GruNet-2, v_{pls} and v_{pais}, are fused by FusionNet-1 constructed with a multilayer perceptron. Then, the output vector of FusionNet-1 and the output vector of CovnNet, v_{hpb}, are sent to FusionNet-2, a multilayer perceptron

Algorithm 1: Training MEMTD

```
   Input: Sample sets
          $S^i = \{(s_1^i, y_1), (s_2^i, y_2), \ldots, (s_N^i, y_N)\}, i \in \{1, 2, 3\}, y_n \in \{1, 2, \ldots, K\}$
          where $s_n^1, s_n^2, s_n^3$ and $y_n$ represent the $x_{hpb}, x_{pls}, x_{pais}$ and label of the
          n-th sample respectively. The number of samples is $N$, and the number
          of traffic categories is $K$.
   Output: A trained instantiated MEMTD network, $Net$.
   Require: $ConvNet, GruNet, FusionNet$-1 and $FusionNet$-2.
1  $Net_1 \leftarrow ConvNet$
2  $Net_2 \leftarrow GruNet$ // i.e., GruNet-1
3  $Net_3 \leftarrow GruNet$ // i.e., GruNet-2
4  for $i$ in $\{1, 2, 3\}$ do
5      Train($Net_i, S^i$) // pretraining stage
6      $Net_i \leftarrow Net_i$ dropped MLP layer and softmax
7      Freezing parameters of $Net_i$
8  end
9  $Net \leftarrow$ Combine($Net_{1,2,3}, FusionNet$-1, 2)
10 Train($Net, S^{1,2,3}$) // fine-tuning stage
```

with a softmax layer added to the tail. Finally, the softmax layer outputs a vector representing the probability of the traffic category, from which the predicted label can be obtained. This label is the result of detecting encrypted malware traffic.

3.5 MEMTD Training Process

The training process of multimodal DL methods has two crucial stages: learning the features within each modality and representing the associations between modalities automatically. As displayed in Algorithm 1, the training process of the MEMTD approach includes pretraining and fine-tuning, which correspond to the aforementioned two stages.

First, three neural networks, Net_1, Net_2, and Net_3, are instantiated and correspond to three modal feature extraction networks, ConvNet, GruNet-1, and GruNet-2, respectively. Second, these networks are trained with HPBs, PLS, and PAIS sample sets to classify traffic families in the pretraining phase. Step 5 in Algorithm 1, Train(Net_i, S^i), represents the network training with the corresponding modal data and labels of the sample set. The trained ConvNet, GruNet-1, and GruNet-2 can be employed as single-modal classifiers to detect encrypted malicious traffic and be used to demonstrate the effectiveness of modalities in experiments. Finally, ConvNet, GruNet-1, and GruNet-2 remove the multilayer perceptron and are combined with FusionNet-1 and FusionNet-2 to form a multimodal DL network. MEMTD needs only to adopt the previous layers of the trained single-modal networks to represent the feature vectors because the lower layers of a network can extract the more general features, which can be transferred to other tasks [17]. In the fine-tuning stage, the MEMTD network needs only to train the parameters of the fusion layer and learn the fusion of multimodal feature vectors and final classification. Moreover, the loss function adopted in pretraining and fine-tuning is the cross-entropy function.

4 Evaluation

4.1 Dataset

To evaluate the performance of the MEMTD approach, we employ the Stratosphere Research Laboratory dataset from the cybersecurity group of the Artificial Intelligence Centre at the Czech Technical University in Prague as raw traffic [18]. This dataset was collected through a project responsible for long-term malware capture and provides malicious and benign traffic to the Stratosphere Intrusion Prevention System.

In the experiment, 47 pcap files were adopted, including 17 pcap files with benign traffic and 30 pcap files generated by 8 different malware families. Table 1 shows the number of secure sockets layer (SSL)/TLS flows in each family. We deliberately do not balance the number of flows between different families to imitate the actual network situation.

Table 1. Detailed dataset description

ID	Families	Pcaps	Flows
1	Bunitu	4	5,259
2	Caphaw	1	2,057
3	Dridex	3	111
4	HTBot	6	18,813
5	Miuref	3	2,499
6	Neris	3	4,365
7	TrickBot	4	522
8	Zbot	6	7,398
9	Normal	17	24,885
Total		47	65,909

4.2 Experiment Settlement

For a complete analysis, this paper includes the following methods as baselines:

- FS-Net [6] inputs the PLS into a DL model with a reconstruction mechanism for encrypted traffic classification, which is implemented by Bi-GRU.
- [9] proposed the building of a convolutional LSTM (CLSTM) neural network to classify encrypted traffic, utilizing the payload bytes of three consecutive packets in a flow. For convenience, CLSTM is used to represent [9].
- The three single-modal feature extraction networks in the MEMTD method, i.e., ConvNet, GruNet-1, and GruNet-2, are employed as independent single-modality DL networks to detect malware traffic.

- Different variants of the MEMTD approach, MEMTD-T and MEMTD-L, utilize v_{hpb} as the input of Fusion-1 and input v_{pls} and v_{pais} into Fusion-2, respectively. Furthermore, MEMTD without the intermediate fusion mechanism is instantiated as MEMTD-S.

We evaluate all methods based on ACC and macro-F1 indicators. The reason for using macro-F1 indicator is that it can evaluate the detection performance of the method in imbalanced data. The definitions are as follows:

$$ACC = \frac{\sum_{i=1}^{C}(TP_i + TN_i)}{\sum_{i=1}^{C}(TP_i + TN_i + FP_i + FN_i)} \tag{6}$$

$$precision_i = \frac{TP_i}{TP_i + FP_i}, recall_i = \frac{TP_i}{TP_i + FN_i} \tag{7}$$

$$macro\text{-}F1 = \frac{1}{C}\sum_{i=1}^{C} F1_i, F1_i = 2 \times \frac{recall_i \times precision_i}{recall_i + precision_i} \tag{8}$$

where TP_i, TN_i, FP_i, FN_i, respectively, represent the number of true positive, true negative, false positive and false negative entries in the *i-th* family. Moreover, the number of traffic families is $C = 9$.

4.3 Experimental Results and Analysis

To evaluate the performance of MEMTD, we compare the proposed approach with other DL detection methods in the comparison experiments section. Then, we analyze several properties of MEMTD, which contribute to improving detection performance:

- **The three input modalities** contain distinguishing information, which is conducive for encrypted malware traffic detection. To demonstrate the validity of the adopted modalities, pretrained ConvNet, GruNet-1, and GruNet-2 are utilized to detect malicious traffic in the ablation experiments section.
- **The intermediate fusion mechanism** can enhance the presentation of intermodal dependence and relieve the difficulty of fusing features from heterogeneous data. For comparison, the variants of the MEMTD approach are utilized to research the effect of different fusion orders in the feature fusion experiments section. They are implemented as MEMTD-T, MEMTD-L, and MEMTD-S as displayed in Table 3.
- **Imbalanced malicious traffic detection** is more relevant to the needs of actual cyberspace security, and the MEMTD approach can accomplish this task. Learning a robust and comprehensive representation of encrypted traffic is key to traffic classification, and the imbalanced distribution of malicious traffic emphasizes this requirement. We discuss further in the imbalance detection analysis section.

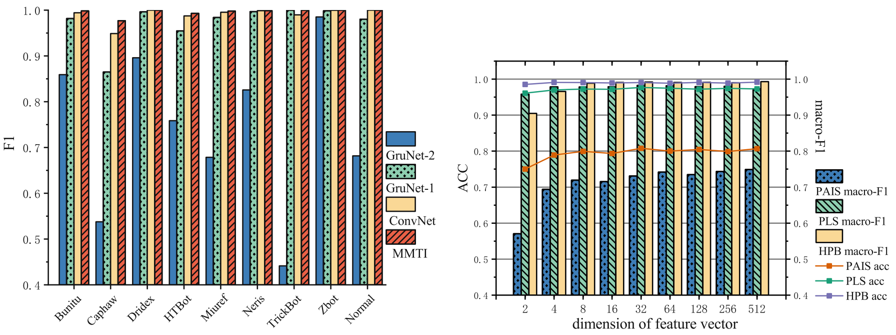
Table 2. Experimental results

ID	Families	CLSTM		FS-Net		**MEMTD**	
		Precision	Recall	Precision	Recall	Precision	Recall
1	Bunitu	0.8932	0.8970	0.9480	**1.0000**	**1.000**	0.9994
2	Caphaw	0.5397	0.4873	0.9842	0.9780	**0.9982**	**0.9973**
3	Dridex	0.7177	0.6696	**1.0000**	0.8666	**1.0000**	**1.0000**
4	HTBot	0.7164	0.7617	**1.0000**	**1.0000**	0.9884	0.9995
5	Miuref	0.9252	0.9432	0.9280	0.7813	**1.0000**	**1.0000**
6	Neris	0.8960	0.8581	0.9826	0.9817	**1.0000**	**1.0000**
7	TrickBot	0.9268	0.9344	0.9832	**1.0000**	**1.0000**	**1.0000**
8	Zbot	0.8061	0.7690	**1.0000**	0.9978	**1.0000**	**1.0000**
9	Normal	0.7858	0.5789	0.9592	0.9825	**1.0000**	**1.0000**
ACC		0.8118		0.9936		**0.9994**	
macro-F1		0.7815		0.9638		**0.9996**	

Comparison Experiments. Table 2 illustrates the performance comparison between the MEMTD approach and other DL methods in terms of encryption malware traffic detection. The conclusions obtained are presented below.

a. The MEMTD approach obtains the best performance compared to other methods and can effectively identify each malware family. According to Table 2, the macro-F1 indicator of MEMTD reached 99.96%, which means that even if a certain malware family has much less traffic than those other families (e.g., Dridex or TrickBot), MEMTD can also identify them with high performance.
b. The introduction of other modal inputs can improve the performance of end-to-end DL models. The MEMTD approach inputs the same PLS as the input of FS-Net and adds the other two modalities, HPBs and PAIS. The MEMTD improved macro-F1 by 3.58% without the reconstruction mechanism of FS-Net. This result demonstrates that the introduction of multimodal learning can break the performance bottleneck of complex neural networks in a single modality.
c. The MEMTD approach can extract comprehensive traffic information without introducing a complex neural network structure. In the different modal feature extraction step, MEMTD does not combine a CNN and a recurrent neural network (RNN) to form a complex feature representation layer. However, the performance of the MEMTD approach is much better than that of the CLSTM method that combines CNN and LSTM.

Ablation Experiments. The classification performance of the MEMTD approach and its three pretrained single-modal deep networks are illustrated in Fig. 3(a). The following conclusions can be drawn.



(a) Experimental results of the single-modal deep neural networks in MEMTD.

(b) Experimental results of networks with different feature vector dimensions.

Fig. 3. Analysis of three modalities and their corresponding feature vector dimensions.

Table 3. Experimental results of different fusion strategies

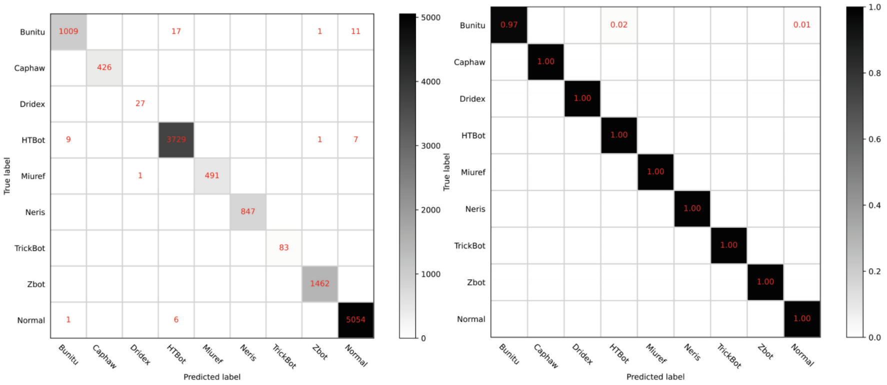
Methods	Fusion-1 inputs	Fusion-2 inputs	macro-F1	ACC
MEMTD	v_{pls} and v_{pais}	v_{fus1} and v_{hpb}	**99.96%**	**99.94%**
MEMTD-T	v_{hpb} and v_{pls}	v_{fus1} and v_{pais}	99.41%	99.91%
MEMTD-L	v_{hpb} and v_{pais}	v_{fus1} and v_{pls}	99.89%	99.87%
MEMTD-S	–	v_{hpb}, v_{pls}, and v_{pais}	99.55%	99.80%

a. The experimental results show that ConvNet, GruNet-1, and GruNet-2 can identify encrypted malware family traffic, demonstrating that HPBs, PLS, and PAIS include distinguished classification information. The most distinguishable information is that on the distribution of packet payload bytes in the SSL/TLS handshake phase. In addition, the traffic generated by different malicious behaviors differs more greatly in terms of packet length than in terms of packet arrival time.

b. The combined usage of modalities complements the feature representation of encrypted traffic. In each malware family detection task, the F1 indicator of MEMTD exceeds the three single-modal classifiers. This result indicates that the classification information extracted by ConvNet, GruNet-1, and GruNet-2 is not lost in the feature fusion step of the MEMTD approach. Moreover, the association and dependence between modalities are learned by fusion layers in the fine-tuning stage to supplement traffic features and improve detection performance.

c. The selection of the packets employed to input CNN affects performance. As shown in Fig. 3(a), ConvNet adopting $x_hpb = (p_ch, p_sh)$ as input has an F1 indicator of more than 0.94 in each family. The macro-F1 indicator of the CLSTM method utilizing three random consecutive packet payloads as input is only 0.7815, although the structure of CLSTM is more complicated than that in ConvNet. Therefore, the distinctive features of encrypted malicious traffic are in the TLS handshake phase, specifically, in client hello and server hello messages, not in other locations.

Feature Fusion Experiments. The dimensions of the feature vectors, v_{hpb}, v_{pls}, and v_{pais}, determine the richness of the corresponding modal information. Their impact on detection is evaluated through experiments, as shown in Fig. 3(b). Table 3 shows the experimental results of the variants of the MEMTD approach, where v_{fus1} represents the output vector of FusionNet-1. The classification performance of these networks with different dimensions of feature vectors and different variants of MEMTD indicates the following conclusions.



(a) Confusion Matrix.

(b) Normalised Confusion Matrix.

Fig. 4. Confusion matrix of MEMTD.

a. As the dimensions of the feature vectors, v_{hpb}, v_{pls}, and v_{pais}, increase in Fig. 3(b), the F1 indicator and ACC rise. Nonetheless, the improvement tends to be gentle with the exponential growth of the dimension of vectors. After the dimensions exceed 32, the three feature vectors no longer significantly enhance detection performance. Considering the balance of performance and training time, the MEMTD approach adopts 32 as the dimension of v_{hpb}, v_{pls}, and v_{pais}.
b. The proposed MEMTD intermediate fusion mechanism enhances detection performance and outperforms all other variants. After fusing the vectors representing the features of the full flow, v_{pls} and v_{pais}, by FusionNet-1, FusionNet-2 combines v_{fus1} and v_{hpb} to classify the flow. This strategy allows the MEMTD approach to fuse feature vectors from different sources effectively.
c. The fusion order of the feature vectors, v_{hpb}, v_{pls}, and v_{pais}, affects the detection performance. In particular, the experimental results in Table 3 elaborate that the input of the last feature fusion layer has a more significant effect. The

modality validity experiments in Fig. 3(a) demonstrate that the discriminative information in HPBs, PLS, and PAIS decreases in order. The macro-F1 indicator of MEMTD-L, which inputs PLS to FusionNet-2, is lower than that of MEMTD. In addition, the performance of MEMTD-T, which applies PAIS as the input of FusionNet-2, is poorer than that of MEMTD-S, which utilizes the simple concatenation of feature vectors. Therefore, this paper proposes to placing the feature vector of the most discriminative modality into the final fusion layer in the multimodal DL method.

Imbalanced Detection Analysis. The confusion matrix of MEMTD performance is displayed in Fig. 4. Figure 4(a) shows the MEMTD classification result for imbalanced malware traffic detection. Interestingly, MEMTD has excellent identification performance against malware families with few samples, Dridex and TrickBot. In contrast, the malware with the third-largest number of samples in the dataset, Bunitu, has the worst identification result, as shown in Fig. 4(b). This experimental results indicate that malicious traffic detection differs from general imbalanced classification tasks. The number of traffic samples of a particular malware does not directly determine its traffic detection ability. Therefore, the MEMTD approach does not utilize cost-sensitive learning methods to alleviate the impact of imbalanced datasets but instead focuses on capturing intra- and intermodal dependencies to enhance the representation of encrypted traffic.

5 Conclusion

This paper proposes a multimodal DL method named MEMTD to detect encrypted malware traffic. The MEMTD approach extracts three modal features from the raw network traffic, where intra- and intermodality dependencies can be captured by learning. In addition, the MEMTD approach employs the translation-invariant 1D-CNN structure to build ConvNet, which extracts feature vectors from packets in a flow containing client hello and server hello messages. For PLS and PAIS, the MEMTD approach adopts a modal feature extraction network, GruNet, constructed by Bi-GRU, to learn feature representation. After pretraining, the three modal inputs can be abstracted into feature vectors containing information within the modalities. To thoroughly learn the dependencies between modalities, the MEMTD approach adopts an intermediate fusion mechanism to fuse feature vectors and classify malware traffic. We demonstrate the effectiveness of the MEMTD approach on an imbalanced dataset that mimics actual network states, the experimental results of which show that the MEMTD approach outperforms other single-modal DL methods and is also effective for imbalanced malware traffic.

Acknowledgment. This work was supported by the National key research and development program of China (Grant No. 2018YFB1800705).

References

1. Aslan, Ö.A., Samet, R.: A comprehensive review on malware detection approaches. IEEE Access **8**, 6249–6271 (2020)
2. Anderson, B., Paul, S., McGrew, D.: Deciphering malware's use of TLS (without decryption). J. Comput. Virol. Hacking Tech. **14**(3), 195–211 (2018). https://doi.org/10.1007/s11416-017-0306-6
3. Shekhawat, A.S., Di Troia, F., Stamp, M.: Feature analysis of encrypted malicious traffic. Expert Syst. Appl. **125**, 130–141 (2019)
4. Niu, W., Zhuo, Z., Zhang, X., Xiaojiang, D., Yang, G., Guizani, M.: A heuristic statistical testing based approach for encrypted network traffic identification. IEEE Trans. Veh. Technol. **68**(4), 3843–3853 (2019)
5. Fang, Y., Xu, Y., Huang, C., Liu, L., Zhang, L.: Against malicious SSL/TLS encryption: identify malicious traffic based on random forest. In: Yang, X.-S., Sherratt, S., Dey, N., Joshi, A. (eds.) Fourth International Congress on Information and Communication Technology. AISC, vol. 1027, pp. 99–115. Springer, Singapore (2020). https://doi.org/10.1007/978-981-32-9343-4_10
6. Liu, C., He, L., Xiong, G., Cao, Z., Li, Z.: FS-Net: a flow sequence network for encrypted traffic classification. In: IEEE INFOCOM 2019-IEEE Conference on Computer Communications, pp. 1171–1179. IEEE (2019)
7. Liu, X., et al.: Attention-based bidirectional GRU networks for efficient https traffic classification. Inf. Sci. **541**, 297–315 (2020)
8. Dong, C., Zhang, C., Zhigang, L., Liu, B., Jiang, B.: CETAnalytics: comprehensive effective traffic information analytics for encrypted traffic classification. Comput. Netw. **176**, 107258 (2020)
9. Zou, Z., Ge, J., Zheng, H., Wu, Y., Han, C., Yao, Z.: Encrypted traffic classification with a convolutional long short-term memory neural network. In: 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), pp. 329–334. IEEE (2018)
10. Huang, H., Deng, H., Chen, J., Han, L., Wang, W.: Automatic multi-task learning system for abnormal network traffic detection. Int. J. Emerging Technol. Learn. **13**(4), 4–20 (2018). https://doi.org/10.3991/ijet.v13i04.8466 https://online-journals.org/index.php/i-jet/article/view/8466
11. Congyuan, X., Shen, J., Xin, D.: A method of few-shot network intrusion detection based on meta-learning framework. IEEE Trans. Inf. Forensics Secur. **15**, 3540–3552 (2020)
12. Tong, X., Tan, X., Chen, L., Yang, J., Zheng, Q.: BFSN: a novel method of encrypted traffic classification based on bidirectional flow sequence network. In: 2020 3rd International Conference on Hot Information-Centric Networking (HotICN), pp. 160–165. IEEE (2020)
13. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process. Mag. **34**(6), 96–108 (2017)
14. Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M., Burgard, W.: Multimodal deep learning for robust RGB-D object recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 681–687. IEEE (2015)
15. Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., Pal, C.: Recurrent neural networks for emotion recognition in video. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 467–474 (2015)

16. Aceto, G., Ciuonzo, D., Montieri, A., Pescapè, A.: MIMETIC: mobile encrypted traffic classification using multimodal deep learning. Comput. Netw. **165**, 106944 (2019)
17. Kaya, A., Keceli, A.S., Catal, C., Yalic, H.Y., Temucin, H., Tekinerdogan, B.: Analysis of transfer learning for deep neural network based plant classification models. Comput. Electron. Agric. **158**, 20–29 (2019)
18. Stratosphere: Stratosphere laboratory datasets (2015). https://www.stratosphereips.org/datasets-overview. Accessed 13 Mar 2020