IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2017.Doi Number*

# Edge Intelligence Based Identification and Classification of Encrypted Traffic of Internet of Things

**Yue Zhao[1], Yarang Yang[2], Bo Tian[1], Jin Yang[3], Tianyi Zhang[4], and Ning Hu[5, 6]**

[1]Science and Technology on Communication Security Laboratory, Chengdu 610041, China
[2]College of Physics and Electrical Engineering, Kashi University, Kashi 844006, China
[3]College of Cyber Security, Sichuan University, Chengdu 610065, China
[4]Graduate School of Advanced Integration Science, Chiba University, Chiba 2638522, Japan
[5]Pengcheng Laboratory Cyberspace Security Research Center, Shenzhen 518000, China
[6]Cyberspace Institute of Advanced Technology Guangzhou University, Guangzhou 510006, China

Corresponding author: Ning Hu (e-mail: huning@gzhu.edu.cn).

This work is supported by the National Natural Science Foundation of China under Grant No. 61976064 and No. 61872254, and the Key Research and Development Project of Sichuan Province of China under Grant No. 2020YFG0292.

**ABSTRACT** A detection model of Internet of Things encrypted traffic based on edge intelligence is proposed in the paper, which can reduce the communication times of distributed Internet of Things gateways in the process of edge intelligence as well as the encrypted traffic detection model establishment time, in order to solve the problems that it is difficult to carry out efficient classification and accurate identification of the encrypted traffic of Internet of Things. In this paper, four new classification and identification methods for encrypted traffic are put forward, namely time-sequence behavior analysis, dynamic behavior analysis, key behavior analysis and two-round filtering analysis. The experimental results show that when the sample size is 1600, the encrypted traffic detection model establishment time is less than 100 seconds, and the accuracy of all the four new traffic classification methods is more than 92% and the recall rates of them are more than 83%.

**INDEX TERMS** Internet of things, edge intelligence, encrypted traffic, identification and classification, IoT gateway.

## I. INTRODUCTION

According to recent reports, the popular communication protocols of Internet of Things (IoT), such as message queuing telemetry transport (MQTT) and advanced message queuing protocol (AMQP), use transport layer security (TLS) protocol to prevent data breaches during transmission [1, 2]. Traffic encryption has become one of the important means to protect the privacy of the IoT. It has even become a mandatory requirement of the law even in financial, transportation and other specific industries.

However, one coin has its two sides. Traffic encryption brings vulnerabilities into IoT. The malwares are concealing in the increasing networked applications by encrypting their traffics. It is reported that over 30% of the malwares utilize encryption protocols [3, 4], which leads to the failure of the traditional malwares detection based on deep packet inspection (DPI) technologies.

The common practice of identifying encryption protocols in IoT is to thoroughly analyze the encapsulation formats and interaction processes of the protocols, find out the characteristics and rules that can be used to distinguish them, and learn the most distinctive characteristics for each application protocol in encrypted traffics [5, 6]. The characteristics of an encrypted traffic are in accordance with the used encryption protocol. It is feasible to model the statistical characteristics of the encrypted traffics by machine learning to identify the corresponding encryption protocols. This has been a research hotspot since content analysis of the traffics is not required at all. In [7], several methods are proposed to identify encrypted traffics based on machine learning and describe their application scenarios.

However, machine learning has several disadvantages in encrypted traffic identification. Firstly, the characteristics of an encryption protocol are usually not unique in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3056216, IEEE Access



whole network environment, and other protocols also have similar characteristics, which significantly reduce the performance of machine leaning in identifying encryption protocols in backbone networks. Secondly, the establishment time of the detection model is greatly depended on the machine learning algorithm. There are over 100 available characteristics [8] and over a dozen machine learning algorithms as well [9]. It requires a lot of efforts to decide the most distinctive characteristics for each encryption protocol in practical use and to design the model with appropriate machine learning algorithm. Finally, it is time consuming to for the machine learning models to extract the characteristics from the encrypted traffics. Due to the complexity of the algorithms, it may be even slower than manual characteristic labeling and thus be less efficient.

Cisco Systems Inc. has published several papers in recent years on how to identify malwares that use TLS protocol without decrypting their traffics, which attract great concern in the IoT industry. [10] analyzes the differences among the TLS traffics, the DNS traffics and the HTTP traffics of legal applications and malwares in millions of data flows. And then the distinctive characteristics are obtained and formed into training datasets, to get a better classifier based on supervised machine learning, whereby identify encrypted malware traffic. [11] analyzes thousands of samples from 18 malware families, as well as tens of thousands of malicious connections from millions of encrypted data flows of enterprise networks. The results show that the malware traffics are distinct to the legal traffics and the malwares usually use older or weaker ciphers. Based on the fact, it is possible to identify encrypted malware traffics in most situations. In [12], it is said that Cisco Systems Inc. adopted deep learning to model and classify malware traffics, and sort them into different malware families in terms of the traffic characteristics, whereby Cisco Systems Inc. claims the classification precision can reach 90.3%.

The studies of Cisco Systems Inc., however, didn't address the deployment of deep learning-based models on resource-constrained IoT devices. Edge intelligence technology integrates the complementary advantages of local computing and high computing by coordinating terminal devices with edge servers, so as to significantly reduce the delay and energy consumption of deep learning model reasoning [13]. Therefore, it is suitable to apply edge intelligence to the IoT. In [14], a master-slave structured edge intelligence model is proposed, where the IoT terminals construct independently a machine learning model and upload its updated parameters to the edge severs. Those edge severs collect the updated parameters of different models from various terminal devices in order to update the overall model, and then distribute the updated parameters collected from the overall model to users. This model improves the accuracy and efficiency of detecting abnormal traffic in the IoT. The edge intelligence-based machine learning models also need to be adapted to the transmission rate limits of the edge severs. A Gossip algorithm for network environments is put forward in [15], which can achieve the convergence of the model parameters across the distributed terminals through point-to-point data exchange. To implement the algorithm in IoT, the IoT devices are divided into clusters. The model parameters within a cluster are updated with traditional Gossip algorithm. The sink nodes are designated for each cluster to share the parameters, in order to update the global model.

The paper studies classification and identification technology of Internet of Things encrypted traffic on the basis of edge intelligence, and builds an applied experimental environment for it. The remainder of this paper is organized as follows. Section II describes the system model and edge intelligence method. Section III presents four novel classification and identification methods for IoTs encrypted traffic applications. Section IV describes the experimental environment, and carries out performance analysis. Section V concludes this paper.

## II. System Model and Edge Intelligence

The edge intelligence-based encrypted traffic detection model is showed in Fig. 1. The sink nodes with high communication and computing capabilities are assigned as IoT gateways to collect required information from the encrypted traffics [16]. IoT gateways provide connectivity and usually include remote control and monitoring applications. Edge intelligence enables a so-called model updating and aggregation process [17], where IoT gateways build local machine learning models respectively, share local model parameters dynamically in real time and build a global model collectively. The updated parameters of the global model are then feed to the IoT gateways respectively.

The distributed IoT gateway uses edge intelligence technology to iterate the model parameters time after time [18]. Suppose the set of $n$ IoT gateways is $\boldsymbol{N}=\{N_1, N_2,\ldots, N_n\}$, and the number of samples corresponding to each IoT gateway is $D_i$. IoT gateway $i$ calculates the updated parameters of the $k$-th round of iteration as $\nabla_i(\theta^k)$, and sends the updated parameters to other IoT gateways. Then the updated model parameters of the global model go back to all IoT gateways. The sum of the gradients uploaded by each IoT node is $\nabla_N^k$, and the global model executes optimization and updates the model parameters according to the aggregate gradients received from all IoT gateways.

Let the optimization algorithm of updating parameters be gradient descent algorithm [19], and it goes as

$$\theta^k = \theta^{k+1} - \alpha\nabla_N^k \quad (1),$$

where $\alpha$ is learning efficiency. It is difficult to obtain $\theta^{k+1}$-$\theta^k$ in the actual distributed network, while the parameter

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3056216, IEEE Access



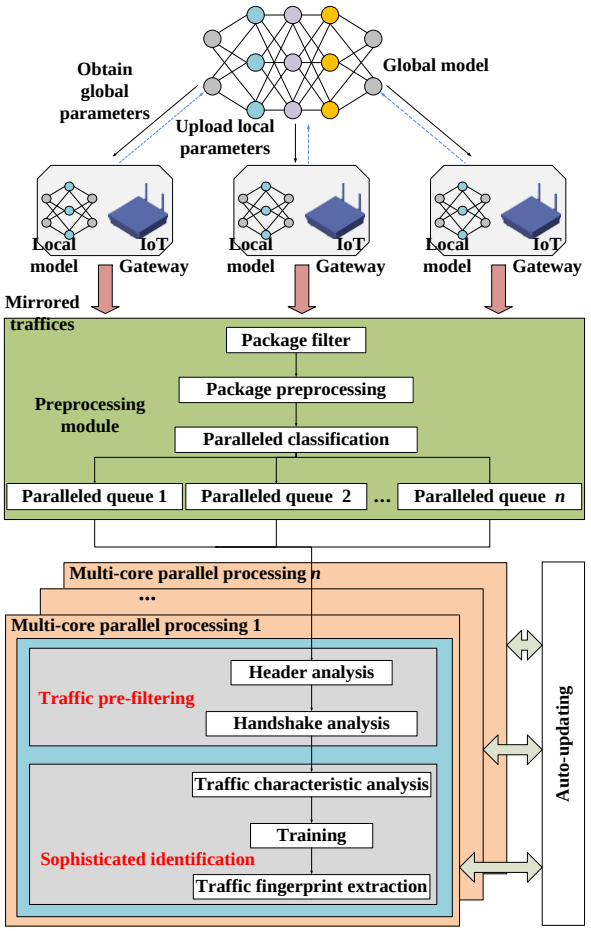


**FIGURE 1. Edge intelligence based encrypted traffic detection model.**

changes tend to be flat in the process of edge intelligence, so $\theta^{k+1}$-$\theta^{k}$ can be approximated as

$$\theta^{k+1}-\theta^{k} \approx \sum_{d=1}^{D} \xi_d (\theta^{k+1-d} - \theta^{k-d}) \quad (2),$$

where both $\xi_d$ and $D$ are constant coefficients. As a matter of convenience, assuming $\xi_d = \frac{1}{D}$ and $D$=1, such that

$$\| \nabla(\theta^{k}) \|^2 \leq \frac{1}{\alpha\beta n^2} \| \sum_{d=1}^{D} \xi_d (\theta^{k+1-d} - \theta^{k-d}) \|^2 \quad (3),$$

where $\beta$ is the scale coefficient, which indicates the proportion of IoT gateways involved currently in the iteration of model parameters. IoT gateways check the gradient according to (3), that is, the nodal IoT gateway performs self-test after a round of learning ends. If (3) is satisfied, the current round of communication is skipped, the gradient is accumulated at local level, and the next round of learning continues to be carried out.

A lower bandwidth often increases duration of each round of the IoT gateway learning so that the overall learning time is prolonged, while the edge intelligence method proposed in the paper has less dependence on network bandwidth. The number of edge intelligence rounds affected by bandwidth reduces as the frequency of communication interaction in the process of gradient descent-based edge intelligence goes down.

With data mirroring, filtering and preprocessing technologies, the IoT gateways are able to process Gigabit-/10-Gigabit-class IoT traffics in real time [20]. The preprocessing module of the model filters the collected traffics utilizing DPI technologies. It removes non-encrypted traffics through an information entropy algorithm for encrypted traffic identification based on statistical testing [21], reducing workloads of analysis and storage in the following processes. Then, based on MapReduce parallel computing architecture [22], the modules classified the encrypted traffics and form them into *n* parallel queues for further processing. The MapReduce architecture is efficient in a clustered environment, in this case IoT. The number of the IoT gateway is linearly positively correlated to the processing performance for the encrypted traffics. With a proper increase in the IoT gateway number and the multi-core parallel processing structure deployed in the following module, the proposed model improves the classification performance for large-scale traffics greatly.

After establishing the communication connection, the IoT devices first carry out key agreements and certificate exchange, which is called as handshake stage, afterwards only then uses the session keys and the encryption suite to implement the encryption transmission of the application data. The pre-filtering module of the model analyzes the plain text in the package headers and the handshake process of MQTT or AMQP protocol. The analyzing results include information indicating the protocol characteristics, such as client version, cypher suites, extensions, etc.

Many studies show that encrypted traffic from a certain IoT application carries behavior characteristics that are inherent to the patterns in its data transmission and package processing [23, 24]. The sophisticated identification module of the model extracts the behavior characteristics of the encrypted traffics without decryption, including time distribution, intervals and transmission rates of the uplink/downlink data packages in the traffics. Then the module identifies encrypted IoT traffics based on a fingerprint database of the traffics and deep recognition algorithms. The fingerprint of an encrypted traffic is a set of essential characteristics. It is unique, distinctive and noise insensitive. With feature fusion machine learning models, multi-dimensional fingerprints can be generated for the encrypted traffics.

The auto-updating module of the model works independently. It decides whether to update the model once an unknown encrypted IoT traffic is identified, and carries out the updating process afterwards.

## III. Methods for Detection and Classification of Encrypted IoT Traffics

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3056216, IEEE Access



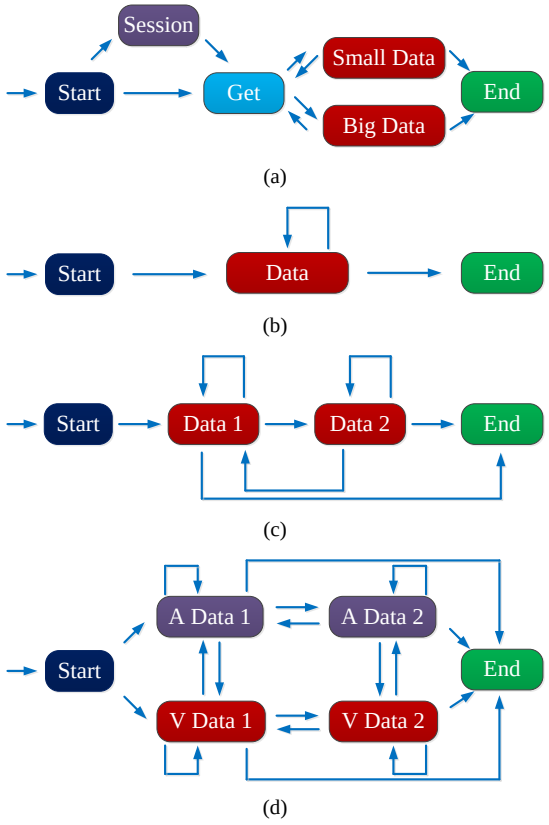


**FIGURE 2. The time-sequence changes of encrypted traffics in four typical applications, i.e. (a) Web application traffics. (b) RTP Traffics. (c) Environment monitoring traffics. (d) Video-audio multimedia**

### *A. Method base on Time-sequence Behavior Analysis*

The method of encrypted IoT traffics identification based on HMM (Hidden Markov Model) extracts the behavior characteristics from a time-sequence change perspective instead of a local one [25]. The behavior characteristics are implied in the process of data packages flowing into and out of the devices, which can be modeled with HMM. We model the timing and statistical characteristics of the traffics with HMM, and analyze the behaviors of the encrypted IoT traffics in the data flows to correlate the traffics to their source applications.

Specifically, the proposed model first labels the packages in the traffics as incoming or outgoing of an application. The observed state of the model is defined by the accumulated length of the continuous incoming or outgoing packages in traffic, in order to reduce the number of the observed states related to a hidden state. The model parameters are obtained with training datasets utilizing forward-backward algorithm. Then the model utilizes Viterbi algorithm in the classifying process to find the most probable state path [26], and relates traffic to its source application.

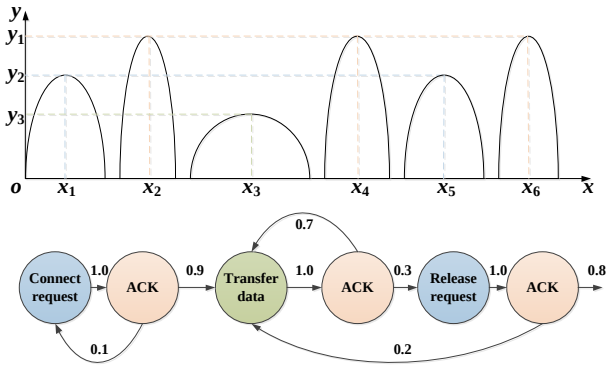


**FIGURE 3. A traffic model of IoT based on HMM.**

Fig. 2 shows the time-sequence changes of encrypted traffics in four typical applications, namely web application traffics, real-time transport protocol (RTP) traffics, environmental monitoring traffics (produced by humidity, temperature or smoke sensors), and video-audio multimedia streaming services traffics (produced by cameras). There are significant differences in the time-sequence change of encrypted traffic for each kind of applications. Therefore, the recurrent neural networks are used to model and analyze the time-sequence change characteristics of encrypted traffic. Fig.3 shows a traffic model of IoT based on HMM. The *x* axis represents the accumulated length of the continuous packages. The *y* axis represents the probability density at each length value, that is, the percentage of the package sets with a length value to overall package sets. The length value is inherent to the modeled protocol *p* and depends on the states. Its maximum appears in the *transfer data* state, while the minimum in the *connect request* or *ACK* state. The relation between the length value and the states implies the statistical characteristics of protocol *p*. In the finite state machine model of protocol *p*, the states are time ordered with given state transition probability. The state transition probability also represents the statistical characteristics of protocol *p*. We take the probability density and the state transition probability comprehensively as the statistical characteristics for identifying traffics of different protocols.

### *B. Method base on Dynamic Behavior Analysis*

Adaptive classification based on weighted ensemble learning analyzes the encrypted traffics dynamically instead of statically. The traffic changes dynamically with the network environment, in the way decided by its protocol and source application. For example, the IoT gateway introduces the latency and package loss into the passing traffics, and it adopts retransmission and flow control strategies to avoid transmission interruption [27]. When the traffics are changed by an IoT gateway, it gives out more information about its protocol and source application.

We proposed an adaptive classification method on weighted ensemble learning. The classifier can be defined

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3056216, IEEE Access



by an expectation function, which is $f$: $x \rightarrow y$. Since $x$ is known, the classification depends on

$$P(y \mid x) = \frac{P(y) \mid P(x \mid y)}{P(x)} \quad (4).$$

Thus we can get

$$f(X) = \arg\max P(y \mid x) = \arg\max_{y \in \mathbf{Y}} \frac{P(y)\prod_{i=0}^{n} P(x_i \mid y)}{P(x)} \quad (5),$$

where $P(x)$ is the probability of characteristic $x$, $P(x) = \prod_{i=0}^{n} P(x_i)$, $P(x_i)$ is the probability of $x_i$ at a training dataset. $y$ represents the categories of the characteristics, consisting constants. $P(y|x)$ varies with $P(x_i|y)$ which depends on $x_i$. $P(y|x)$ also varies with prior probability $P(y)$.

To better describe the changes of the encrypted traffics and get finer classifiers, the proposed method extracts the characteristics of the encrypted traffics at the points where they are changed. The classifiers are trained and selected by their performance before integrated together to guarantee the generalization ability. To train the classifiers, the training dataset is divided into $n$ blocks with identical size, i.e., $S_1$, $S_2$, … , $S_n$. $S_n$ consists of the latest packages. $C_i$ is the classifier trained by $S_i$. $G_k$ is the classifier trained by the last $k$ blocks, i.e., $S_{n-k+1} \cup \cdots \cup S_n$ . $E_k$ is the classifier that integrates the last $k$ classifiers, i.e., $C_{n-k+1}$, … , $C_n$. $W_i$ is the weight of $C_i$, which is inversely proportional to the expectational error for the classifying result for $C_i$.

Here we suppose the category distribution of $S_n$ is the closest with the training dataset, the weight of $C_i$ can be approximated by the classifying error of $C_n$. Specifically, for a training data block $S_n$ consisting of training samples $(x,c)$, the mean square error of $C_i$ is represented as

$$MSE_i = \frac{1}{|S_n|} \sum_{(x,c) \in \mathbf{S_n}} (1 - f_c^i(x))^2 \quad (6),$$

where $c$ is the actual category of characteristic $x$, $f_c^i(x)$ is the probability of correlating $x$ to category $c$, and $1 - f_c^i(x)$ is error rate of $C_i$ for a training sample $(x, c)$. $w_i$ of $C_i$ is inversely proportional to $MSE_i$. The mean square error of the classification error probability $p(c)$ of the random classifier can be represented as

$$MSE_r = \sum_c p(c)(1 - p(c))^2 \quad (7).$$

Since the random classifier does not hold valuable information of the training samples, we use $MSE_r$ to decide whether to integrate it or not. That is to integrate the random classifiers with classification error probability less than $MSE_r$, and discard the others. $w_i$ of $C_i$ being integrated can be calculated by

$$w_i = MSE_r - MSE_i \quad (8).$$

### C. Method based on Key Behavior Analysis

The IoT devices are generally constrained in computing and communication resources. It is necessary to focus the behavior analysis on the most valuable characteristics of the traffics instead of learning from a whole training dataset. To this end, we use the classification accuracy and the iterative rate of the model in the training stage to select valuable characteristics, which significantly reduces the resources required by the proposed model. We use sequential forward selection algorithm to obtain a characteristic set, and select the most valuable characteristics that comprise the fingerprint of the encrypted traffics by a criterion value defined by classification accuracy. The set of the most valuable characteristics obtained by key behavior analysis method is the fingerprint of encrypted traffic.

Suppose there are $n$ characteristics in the global characteristic set, and thus $2^{n-1}$ non null characteristic values. Our goal is to get the most valuable characteristics from the set. An improved sequential forward selection algorithm is proposed. The algorithm adds a characteristic to a valuable characteristic set and calculates the criterion value of the current valuable characteristic set. The valuable characteristic set with the max criterion value is then taken for behavior analysis. The algorithm details are as follows:

The global characteristic set is represented as $\boldsymbol{F}$={$f_1$, $f_2$, …, $f_n$}. The initial valuable characteristic set is $\boldsymbol{F}_0$= Ø, and the valuable characteristic set is represented as $F_k$ which includes $k$ characteristics from $\boldsymbol{F}$. There are $n$-$k$ unselected characteristics in $\boldsymbol{F}$, which are represented as $F_j$ ($j$=1, 2, … , $n$-$k$). The algorithm adds one characteristic of $F_j$ to $F_k$ every time, and calculates the criterion value $J$ of each resulting $F_k$ respectively.

If $J(F_k + x_1) \geq J(F_k + x_2) \geq \cdots \geq J(F_k + x_{n-k})$, $x_1$ is added to $F_k$ and $F_{k+1}$=$F_k$+$f_1$. The algorithm continues until the maximum $J$ is reached to reduce the computation. The time complexity of the algorithm is no more than $n(n$-1)/2. Table 1 gives an instance of the algorithm running process.

### D. Method based on Two-round Filtering Analysis

The common classification methods for the encrypted IoT traffics are based on single machine learning algorithm [28]. We propose a two-round filtering analysis method for encrypted IoT traffics. The proposed method improves the

**TABLE 1. An instance of the improved sequential forward selection algorithm.**

| Number of Iterations | $F_k$ | $J$ | $F_{k+1}$ |
|---|---|---|---|
| 1 | $f_1$ | 30 | $f_3$ |
| | $f_2$ | 20 | |
| | $f_3$ | 35 | |
| | $f_4$ | 25 | |
| 2 | $f_1f_3$ | 40 | $f_2f_3$ |
| | $f_2f_3$ | 50 | |
| | $f_3f_4$ | 45 | |
| 3 | $f_1f_2f_3$ | 40 | Stop($f_2f_3$) |
| | $f_2f_3f_4$ | 45 | |

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3056216, IEEE Access



accuracy and the timeliness of traffic classification at the IoT gateways, while reducing the energy consumption.

The first layer machine learning algorithm aims to a fast filtering of the encrypted IoT traffics. The second layer is to obtain sophisticated classification of the filtered traffics and correlate them to their source applications.

The most common supervised learning algorithms [29] perform differently in encrypted traffics classification. The accuracy of naive Bayes algorithm is low. For the neural network algorithm, the learning speed, the tolerance and interpretation of missing values and irrelevant attributes are relatively low. The classification speed of K-nearest neighbor (KNN) algorithm is relatively low. And the learning speed, the interpretation ability and the model parameter processing ability of support vector machine (SVM) algorithm are relatively low.

Compared with the mentioned algorithms, the decision tree algorithm is superior in learning speed, classification speed, model parameter processing ability, etc. It is efficient in encrypted traffic detection and identification in binary classification scenarios. However, the over-fitting problem of the decision tree algorithm cannot be thoroughly solved so far. Therefore, we combine the decision tree algorithm with the random forest algorithm to provide a two-round filtering analysis method for encrypted IoT traffics. The random forest algorithm performs well in multi classification scenarios, is fast in training and prediction, and is less prone to over fitting. As shown in Fig. 4, the decision tree algorithm is firstly used to obtain the encrypted IoT traffics, and the random forest algorithm is then used to correlate the obtained traffics to their source applications.

## IV. Experimental Results

We use two sets of experimental data to test the proposed model and methods. Dataset 1 is formed by IoT data obtained from public sources. Dataset 2 is formed by the IoT traffics captured by OpenWrt software installed on IoT gateways [30]. The experimental data includes web application traffics, RTP traffics, environmental monitoring traffics, and video-audio multimedia streaming services traffics.

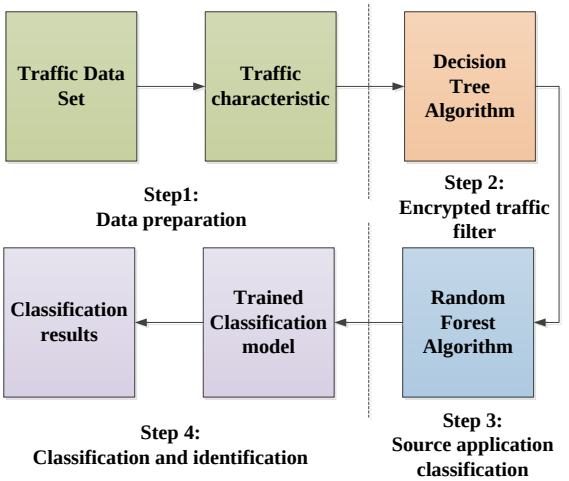


**FIGURE 4. The Two-round filtering analysis for the encrypted IoT traffics.**

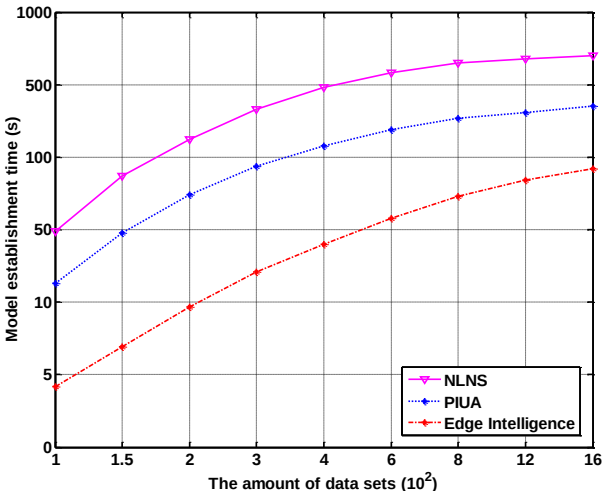


**FIGURE 5. The model establishment time of different algorithms.**

With various data sets of different sizes, the paper compares and analyzes the IoT encryption traffic model establishment time in different algorithms including PIUA in reference [6], NLNS in reference [12] and the edge intelligence method proposed in this paper. Each algorithm is executed 10 times in order to get an average value of the model establishment time, which is shown in Fig. 5. It is found that the model establishment time of edge intelligence based traffic fingerprint extraction is the shortest, which utilizes distributed parallel computing to speed up the processing speed and realize reduction of its execution time, while that of NLNS method, due to the integration of multiple characteristic selection algorithms, is the longest. In addition, the number of iterations of model parameters generated by the edge intelligence method is the smallest so that the classification model is simplified. When the sample size is 1600, the execution time of characteristic selection is less than 100 seconds.

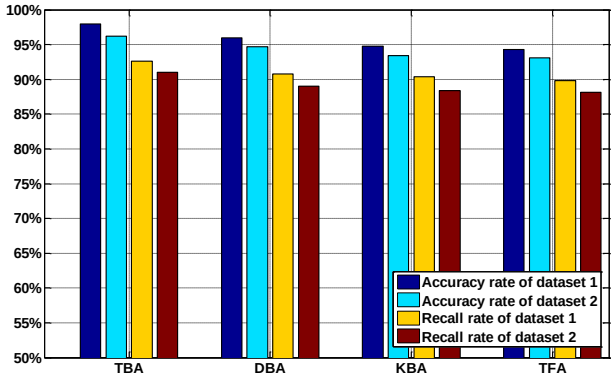


**FIGURE 6. The accuracy rate and the recall rate of the proposed methods.**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3056216, IEEE Access



In the experiment, we use the proposed methods to classify and identify the encrypted IoT traffics in the test datasets. The four methods proposed in this paper are used separately, and their performances are tested and analyzed respectively. The accuracy rate and the recall rate of each method are shown in Fig. 6. The experimental results show that, the method based on time-sequence behavior analysis has the best performance, with accuracy rate up to 98%. The method based on two-round filtering analysis is the worst, but still with accuracy rate up to 92%. The experimental results obtained from dataset 1 are much better than that from dataset 2, because the traffics in dataset 2 are more comprehensive and evenly distributed in categories.

Fig. 7 and Fig.8 show the model establishment time and the traffics classification time of four methods. The method based on dynamic behavior analysis and the method based on two-round filtering analysis are faster in traffic fingerprint extraction, and the method based on time-sequence behavior analysis is the slowest. The latter is slowed by its recurrent neural networks algorithms, which could be improved utilizing parallel computing. As can be seen from the figures, the online classification time of encrypted traffic is much faster than the establishment time of encrypted traffic detection model. We can see that the

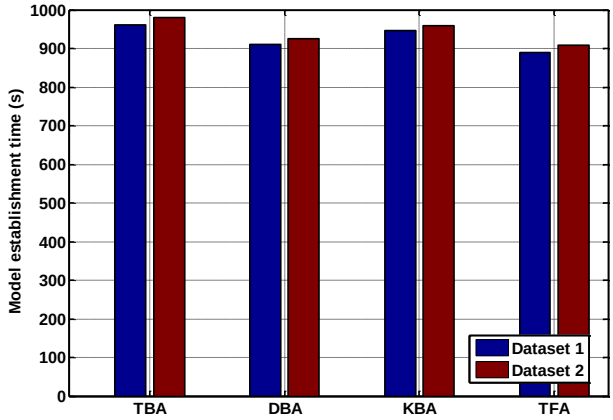


**FIGURE 7. The model establishment time of the proposed methods.**

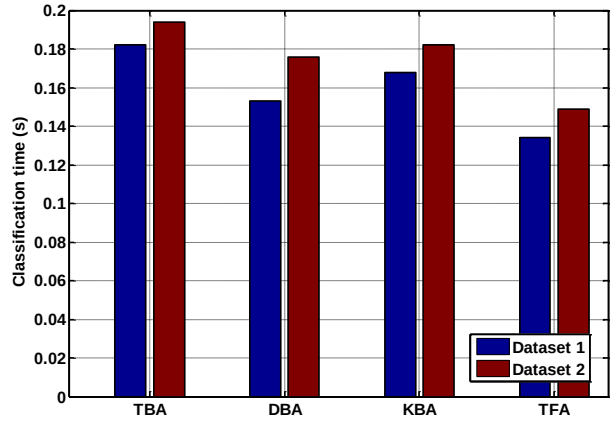


**FIGURE 8. The classification time of the proposed methods.**

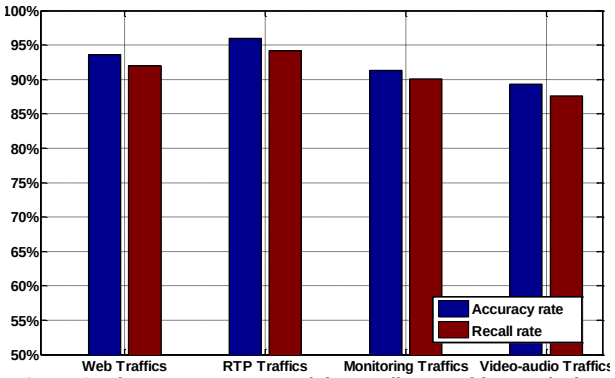


**FIGURE 9. The accuracy rates and the recall rates of four typical applications.**

method based on key behavior analysis is the most balanced choice. It leads in modeling and classifying speed, simplifies the model by reducing the characteristics, and improves both the accuracy rate and the recall rate of classification.

With the method based on key behavior analysis, four typical applications in dataset 2 are used to train and build the detection model. As shown in Fig. 9, the accuracy rates and recall rates of encrypted traffic classification of the four applications are different. It is illustrated that both web application traffics and RTP traffics have better detection effect. Generally, their accuracy rates are above 93.5%, and recall rates are above 92.0%, while the detection efficiency of environment monitoring traffics and video-audio multimedia streaming services traffics is slightly lower. This is because there are differences in the traffic characteristics of the applications themselves, resulting in the degrees of difficulty in extracting their traffic fingerprint are not the same, so the performance of encrypted traffic classification of them are not the same.

## V. CONCLUSION

Regarding to IoT encrypted traffic, including the open data of the IoT traffic acquired from the network and the IoT gateway traffic collected by IoT gateway, the paper proposes four new classification & identification methods, namely time-sequence behavior analysis, dynamic behavior analysis, key behavior analysis and two-round filtering analysis, to study classification and recognition technology, whereby significantly reducing the execution duration of the characteristic location of IoT encrypted traffic with edge intelligence. These newly-proposed traffic identification technologies are able to well classify and identify encrypted web traffic, RTP traffic, environmental monitoring traffic, and VoIP traffic with accuracy more than 92% and recall rate more than 83%. Among them, the dynamic behavior analysis can effectively detect the changes of encrypted traffic in complex networks, and update the adaptive classifier in real time; the key behavior analysis algorithm can produce fewer characteristics, improve the accuracy and stability of classification with a less time than other

methods in establishing model and classifying traffic. These methods proposed in the paper can provide technical support for the detection, analysis and traceability of encrypted traffic in the IoT. It is necessary to make certain the application protocol and the application that generate the encrypted traffics in order to prevent illegal elements from using encryption protocol for information transmission or network attack, and meet the current requirements of the network supervision in an efficient way.

## ACKNOWLEDGMENT

The authors would like to thanks Kaijun Wu, Yao Hao, and Hong Su for their kind help and valuable suggestions. They would also like to thanks the anonymous reviewers for their insightful comments which have significantly improved the quality of the paper. This work is supported by the National Natural Science Foundation of China under grants No.61976064, No.61872254, and the Key Research and Development Project of Sichuan Province of China under grant No. 2020YFG0292.

## REFERENCES

[1] Y. Lee, Y. Kim and J. Kim, “Implementation of TLS and DTLS on Zephyr OS for IoT Devices,” in Proc. International Conference on Information and Communication Technology Convergence, pp. 1292-1294, Jeju, South Korea, Oct. 2018.

[2] N. Nikolov and O. Nakov, “Research of Secure Communication of Esp32 IoT Embedded System to.NET Core Cloud Structure using MQTTS SSL/TLS,” in Proc. IEEE XXVIII International Scientific Conference Electronics, pp. 1-4, Sozopol, Bulgaria, Sept. 2019.

[3] D. Herrald and R. Kovar, “The Hidden Empires of Malware,” [Online]. Available: https://www.sans.org/summit-archives/file/summit-archive-1517253771.pdf.

[4] F. Wang and J. Feng, “Research on Network Traffic Identification Technology for Big Data Platform,” in Proc. International Conference on Communication Software and Networks, pp. 584-588, Chengdu, China, Jul. 2018.

[5] N. Msadek, R. Soua and T. Engel, “IoT Device Fingerprinting: Machine Learning based Encrypted Traffic Analysis,” in Proc. IEEE Wireless Communications and Networking Conference, pp. 1-8, Marrakesh, Morocco, Apr. 2019.

[6] P. Junges, J. François and O. Festor, “Passive Inference of User Actions through IoT Gateway Encrypted Traffic Analysis,” in Proc. IFIP/IEEE Symposium on Integrated Network and Service Management, pp. 7-12, Arlington, USA, Apr. 2019.

[7] W. Pan, G. Cheng, X. Guo and S. Huang, “Review and Perspective on Encrypted Traffic Identification Research,” Journal on Communications, vol. 37, no. 9, pp. 154-167, Sept. 2016.

[8] D. Peng, Y. Qiao and J. Yang, “Analyzing Traffic Characteristics between Backbone Networks Based on Hadoop,” in Proc. IEEE Conference on Cloud Computing and Intelligence Systems, pp. 149-153, Shenzhen, China, Nov. 2014.

[9] M. Duan, K. Li, X. Liao and K. Li, “A Parallel Multi-classification Algorithm for Big Data Using an Extreme Learning Machine,” IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 6, pp. 2337-2351, Jun. 2018.

[10] B. Anderson and D. McGrew, “Identifying Encrypted Malware Traffic with Contextual Flow Data,” in Proc. ACM Workshop on Artificial Intelligence and Security, pp. 35-46, Vienna, Austria, Oct. 2016.

[11] B. Anderson, S. Paul and D. McGrew, “Deciphering Malware's Use of TLS (without Decryption),” Journal of Computer Virology and Hacking Techniques, vol. 14, pp. 195-211, Jul. 2017.

[12] B. Anderson and D. McGrew, “Machine Learning for Encrypted Malware Traffic Classification: Accounting for Noisy Labels and Non-Stationarity,” in Proc. ACM Conference on Knowledge Discovery and Data Mining, pp. 1723–1732, New York, USA, Aug. 2017.

[13] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo and J. Zhang, “Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing,” Proceedings of the IEEE, vol. 107, no. 8, pp. 1738-1762, Aug. 2019.

[14] J. Wang, J. Zhang, W. Bao, X. Zhu, B. Cao and P. Yu. “Not Just Privacy: Improving Performance of Private Deep Learning in Mobile Cloud,” in Proc. ACM Conference on Knowledge Discovery & Data Mining, pp. 2407-2416, London, UK, Aug. 2018.

[15] Y. Li, J. Park, M. Alian, Y. Yuan, Z. Qu, P. Pan, R Wang, A. Schwing, H. Esmaeilzadeh and N. Kim, “A Network-Centric Hardware/Algorithm Co-Design to Accelerate Distributed Training of Deep Neural Networks,” in Proc. IEEE/ACM International Symposium on Microarchitecture, pp. 175-188, Fukuoka, Japan, Oct. 2018.

[16] F. Banaie, J. Misic, V. Misic, M. Moghaddam and S. Seno, “Performance Analysis of Multithreaded IoT Gateway,” IEEE Internet of Things Journal, vol. 6, no. 2, pp. 3143–3155, Apr. 2019.

[17] M. Frej and K. Elleithy, “Secure Data Aggregation Model (SDAM) in Wireless Sensor Networks,” in Proc. IEEE International Conference on Machine Learning and Applications, pp. 330-334, Miami, USA, Dec. 2015.

[18] C. Gong, F. Lin, X. Gong and Y. Lu, “Intelligent Cooperative Edge Computing in Internet of Things,” IEEE Internet of Things Journal, vol. 7, no. 10, pp. 9372-9382, Oct. 2020.

[19] H. Hui, C. Zhou, S. Xu, and F. Lin, “A Novel Secure Data Transmission Scheme in Industrial Internet of Things,” China Communications, vol. 17, no. 1, pp. 73-88, Jan. 2020.

[20] H. Duan, Y. Zheng, C. Wang and X. Yuan, “Treasure Collection on Foggy Islands: Building Secure Network Archives for Internet of Things,” IEEE Internet of Things Journal, vol. 6, no. 2, pp. 2637-2650, Apr. 2019.

[21] J. Yoon, “Using a Deep-Learning Approach for Smart IoT Network Packet Analysis,” in Proc. IEEE European Symposium on Security and Privacy Workshops, pp. 291-299, Stockholm, Sweden, Jun. 2019.

[22] Q. Wang, B. Lee, N. Murray and Y. Qiao, “MR-IoT: An Information Centric MapReduce Framework for IoT,” in Proc. IEEE Annual Consumer Communications & Networking Conference, pp. 1-6, Las Vegas, NV, Jan. 2018.

[23] G. Aceto, D. Ciuonzo, A. Montieri and A. Pescapé, “Mobile Encrypted Traffic Classification Using Deep Learning: Experimental Evaluation, Lessons Learned, and Challenges,” IEEE Transactions on Network and Service Management, vol. 16, no. 2, pp. 445-458, Jun. 2019.

[24] P. Wang, S. Li, F. Ye, Z. Wang and M. Zhang, “PacketCGAN: Exploratory Study of Class Imbalance for Encrypted Traffic Classification Using CGAN,” in Proc. IEEE International Conference on Communications, pp. 1-7, Dublin, Ireland, Jun. 2020.

[25] T. Chadza, K. Kyriakopoulos and S. Lambotharan, “Contemporary Sequential Network Attacks Prediction using Hidden Markov Model,” in Proc. IEEE International Conference on Privacy, Security and Trust, pp. 1-3, Fredericton, Canada, Aug. 2019.

[26] Y. Zhao, Y. Yang, B. Tian, and T. Zhang, “An Invocation Chain Test and Evaluation Method for Fog Computing,” Wireless Communications and Mobile Computing, vol. 2020, Article ID. 8812017, pp. 1-11, Aug. 2020.

[27] Y. Zhao, B. Tian, Z. Chen, J. Yang and S. Li, “A Relay-assisted Secure Handover Mechanism for High-speed Trains,” KSII Transactions on Internet and Information Systems, vol. 13, no.2, pp. 582-596, Jan. 2018.

[28] J. Yang, T. Li, G. Liang, W. He and Y. Zhao, “A Simple Recurrent Unit Model Based Intrusion Detection System with DCGAN,” IEEE Access, vol. 7, pp. 83286-83296, Jun. 2019.

[29] R. Saravanan and P. Sujatha, “A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3056216, IEEE Access



Approaches in Data Classification," in Proc. International Conference on Intelligent Computing and Control Systems, pp. 945-949, Madurai, India, Jun. 2018.

[30] GitHub, "OpenWrt," [Online]. Available: https://github.com/openwrt/openwrt.

**YUE ZHAO** received the BS degree in communication engineering in 2006 from the North China Institute of Science and Technology, Langfang, China, and the PhD degree in information and communication systems in 2012 from Southwest Jiaotong University, Chengdu, China. From September 2010 to September 2011, he was a visiting student in the Department of Electrical and Computer Engineering, University of Florida. He is currently a senior engineer of science and technology on communication security laboratory, Chengdu, China. His research interests include wireless network and information security.

**YARANG YANG**, male, born in 1972, graduated from Central China Normal University as an master, works as an associate professor in College of Physics and Electrical Engineering, Kashi university, has engaged in wireless network and information security research for decades of years, published more than 10 papers.

**BO TIAN** was born in 1970. He received a M.S. degree in communication engineering from University of Electronic Science and Technology of China (UESTC), in 1997. Currently, he is a professor status high level engineer of Science and Technology on Communication Security Laboratory, Chengdu, China. His main research interests include cyberspace security and communication information system.

**JIN YANG** received the M.S. and Ph.D. degrees in computer science from Sichuan University, Sichuan, China, in 2004 and 2007 respectively. He is currently an Associate Researcher with the College of CyberSecurity, Sichuan University, China. His main research interests include network security, knowledge discovery, and expert systems.

**TIANYI ZHANG** received the B.E. degree in Electrical Engineering from North China Institute of Science and Technology in 2006, received the M.E. degree in Electrical Engineering from Tohoku Gakuin University in 2013. He is currently working in the Graduate School of Advanced Integration Science, Chiba University. His current research interests include error control coding, communication system and distributed storage system. He is a member of the IEEE and IEICE.

**NING HU** is currently a professor with the Cyberspace Institute of Advanced Technology, Guangzhou University, Guangdong Province, China. He is also a part-time researcher of Pengcheng Laboratory of Shenzhen. He received his B.S., M.S. and Ph.D degrees in computer science college of National University of Defense Technology (NUDT), China in 1994, 1997 and 2010, respectively. He has authored over 50 journal and conference papers in these areas. His current research interests include software defined network, Industrial IoT and network security. His research has been supported by the National Natural Science Foundation of China. He is a Senior Member of the China Computer Federation. E-mail: huning@gzhu.edu.cn