applied sciences

MDPI

Article

A VPN-Encrypted Traffic Identification Method Based on Ensemble Learning

Jie Cao [1,2], Xing-Liang Yuan [1], Ying Cui [3], Jia-Cheng Fan [1] and Chin-Ling Chen [4,5,*]

[1] School of Computer Science, Northeast Electric Power University, Jilin 132012, China; caojie@neepu.edu.cn (J.C.); 2201990048@neepu.edu.cn (X.-L.Y.); 2202100995@neepu.edu.cn (J.-C.F.)
[2] School of Information Engineering, Guangdong ATV College of Performing Arts, Zhaoqing 526631, China
[3] Zhuhai Power Supply Bureau of Guangdong Power Grid Co., Ltd., Zhuhai 519000, China; cuiying794758706@126.com
[4] School of Information Engineering, Changchun Sci-Tech University, Changchun 130600, China
[5] Department of Computer Science and Information Engineering, Chaoyang University of Technology, Taichung 41349, Taiwan
* Correspondence: clc@mail.cyut.edu.tw

Abstract: One of the foundational and key means of optimizing network service in the field of network security is traffic identification. Various data transmission encryption technologies have been widely employed in recent years. Wrongdoers usually bypass the defense of network security facilities through VPN to carry out network intrusion and malicious attacks. The existing encrypted traffic identification system faces a severe problem as a result of this phenomenon. Previous encrypted traffic identification methods suffer from feature redundancy, data class imbalance, and low identification rate. To address these three problems, this paper proposes a VPN-encrypted traffic identification method based on ensemble learning. Firstly, aiming at the problem of feature redundancy in VPN-encrypted traffic features, a method of selecting encrypted traffic features based on mRMR is proposed; secondly, aiming at the problem of data class imbalance, improving the Xgboost identification model by using the focal loss function for the data class imbalance problem; Finally, in order to improve the identification rate of VPN-encrypted traffic identification methods, an ensemble learning model parameter optimization method based on optimal Bayesian is proposed. Experiments revealed that our proposed VPN-encrypted traffic identification method produced more desirable VPN-encrypted traffic identification outcomes. Meanwhile, using two encrypted traffic datasets, eight common identification algorithms are compared, and the method appears to be more accurate in identifying encrypted traffic.

Keywords: VPN-encrypted traffic identification; ensemble learning; Xgbooost; feature selection; Bayesian optimization

Citation: Cao, J.; Yuan, X.-L.; Cui, Y.; Fan, J.-C.; Chen, C.-L. A VPN-Encrypted Traffic Identification Method Based on Ensemble Learning. *Appl. Sci.* **2022**, *12*, 6434. https://doi.org/10.3390/app12136434

Academic Editor: Luis Javier Garcia Villalba

Received: 2 June 2022
Accepted: 22 June 2022
Published: 24 June 2022

1. Introduction

1.1. Background

For a long time, network security has been a concern [1,2]. Network security plays a key role in maintaining the security, economic development, and social stability of each country [3–5]. In the field of network security, traffic identification is one of the basic and key technologies for optimizing network services [6,7]. It divides traffic into multiple priorities or multiple service classes, which represents the first step in detecting abnormal network activity. In recent years, various data transmission encryption technologies are widely used. Criminals often need to transmit specific data packets during network intrusions and malicious attacks. Anomalous traffic is often identified and intercepted by firewalls and intrusion detection systems [8], and a virtual private network (VPN) is a technique for bypassing these network security defenses [9]. They use the encryption features of VPN to evade the detection of network security facilities [10]. Encrypted traffic identification poses

a great challenge to the current traffic identification technology, and makes it an important part of traffic identification [11,12]. Therefore, the identification of VPN-encrypted traffic is of great significance for detecting malicious network behaviors and maintaining network security. In the past research on traffic identification technology, the research on encrypted traffic identification has made some progress [13–15].

According to different identification depths, encrypted traffic can be divided into two types. The first type is to divide the known encrypted traffic into various application types, such as chat applications, video streaming applications, web applications, etc. The second type is to classify the encrypted traffic into a specific network application and select the appropriate identification method according to the depth of identification. Current traffic identification methods generally include three classes: first, traffic identification methods based on port numbers; second, traffic identification methods based on deep packet inspection (DPI); and third, traffic identification methods based on machine learning. However, the first two methods require byte-level matching of traffic data for encrypted traffic identification. Because the encrypted traffic is encapsulated by the protocol, it cannot be identified at the byte level. At the same time, machine learning algorithms are widely used in the field of encrypted traffic identification due to their intelligence and efficiency.

At present, machine learning methods have been widely used in many fields, and there are many applications for encrypted traffic identification in network security. Machine learning can solve the difficulty of identifying encrypted traffic by using classifiers [16]. However, traditional machine learning may not achieve satisfactory performance when dealing with unbalanced or noisy complex data [17,18]. The reason is that it is difficult to capture the multiple characteristics and infrastructure of the data. Among machine learning, Ensemble learning can integrate data fusion, data modeling, and data mining into a unified framework. Ensemble learning models can handle data with high-dimensional features and solve the feature redundancy problem. Ensemble learning also supports the use of loss functions, which are robust to anomalous data [19]. Therefore, this paper intends to use an ensemble learning framework to identify VPN-encrypted traffic and solve the previous problem of difficult encrypted traffic identification. It has positive significance for detecting malicious network behaviors and maintaining network security.

In response to the above problems, this paper proposes a VPN-encrypted traffic identification model based on an ensemble learning model, which mainly includes:

(1) Regarding the problem of redundancy of VPN-encrypted traffic data features, this paper proposes a method for selecting encrypted traffic features based on mRMR. The correlation coefficient between the Time-Related VPN-encrypted traffic features is calculated, and the importance ranking is outputted. Then, redundant features are eliminated, and to obtain the optimal set of Time-Related VPN-encrypted traffic features.

(2) For the imbalance of VPN-encrypted traffic data classes, this paper proposes Xgboost-encrypted traffic identification model based on Focal Loss. The Focal Loss function is designed on based on the cross-entropy loss function to improve the Xgboost model. This can change the calculation weight of VPN traffic samples, and realize the processing of data imbalance between VPN-encrypted traffic data classes.

(3) For the problem of the low identification rate of the Xgboost model for encrypted traffic identification, a Bayesian-based parameter optimization method for VPN-encrypted traffic identification model is proposed. The method enables to solve the problem of the low identification rate of VPN-encrypted traffic identification model and achieve the accuracy optimization of this model.

The rest of this article is structured as follows. Section 2 introduces the theoretical basis of the realization of the VPN-encrypted traffic identification model; Section 3 performs mathematical modeling of the proposed research method and analyzes the experimental performance of the proposed method; Section 4 gives the conclusion of this article.

1.2. Related Works

Encrypted traffic identification plays a vital role in combating illegal information theft and hacker attacks, and protecting sensitive data. Generally, traffic identification methods can be divided into the following three classes.

(1) The research of traffic identification based on port number. literature [20,21] screened different port flows by checking the port number of the data packet, and listed them, to know which protocols are used in the flow. Port identification technology can only identify TCP-type and UDP-type packets. When service traffic is transmitted using dynamic ports or unknown ports, some packets such as ICMP messages and other traffic that do not have port numbers cannot be identified. Traffic identification methods based on port numbers are difficult to identify all traffic with a certain method. Each method has its own advantages and disadvantages and is suitable for specific protocols or applications. At the same time, the method has a low identification rate for encrypted traffic, so it is less practical. Port identification technology uses the port number of IP traffic to complete the identification process, and the traffic needs to be TCP-type and UDP-type packets. The method is not very applicable to the problem of identifying VPN-encrypted traffic.

(2) The DPI-based traffic identification method. The literature [22,23] adopts DPI technology. The method requires packet-by-packet unpacking and matching and comparing with the backend database, and is more accurate in identifying specific application types in the traffic. However, the identification system based on DPI technology needs to keep up with the generation of new protocols and new applications and continuously upgrade the background application database, otherwise, it will not be able to identify effectively. The feature sequence in the data packet is manually extracted and the feature library is formed. With the increase in the type and quantity of traffic, the maintenance cost of the feature library continues to increase. At the same time, if the data packet is unknown or encrypted for transmission, it is difficult to identify its specific application using DPI identification technology.

(3) The study of machine learning methods for traffic identification is divided into supervised identification and unsupervised identification.

Literature [24–26] uses unsupervised identification methods. The goal of unsupervised learning is to reveal the inherent characteristics and laws of data through the learning of unlabeled samples. There is no clear identification purpose when identifying encrypted traffic, and known data information cannot be fully utilized. At the same time, it is difficult to quantify identification. Supervised learning is used to identify VPN-encrypted traffic, it can make full use of data information for learning, and it can evaluate the identification results.

In the aspect of supervised identification: the literature [27–30] discusses the identification methods of different supervised learning algorithms for encrypted traffic, including SVM, C4.5 decision tree, and other methods. It is necessary to perform in-depth exploratory data analysis on the dataset, and then conduct a simple dimensionality reduction process. Finally, the optimal set of features suitable for the corresponding algorithm is selected. The performance of this method depends on the artificially designed features and the private information in the traffic. When the feature selection has insufficient consideration, it will affect the accuracy of the identification.

Supervised identification can identify encrypted traffic, but various classifiers have different advantages and disadvantages. For example, they have high deviations or too large variances, which leads to weak robustness. Multiple supervised learning algorithms can be an ensemble and balanced for VPN encryption. The use of ensemble learning for the research of encrypted traffic identification, compared with the traditional machine learning supervised identification method, adds a lot of nonlinear transformations and does not need to do complex feature engineering and feature transformation, which can greatly improve the VPN-encrypted traffic identification results.

In the supervised learning methods, the ensemble learning algorithms have more linear transformations. It combines multiple weak classifiers to build a strong classifier

for ensemble voting and can obtain a higher identification rate. The ensemble learning method is to first generate multiple learners through certain rules, then use a certain integration strategy to combine, and finally comprehensively judge and output the final result. Literature [31–33] all use ensemble learning methods. The Xgboost used in this article is an ensemble learning framework. The method is an excellent implementation of boosting tree, using the CART regression tree as the basic learner. The tree model complexity is added as a regular term to the optimization objective, and a good balance is found between the loss function and the regular term. Xgboost is a type of serially generated ensemble learning, its next learner is related to the previous learner. Finally, after obtaining many learners, the sum is obtained to obtain the final learning result.

Xgboost framework sorts the data in advance and stores it in block form. It is easy to parallel computing and optimize the architecture, thus it is suitable for the identification of encrypted traffic. However, the traditional Xgboost framework still has some problems in identifying encrypted traffic. There is a problem with the redundancy of encrypted traffic data features due to the high correlation between the original data features. The data classes of encrypted traffic are unbalanced, and the accuracy of the Xgboost model for VPN-encrypted traffic identification needs further improvement. Therefore, this paper proposes to use a VPN-encrypted traffic identification method with an improved Xgboost model.

2. Methodology

This article firstly performs a series of preprocessing on the VPN-encrypted traffic data to obtain a new encrypted traffic data collection. Next, the features of the VPN-encrypted traffic dataset can represent each other, i.e., the correlation between features and features is high. This results in redundancy in encrypted traffic data features. The first-order incremental search method is used to analyze the correlation between Time-Related VPN-encrypted traffic features, and calculate the correlation coefficient between features and their importance ranking. Then, mRMR feature selection is performed on VPN-encrypted traffic to eliminate irrelevant or redundant features and obtain the optimal Time-Related VPN-encrypted traffic feature subset. Then, it is necessary to focus on unbalanced VPN traffic samples that are difficult to classify. The Focal Loss function is designed based on the cross-entropy loss function to improve the Xgboost model, and the calculation weight of VPN traffic samples is changed. It is necessary to obtain the VPN-encrypted traffic identification model based on FL_XGB, which realizes the processing of data imbalance between data classes. Eventually, the objective function of a given optimization is combined with Bayesian optimization to implement a global parameter search strategy. the posterior distribution of the objective function is updated by continuously adding sample points to find the optimal value of the objective function of the combined parameters. The optimized global optimal parameter combinations are imported into the FL-XGB model for training. It can solve the problem of the low identification rate of the VPN-encrypted traffic identification model, and finally output the identification results.

The general framework of the FL_XGB-VPN-encrypted traffic identification model is shown in Figure 1.

2.1. Time-Related mRMR Feature Selection

The mRMR-based feature selection method for VPN-encrypted traffic belongs to the class of filtered feature selection methods. The method uses many trade-offs between relevance and redundancy. Additionally, it uses mutual information as the calculation criterion to measure the redundancy between features and the relevance between features and classes. Thus, the relevance between features and classes is maximized and the redundancy between features and features is minimized and achieves feature selection. The maximum correlation means that the features with the highest correlation to the model are selected, and the higher the correlation is, the better the encrypted traffic can be identified. At the same time, the method can simplify the model and shorten the training time of the

model. The mRMR feature selection method uses mutual information as a measure of the correlation between features and features as well as between features and classes. In the obtained feature set, there are large differences between features and correlations with the target variables. Using the mRMR algorithm for feature selection, a streamlined subset of features with good identification can be obtained, and the optimal number of features can also be determined. The mRMR feature selection method has the framework shown in Figure 2.

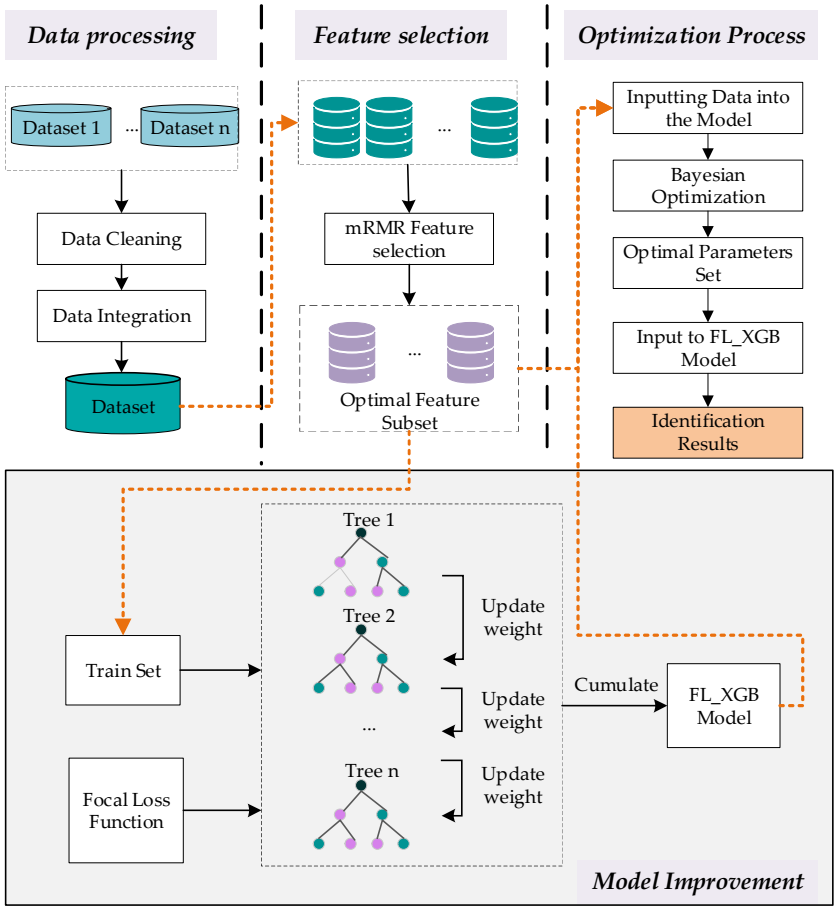


Figure 1. FL_XGB-VPN-encrypted traffic identification model general framework.

The mRMR algorithm is a filtered feature selection method. The method weighs relevance and redundancy differently. Additionally, uses mutual information as a computational criterion to measure the redundancy between features and the relevance between features and class variables. Performs feature selection by maximizing the relevance of features to class variables and minimizing the redundancy between features. The maximum relevance principle refers to the selection of features that have the greatest relevance to the mode. The greater the relevance, the greater the problem-solving capability of the trained model. The specific steps of the mRMR algorithm for feature selection of VPN-encrypted traffic are as follows.

First, the feature set S is initialized to be empty, and the input training dataset contains the feature set F and the class set C. For data containing N-dimensional features, the sorted feature list is obtained by a cyclic process of N iterations. In each iteration, one feature is selected. This feature has the maximum relevance to the target variable compared to other features while having the minimum redundancy with other features that have been selected.

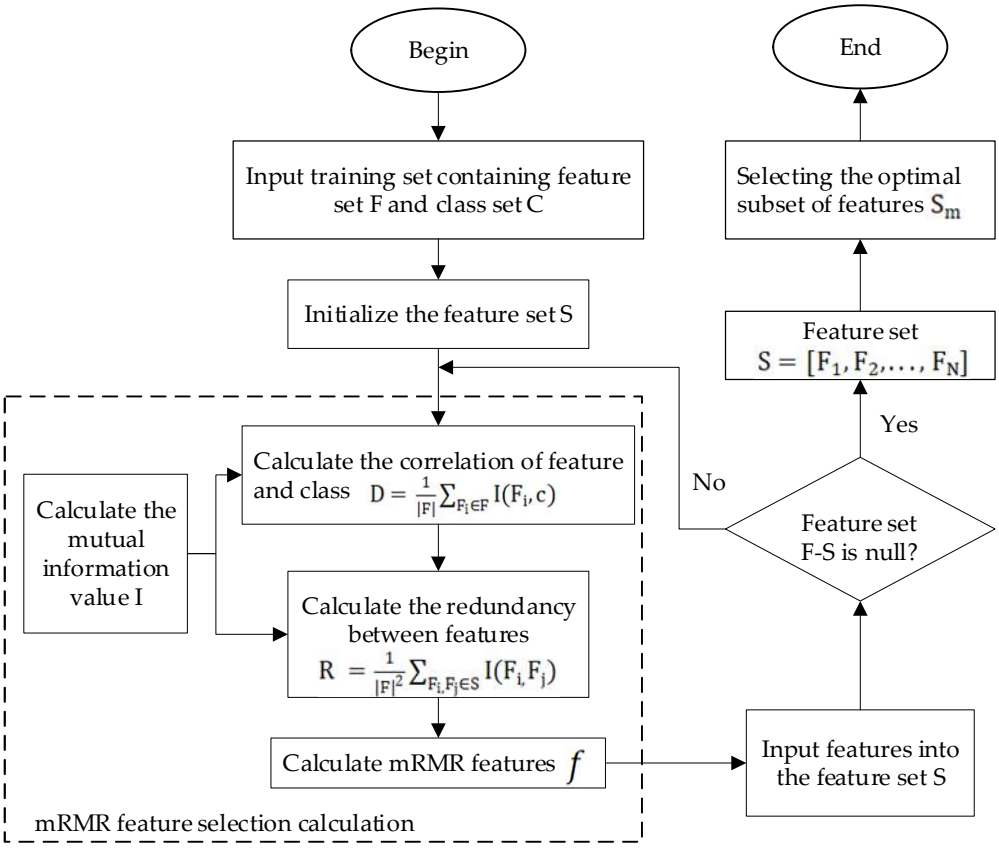


Figure 2. Framework of the mRMR feature selection method.

Mutual information [34] is a measure of information in information theory. It reflects the amount of information contained in a random variable with respect to other variables Y. Assuming that the marginal probability density function of variable X is $P(x)$, the marginal probability density function of variable Y is $P(y)$, and their joint probability density function is calculated to be $P(x, y)$, the value of mutual information is calculated as shown in Equation (1).

$$I(x, y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) log \frac{P(x, y)}{P(x)P(y)} \quad (1)$$

Assume that F is the set containing N-dimensional features and S is the set containing the selected features. For each feature F_i in the feature set to be selected, the correlation between the feature set F composed of all features and the sample class set C is calculated. The correlation between the feature F_i to be selected and the class set C is $I(F_i, c)$. D indicates the relevance of the feature to be selected to the class, and the larger the D the higher the relevance of the feature to the class. The correlation D is calculated as shown in Equation (2).

$$D = \frac{1}{|F|} \sum_{F_i \in F} I(F_i, c) \quad (2)$$

To calculate the redundancy among the features to be selected, the redundancy of all the features in the set F is the average of the mutual information values between feature F_i and feature F_j, and the redundancy is calculated as shown in Equation (3).

$$R = \frac{1}{|F|^2} \sum_{F_i, F_j \in S} I(F_i, F_j). \quad (3)$$

In Equation (3), $I(F_i, F_j)$ indicates the mutual information value of the ith feature and the jth feature. the larger R indicates the higher redundancy between the two features. Find the formula of the criterion that fits the maximum correlation and minimum redundancy between features and types in F. The features f in the feature set F that meet the criteria of maximum correlation between features and classes and minimum redundancy between features are selected and stored in the feature set S. D denotes the maximum correlation between features and classes, and R denotes the minimum redundancy between features.

$$f = max\{D - R\} \tag{4}$$

The algorithm is judged whether the stopping condition is satisfied, that is, whether the feature set F-S is empty or whether the set S is the same as the set F contained elements. If so, the loop is skipped, otherwise, the relevance and redundancy calculation steps are repeated. At the end of the loop, the set $S = [F_1, F_2, \ldots, F_N]$ is obtained in descending order of feature importance. finally, the features in the set $S = [F_1, F_2, \ldots, F_N]$ are selected using the feature-by-feature selection method to obtain the optimal feature subset S_m.

The feature selection method using mRMR reduces the redundancy between features and features while ensuring the maximum correlation between features and classes. The method can solve the problem of feature redundancy existing in the original data.

2.2. VPN-Encrypted Traffic FL-XGB Identification Model

In the problem of VPN-encrypted traffic identification, various types of training data are not evenly distributed. Among them, classes with more data account for a larger proportion of the dataset, and classes with fewer data have a smaller proportion in the dataset. The reason is that the usage frequency of different applications is different, and the amount of traffic data generated is different. Thus, the amount of data between different classes is quite different, and there is a problem of imbalance in the amount of data between data classes. Therefore, this article proposes a custom-defined Focal Loss Xgboost ensemble learning model for the imbalance of encrypted traffic classes.

2.2.1. Overview of the Basic Xgboost Model

Xgboost model has high operational efficiency and prediction accuracy compared with other methods in the field of machine learning and data mining. Xgboost is an improved gradient boosting learning (GDBT) framework, which is a boosting method. Xgboost adds regularizations such as leaf node weights and depth of the tree to the cost function. The regularization term controls the complexity of the model and prevents overfitting. Xgboost has the advantages of being fast, effective, capable of handling large-scale data, and supporting custom loss functions. Classification and regression tree (CART) is a widely used decision tree learning method [35]. Xgboost model consists of CART trees, and its parameters are present in each CART. The determination of the tree structure is to select the optimal split node with the information gain rate as the split criterion. The greedy algorithm is used to enumerate all nodes, the information gain of each node before and after the split is calculated, and the node with the greatest information gain is selected. Essentially it is to do two loops. The first loop is implemented for each feature's split point, and then the gain is calculated, whereas the best split point for the feature is selected. Among them, the split gain uses the difference in the change of the objective function after the split. The second loop object is to select the feature with the highest gain for all features.

The idea of using iterative operations to set some weak learners into a strong learner to achieve accurate identification. The results of the Xgboost model after t iterations is shown in Equation (5).

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{5}$$

In Equation (5), $\hat{y}_i^{(t)}$ is the identification result of sample i after the tth iteration, $\hat{y}_i^{(t-1)}$ is the identification result of sample i after the $(t-1)$th iteration, and $f_t(x_i)$ is the model

identification result of the tth tree. At this point, the objective function $\lambda^{(t)}$ of Xgboost is shown in Equation (6).

$$\lambda^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{6}$$

In Equation (6), y_i is the true value of the ith sample, $\Omega(f_t)$ is the regularization term of the function, and $\sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right)$ is the loss function of the model. $\hat{y}_i^{(t-1)}$ is known, the model has to learn only the tth tree f_t, and then the error function is subjected to a second-order Taylor expansion at $\hat{y}_i^{(t-1)}$ as shown in Equation (7).

$$\lambda^{(t)} \cong \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i {f_t}^2(x_i) \right] + \Omega(f_t). \tag{7}$$

In Equation (7), $g_i = \partial_{y^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ is the first derivative of the loss function, and $h_i = \partial^2_{y^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ is the second derivative of the loss function. Then, after removing the constant term from the formula, the result is obtained as shown in Equation (8).

$$\widetilde{\lambda}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i {f_t}^2(x_i) \right] + \Omega(f_t) \tag{8}$$

Xgboost ensemble learning model, when dealing with the problem of data imbalance, usually requires the use of operations such as active acquisition or sampling to balance the data. However, the operation changes the distribution of the original dataset. In this paper, we take advantage of Xgboost's support for custom loss functions and modify the loss function based on the cross-entropy loss function. On this basis, we solve the problem of VPN-encrypted traffic data imbalance without changing the distribution of the original dataset.

2.2.2. FL-XGB-VPN-Encrypted Traffic Identification Model

The basic Xgboost ensemble learning model suffers from data imbalance when identifying VPN-encrypted traffic data. Various types of training data are not evenly distributed. Among them, classes with more data will take up a larger proportion of the dataset, whereas classes with less data will have a smaller proportion of the dataset. To solve the problem, the Xgboost model is improved with the Focal Loss function based on the cross-entropy loss function. The method can solve the problem of imbalance in the number of difficult and easy samples identified during the training process. Focal Loss is proposed in target detection to solve the problem of imbalance of difficult and easy samples. A parametric balancing factor α is added in front of the cross-entropy loss function to balance the unequal proportions of difficult and easy samples. The parameter γ adjusts the rate at which the weight of simple samples decreases. When γ is 0, it is the cross-entropy loss function. When γ increases, the influence of the factor also increases. The Focal Loss function is shown in Figure 3.

The framework of the FL-XGB-VPN-encrypted traffic identification model is shown in Figure 4.

The basic principle of the FL-XGB-VPN-encrypted traffic identification algorithm is as follows.

First, the training set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$ start training the model initializes t trees, and when t = 1, initialize the sample class label predicted value y_i. XGBoost is a boosted tree model, where an initial tree predicts a value, then gets the deviation of that value from the actual value, and then adds a tree to learn the deviation. The objective function is a combination of the loss function and penalty function. One part is to calculate the difference between the predicted and true values, and the other part is

the regularization term $\Omega(f_t)$. The objective function is $Obj = FL + \Omega(f_t)$, and the Focal Loss function with weight α is shown in Equation (9).

$$FL = -(\alpha y_i log p) + (1 - y_i) log(1 - p) \tag{9}$$

In Equation (9), γ is the focus parameter for focusing on hard-to-identify samples. i is the number of training samples and p is the prediction probability of the label. When $\gamma = 2$ and p is close to 1, the value of $(1 - p)^2$ will be close to 0, and the more accurate the prediction is. When p_t is close to 0, $(1 - p)^2$ will be close to 1, and the more inaccurate the prediction is. In summary, easily identifiable samples will have smaller weights and hard-to-identify samples will have larger weights.

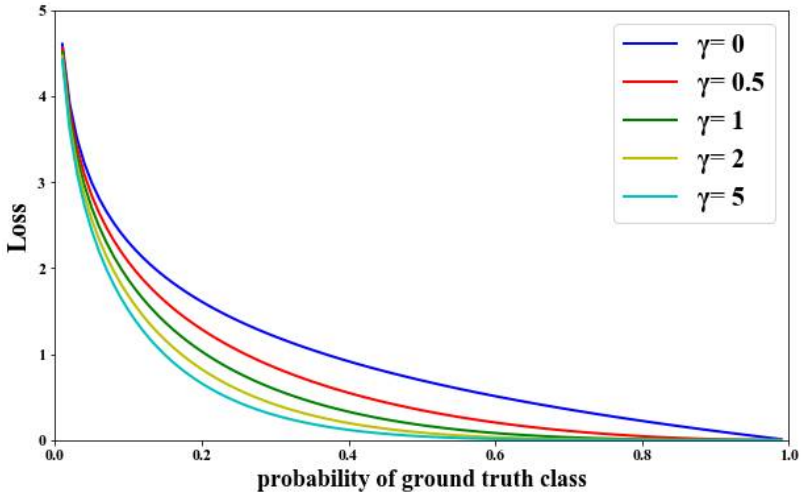


Figure 3. Focal Loss function diagram.

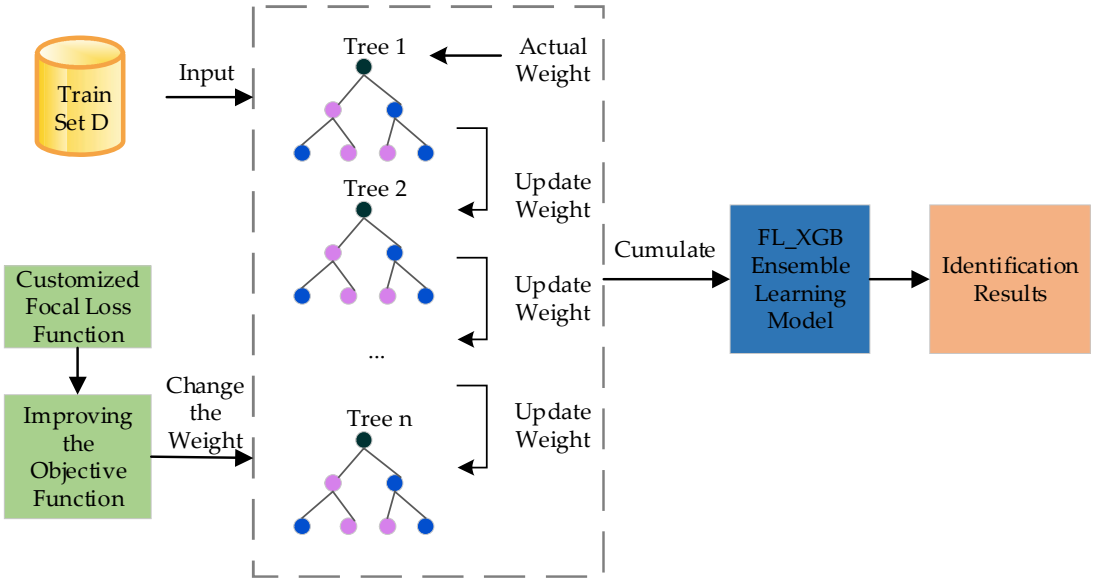


Figure 4. The framework of the FL_XGB model.

The first derivative g_i and the second derivative h_i of the Focal Loss function are shown in Equation (10).

$$g_i = -\left(y_i + (-1)^{y_i} p\right)^{\gamma} \left[(y_i - p) + \gamma(1 - y_i - p) + log\left(1 - y_i - (-1)^{y_i} p\right)\right] \tag{10}$$

Before deriving Equation (10), make:

$$\begin{cases} \beta_1 = \left(y_i + (-1)^{y_i} p\right)^{\gamma} \\ \beta_2 = (y_i - p) \\ \beta_3 = \gamma(1 - y_i - p) \\ \beta_4 = log\left(1 - y_i - (-1)^{y_i} p\right) \end{cases} \tag{11}$$

$$h_i = -[\gamma\left(y_i + (-1)^{y_i} p\right)^{\gamma-1} p(1-p)(\beta_2 + \beta_3 + \beta_4) + \beta_1(-p(1-p) - \gamma p(1-p)\beta_4 - \frac{1}{1 - y_i(-1)^{y_i} p} p(1-p))] \tag{12}$$

According to g_i and h_i, the tth tree $f_t(x_i)$ is trained.

The tth tree $f_t(x_i)$ together with the previous $t-1$ trees $\hat{y}_i^{(t-1)}$ form the strong learner $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$. $\hat{y}_i^{(t)}$ is used to predict the samples in dataset D and obtains the dataset class label predictions y_i. The training set is fed into the strong learner $\hat{y}_i^{(t)}$, i.e., the FL-XGB model, to obtain the identification results.

The pseudo-code of the method is shown in Algorithm 1.

Algorithm 1: FL-XGB-VPN-Encrypted Traffic Identification Algorithm

Input: Train Set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$
Output: Index value of Accuracy, Precision, Recall and F1-Score
Begin
1. For t = 1 to n do
2. If t = 1:
3. Then, initialize sample class label identification values y_i.
4. Based on the FL, Input actual class label Y and predicted value y_i, calculate g_i *and* h_i.
5. according to g_i and h_i training $f_t(x_i)$
6. Strong Learner $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$
7. Use $\hat{y}_i^{(t)}$ training sample obtains identification values y_i
8. End for
9. The train set is fed into $\hat{y}_i^{(t)}$, i.e., the FL-XGB model, to obtain the values of each index.

End

When identifying VPN-encrypted traffic, there is the problem of uneven distribution of all types of data, i.e., the data class imbalance problem. To address this problem, this paper uses the Focal Loss function to improve the Xgboost traffic identification model. The loss of easily identifiable samples is reduced by modifying parameter a, and the loss of samples received is broadened. This method achieves to retain the loss values of the hard-to-identify samples while reducing the loss values of the easy-to-identify samples. We use the FL-XGB-encrypted traffic identification model to solve the problem of imbalance between the number of hard-to-identify and easy-to-identify samples during model training.

2.3. Optimization of FL-XGB-VPN-Encrypted Traffic Identification Model

The problem of low identification rate still exists when using the FL-XGB model for VPN-encrypted traffic. The accuracy of the FL-XGB identification model for VPN-encrypted traffic is affected by parameters, and parameter optimization is needed in combination with parameter search strategies. Bayesian optimization is used to update the posterior distribution of the objective function by continuously adding sample points.

The Bayesian optimization algorithm is based on Bayes' theorem, which is expressed as shown in Equation (13).

$$p(f|D_{1:t}) = \frac{p(D_{1:t}|f)p(f)}{p(D_{1:t})} \tag{13}$$

In Equation (13), f is the unknown objective function, $D_{1:t} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_t, y_t)\}$ is the set of evaluated points, x_t is the decision vector, $y_t = f(x_t) + \varepsilon$ is the observed value, ε

is the observed error, $p(D_{1:t}|f)$ is the likelihood distribution of y, $p(D_{1:t})$ is the marginalized f marginal likelihood distribution, $p(f)$ is the prior probability of f, and $p(f|D_{1:t})$ is the posterior probability of f. After correcting the prior probability by the set of evaluated points, the posterior probability distribution is the confidence level of the parameters in the unknown objective function or parametric model. The Bayesian optimization algorithm uses a probabilistic agent model to fit the true objective function, and selects the next evaluation point based on the acquisition function. The commonly used probabilistic agent models include the Beta-Bernoulli model, linear model, Gaussian process, random forest, etc. Among them, Gaussian process is highly flexible, scalable and analyzable, it is the most widely used probabilistic agent model in Bayesian optimization. The Gaussian process is a paradigmization of the multivariate Gaussian probability distribution, consisting of a mean function and a semi-positive definite covariance function, as shown in Equation (14).

$$y \sim gp(\mu_t(x), k(x, x')) \tag{14}$$

In Equation (14), $\mu_t(x)$ is the mean function and $k(x, x')$ is the covariance function. When fitting discrete data (x_t, y_t) using a Gaussian process, $\mu_t(x)$ is usually set to 0 and $k(x, x')$ is usually used as a Matern covariance function, as shown in Equation (15).

$$k(x, x') = {\sigma_f}^2[1 + \sqrt{5}\frac{r}{\sigma_l} + \frac{5}{3}\left(\frac{r}{\sigma_l}\right)^2]e^{-\sqrt{5}\frac{r}{\sigma_l}} \tag{15}$$

In Equation (15), r is the Euler distance of x and x', σ_f is the characteristic deviation, σ_l is the characteristic length. σ_f and σ_l will change automatically as the Gaussian process is fitted, the initial value of σ_l is the standard deviation of x_i, and the initial value of σ_f is the standard deviation of y_i divided by $\sqrt{2}$. The Bayesian optimization method is an acquisition function based on a strategy of lifting probability and lifting amount, as shown in Equation (16).

$$\alpha_t(x; D_{1:t}) = \begin{cases} (v^* - \mu_t(x))\varnothing\left(\frac{v^* - \mu_t(x)}{\sigma_t(x)}\right) + \sigma_t(x)\varnothing\left(\frac{v^* - \mu_t(x)}{\sigma_t(x)}\right), & \sigma_t(x) > 0 \\ 0, & \sigma_t(x) = 0 \end{cases} \tag{16}$$

In Equation (16), $\alpha_t(x; D_{1:t})$ is the acquisition function, v^* is the current optimal function value, $\varnothing(x)$ is the standard normal distribution cumulative density function. $\mu_t(x)$ and $\sigma_t(x)$ are the mean and variance, respectively. In this paper, the acquisition function is selected based on the confidence interval strategy, and comparison of the maximum of the confidence interval. The location of the next extreme value point of the confidence interval is shown in Equation (17).

$$X_{t+1} = arg\ max\mu_t(x) + \sqrt{\beta}\sigma_t(x) \tag{17}$$

In Equation (17), $\sqrt{\beta}$ denotes the constant used to determine the equilibrium exploration and exploitation, and f_{t+1} is the objective function value. The framework of the VPN-encrypted traffic identification model based on Bayesian optimization is shown in Figure 5.

The algorithm improvement steps for optimizing the FL-XGB model using the Bayesian approach are as follows:

(1) Initialize the FL-XGB model, and choose the FL-XGB model as the training target. The objective function f is the average evaluation value of the identification model. The set of input parameters P and the range of parameter values and the objective function f. Randomly set 6 parameters to be optimized, and determine the number of iterations of the algorithm n = 6;

(2) Update the mean $\mu_t(x)$ and variance $\sigma_t(x)$ of the prior $p(f|D)$ based on the current training data D, and calculate the value of the acquisition function based on the mean and variance of $p(f|D)$;

(3) Determination of the next parameter taking point $X_{n+1} = arg\ max\mu_t(x) + \sqrt{\beta}\sigma_t(x)$ according to the extreme value of the acquisition function;
(4) Calculate the model evaluation value f_n using the new parameter values in the FL-XGB model;
(5) Find the optimal parameter value by the model evaluation value in step (4);
(6) Determine whether the number of iterations reaches n;
(7) When the maximum number of set iterations is reached, the parameter optimization process stops. Output a combination of parameter values with the highest evaluation value;
(8) Otherwise, continue with step 2 Bayesian optimization parameters until the termination condition is met;
(9) The optimal combination of parameters obtained in step (7) is input into the FL-XGB model to obtain the optimal identification model. The final identification result of the output model, i.e., the evaluation value. The final identification result of the model, i.e., the evaluation value f_{n+1}, is output.

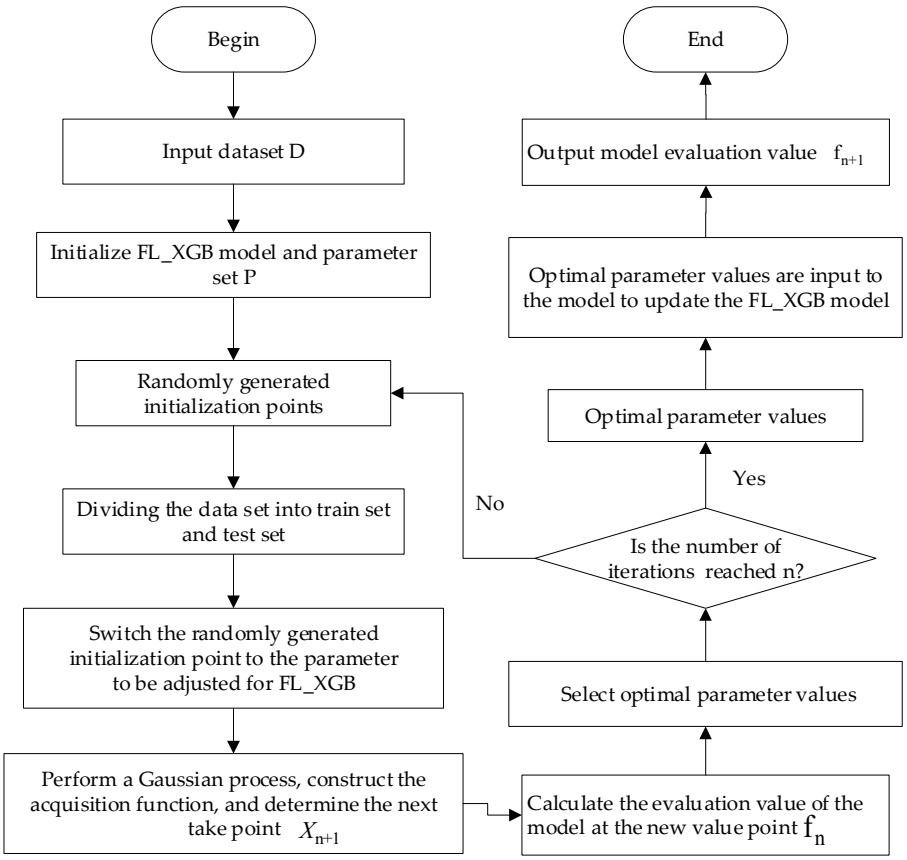


Figure 5. Optimization algorithm framework for encrypted traffic identification model.

In this paper, seven variables of the model are optimized based on the Bayesian optimization algorithm. The probabilistic agent model is selected as a Gaussian process model, and the acquisition function is constructed through a strategy based on the probability of lifting and the amount of lifting. Then, find the optimal value of the objective function of the combined parameters. The objective function is combined with a Bayesian global parameter search strategy to continuously add sample points to update the posterior distribution of the objective function. The global optimal parameter combination is input into the FL-XGB model for training. We found that the FL-XGB-VPN-encrypted traffic identification model optimized by the Bayesian optimization method can directly improve the identification rate of the model by optimizing the parameters. At the same time, the method has the advantages of fewer iterations and faster speed.

3. Example Analysis

3.1. Data Sources

This paper is based on the ISCX VPN-NonVPN dataset [27], which is the time flow feature data. The dataset has 14 kinds of traffic data, including 7 kinds of conventional encrypted traffic and 7 kinds of VPN-encrypted traffic. This dataset has become a common dataset for current research. In this paper, the Scenario A2 folder data of the ISCX VPN-NonVPN dataset is identified and researched. It contains 7 kinds of VPN-encrypted traffic. After integrating the stream data at different times, the final ensemble data is 37,028. Each file contains 23 time-related feature columns. The traffic classes and corresponding application content in the dataset are shown in Table 1, and the time-related features and their descriptions are shown in Table 2.

Table 1. The traffic classes and traffic content.

Traffic	Content
Browsing	Firefox and Chrome
Email	SMPTS, POP3S and IMAPS
Chat	ICQ, AIM, Skype, Facebook and Hangouts
Streaming	Vimero and Youtube
File Transfer	Skype, FTPS and SFTP
VoIP	Facebook, Skype, and Hangouts
P2P	Utorrent and Transmission

Table 2. The time-related features and their descriptions.

Feature	Description
Duration	The duration of the flow.
FIAT	Forward Inter Arrival Time, the time between two packets sent forward direction (mean, min, max, std).
BIAT	Backward Inter Arrival Time, the time between two packets sent backward (mean, min, max, std).
Flow-IAT	Flow Inter Arrival Time, is the time between two packets sent in either direction (mean, min, max, std).
Active	The amount of time a flow was active before going idle (mean, min, max, std).
Idle	The amount of time a flow was idle before becoming active (mean, min, max, std).
FB-psec	Flow Bytes per second.
FP-psec	Flow packets per second.

The dataset is generated by capturing traffic with Wireshark and tcpdump. This process uses an external VPN service provider and connects to the VPN using OpenVPN. In order to generate SFTP and FTPS traffic, an external service provider and Filezilla are used as clients. Filters are used to only capture data packets with source or destination IP, that is, the address of the local client. The capture path of VPN-encrypted traffic is shown in Figure 6.

3.2. Experimental Analysis and Verification

3.2.1. Time-Related mRMR Feature Selection

The use of the time-related mRMR feature selection method facilitates the removal of redundant and irrelevant features and enhances the model's understanding between features and feature values. Unimportant features are deleted according to their importance, and the importance of data features is shown in Table 3 and Figure 7.

After the ranking of feature importance is obtained, feature selection needs to be performed. Too much or too little feature selection can have an impact on the accuracy of the model. Therefore, several different numbers of features n are chosen and passed through several Experiments. The effect of different n values on the Accuracy value of the model is compared each time, and the best number of features n is selected from them. The effect of the features number on the accuracy of the model is shown in Table 4 and Figure 8.

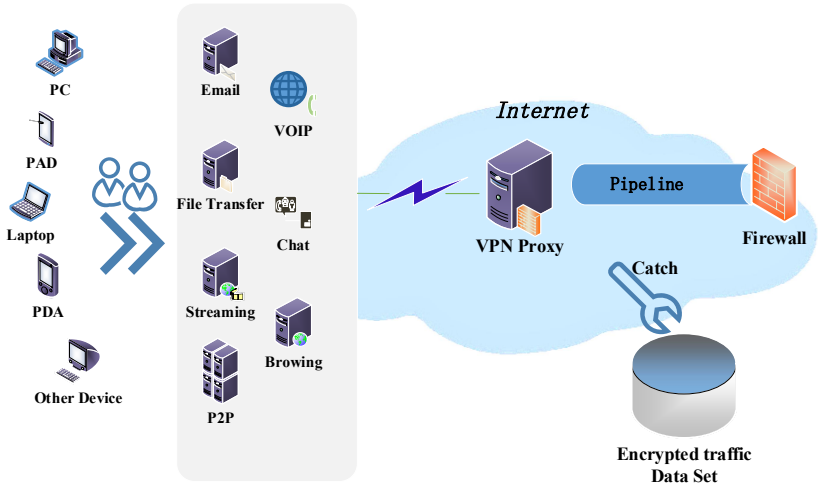


Figure 6. Encrypted traffic capture path.

Table 3. The importance of data features.

Feature Number	Feature Name	Importance Ordering
0	duration NUMERIC	18
1	total_fiat NUMERIC	23
2	total_biat NUMERIC	15
3	min_fiat NUMERIC	16
4	min_biat NUMERIC	14
5	max_fiat NUMERIC	21
6	max_biat NUMERIC	3
7	mean_fiat NUMERIC	19
8	mean_biat NUMERIC	20
9	flowPktsPerSecond NUMERIC	8
10	flowBytesPerSecond NUMERIC	2
11	min_flowiat NUMERIC	11
12	max_flowiat NUMERIC	7
13	mean_flowiat NUMERIC	17
14	std_flowiat NUMERIC	22
15	min_active NUMERIC	5
16	mean_active NUMERIC	9
17	max_active NUMERIC	12
18	std_active NUMERIC	13
19	min_idle NUMERIC	4
20	mean_idle NUMERIC	6
21	max_idle NUMERIC	10
22	std_idle NUMERIC	1

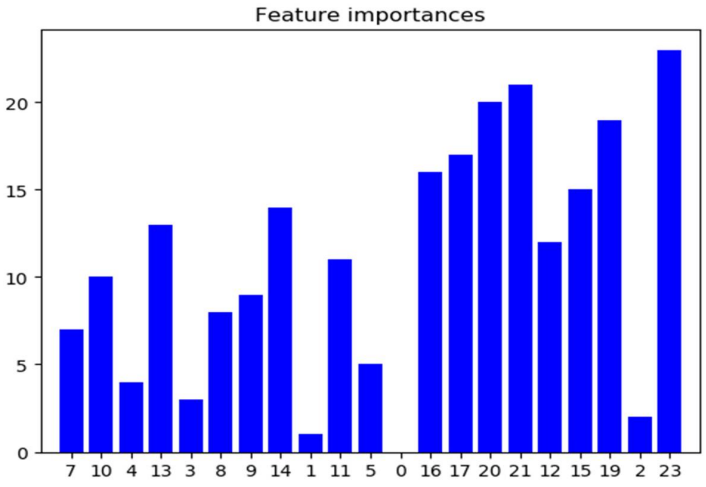


Figure 7. Importance of data features.

Table 4. The effect of the features number on the accuracy of the model.

The Number of Features	Accuracy	Removed Feature Number	Removed Feature Name
23	0.9617	Non	Non
22	0.9630	1	total_fiat NUMERIC
21	0.9633	14	std_flowiat NUMERIC
20	0.9636	5	max_fiat NUMERIC
19	0.9623	8	mean_biat NUMERIC
18	0.9614	7	mean_fiat NUMERIC
17	0.9610	0	duration NUMERIC
16	0.9636	13	mean_flowiat NUMERIC
15	0.9503	3	min_fiat NUMERIC
14	0.9488	2	total_biat NUMERIC
13	0.9376	4	min_biat NUMERIC
12	0.9380	18	std_active NUMERIC
11	0.9388	17	max_active NUMERIC
10	0.9312	11	min_flowiat NUMERIC
9	0.9312	21	max_idle NUMERIC
8	0.9313	16	mean_active NUMERIC
7	0.9150	9	flowPktsPerSecond NUMERIC
6	0.8794	12	max_flowiat NUMERIC
5	0.8772	20	mean_idle NUMERIC
4	0.8658	15	min_active NUMERIC
3	0.8514	19	min_idle NUMERIC
2	0.7038	6	max_biat NUMERIC
1	0.4016	10	flowBytesPerSecond NUMERIC
0	0	22	std_idle NUMERIC

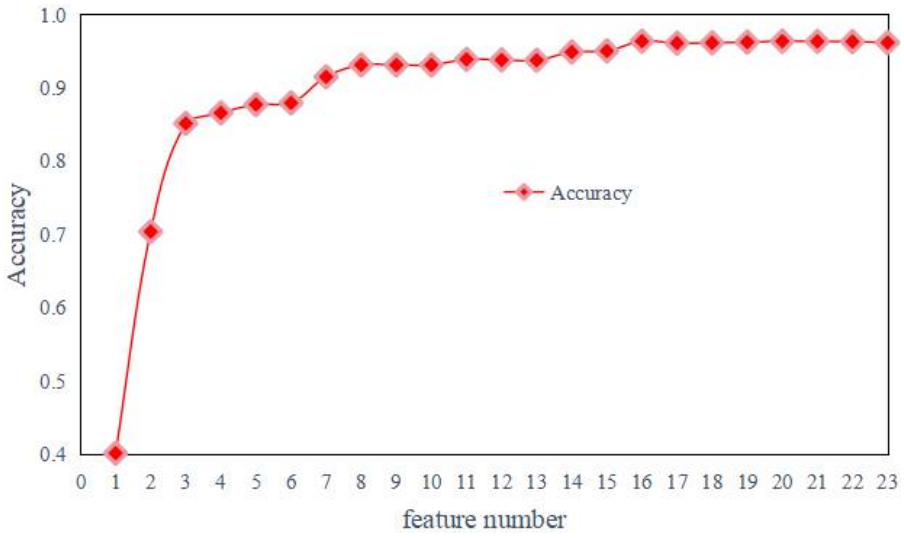


Figure 8. Effect of number of features on model accuracy.

From the experimental results, when the number of features is n = 20, the accuracy obtained by using xgboost is the highest. At the same time, the feature redundancy is the lowest. Therefore, the feature number n is selected as 20 in this paper. According to Table 4, the selected features are the top 20 comprehensively ranked features. We verify the importance of the top 20 features of the composite ranking, as shown in Figure 9.

The encrypted traffic data features before mRMR feature selection are selected to generate a heat map about the correlation coefficients between different features. We verify the correlation of encrypted traffic features before and after feature selection, as shown in Figures 10 and 11.

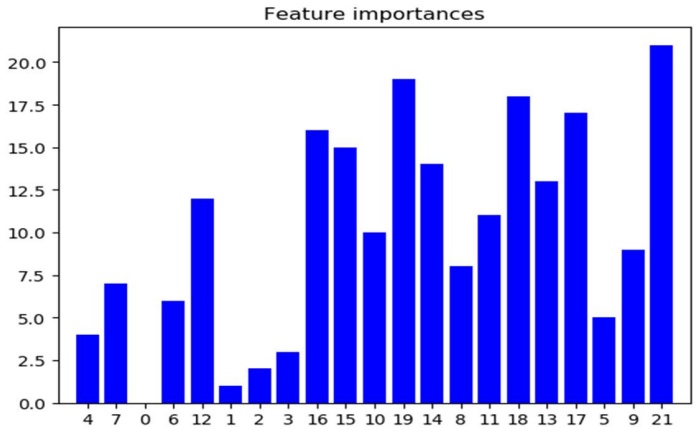


Figure 9. Importance of data features.

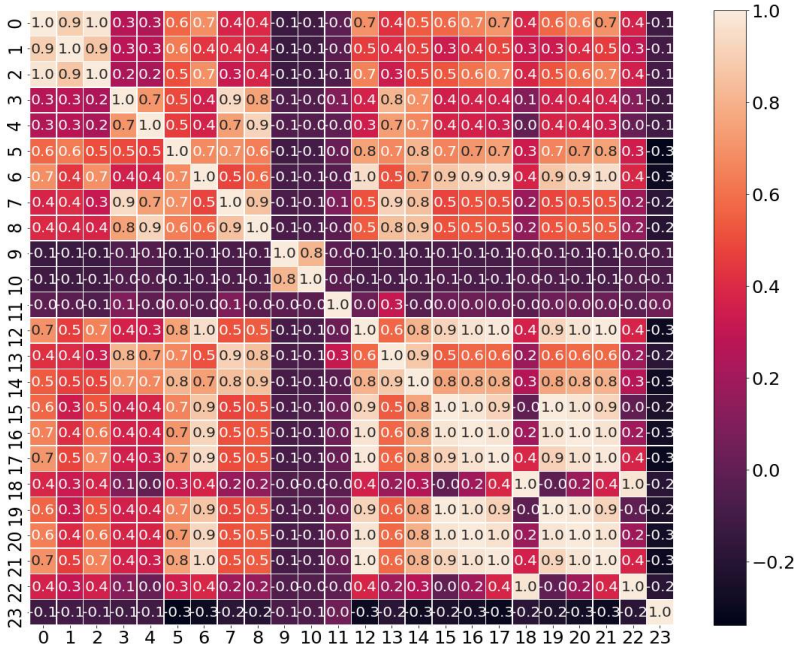


Figure 10. Heat map of coefficient correlation before feature selection.

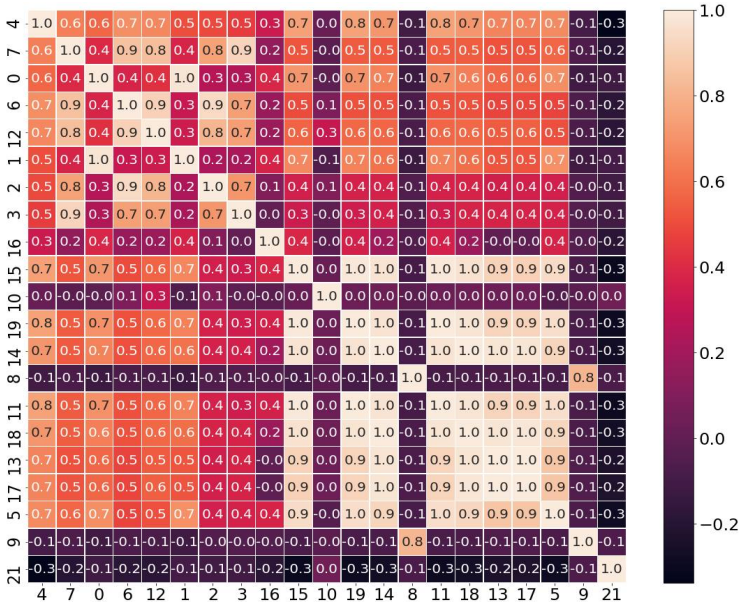


Figure 11. Coefficient correlation heat map after feature selection.

When the absolute value of Pearson's correlation coefficient is close to 0, the correlation between features is low, whereas the correlation between features close to 1.0 is higher. The mRMR feature selection removes feature numbers 1 and 14, respectively. It can be seen from Figure 11 that the correlation between feature numbers 1 and 2 is 0.9, and the correlation between feature numbers 14 and 13 is 0.9. Both pairs of features have a high correlation, which proves the feasibility of the mRMR feature selection results. The correlation between features of the encrypted traffic data after mRMR feature selection is lower than the correlation between features before going through this experiment. So mRMR feature selection on this dataset can make the redundancy between features reduced while ensuring the maximum correlation between features and classes. This method also solves the problem of feature redundancy that exists in the original data.

3.2.2. Construction and Optimization of FL-XGB-VPN-Encrypted Traffic Identification Model

In the VPN-encrypted traffic identification problem, the training data of each class is not uniformly distributed. The classes with more data have a large proportion in the dataset, whereas the classes with less data will have a smaller proportion in the dataset. The reason is that different applications use different frequencies and generate different amounts of traffic data. Therefore, the amount of data varies greatly between different classes, and there is a problem of unbalanced data volume between data classes. Statistic on VPN-encrypted traffic data is shown in Figure 12.

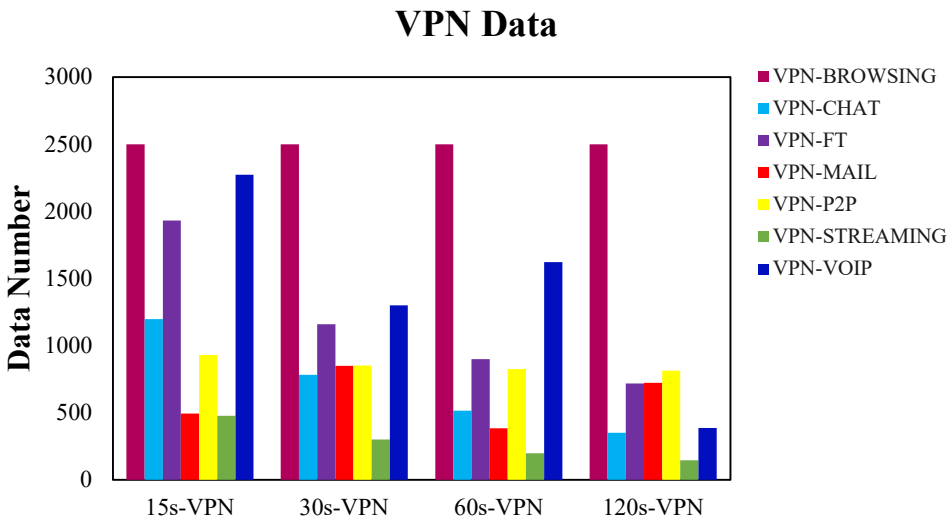
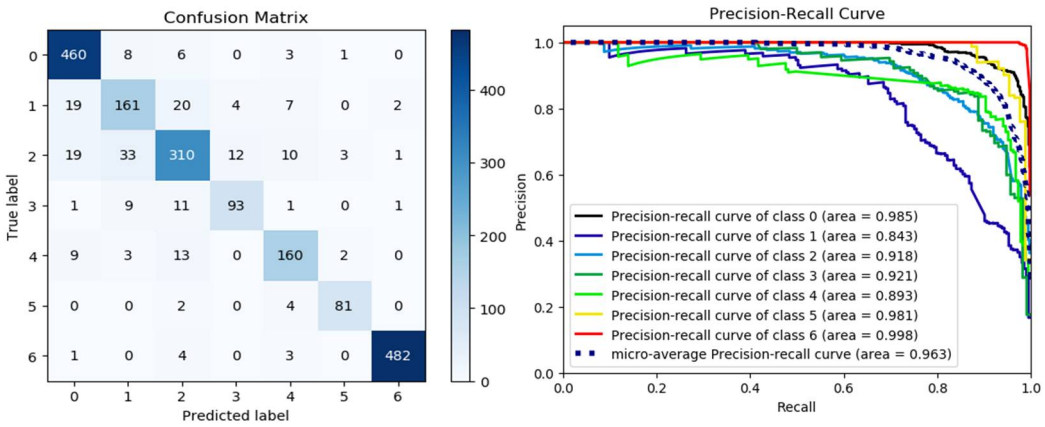


Figure 12. VPN-encrypted traffic data statistics.

For the class imbalance problem in VPN-encrypted traffic data identification, we use the Xgboost ensemble learning model of custom Focal Loss, i.e., the FL-XGB model. The introduction of the Focal Loss function solves the problem of difficult-to-identify and easy-to-identify samples imbalance when identifying VPN traffic data. The method balances the problem of uneven proportions of the two samples, thus providing the necessary conditions for overall traffic identification. Among them, Accuracy is 0.9543, Precision is 0.9522, F1-Score is 0.9503, and Recall is 0.9496. Therefore, the FL-XGB model obtained by using Focal Loss to improve Xgboost can better deal with the problem of imbalance between different types of data. The confusion matrix and precision–recall curve are shown in Figure 13.

Each column of the confusion matrix represents the identified class, and the total number of each column indicates the number of data identified as that class. Each row represents the actual class of the data, and the total amount of data in each row indicate the number of data instances in that class. From the confusion matrix in Figure 13a, the value in each column indicates the number of actual data identified as the class. For example, 460 in the first column of the first row means that 460 instances belonging to class 0 are identified as class 0. Similarly, the 8 in the second column of the first row indicates that

there are 8 class 0 that are incorrectly identified as class 1. From the precision–recall curves in Figure 13b, it can be seen that the precision–recall rate for class 0 is 0.998, for class 1 is 0.843, for class 2 is 0.918, for class 3 is 0.921, for class 4 is 0.893, for class 5 is 0.981, and for class 6 is 0.998. The average precision–recall rate is 0.963.



(**a**) Confusion Matrix (**b**) Precision–Recall Curve

Figure 13. Confusion Matrix and Precision–Recall Curve.

To compare the performance of the Focal Loss function in the FL-XGB model proposed in this paper and the cross-entropy loss function in the original data, the comparison results of the two functions are given in Table 5 and Figure 14. It can be seen that the Xgboost model with the improved Focal Loss function has better performance than the original cross-entropy loss function. The γ in the Focal Loss function adjusts the rate of weight reduction in easily identifiable samples. It is equivalent to using the standard cross-entropy loss function when $\gamma = 0$. When $\gamma = 2$, the model can effectively reduce the weight of easily identifiable samples and increase the weight of the difficult identifiable samples. The model focuses more on the identification of difficult samples, resulting in higher Accuracy, Precision, F1-Score, and Recall. Among them, higher Precision performance indicates that the number of correctly identified samples is higher in total and the model has better sample identification ability.

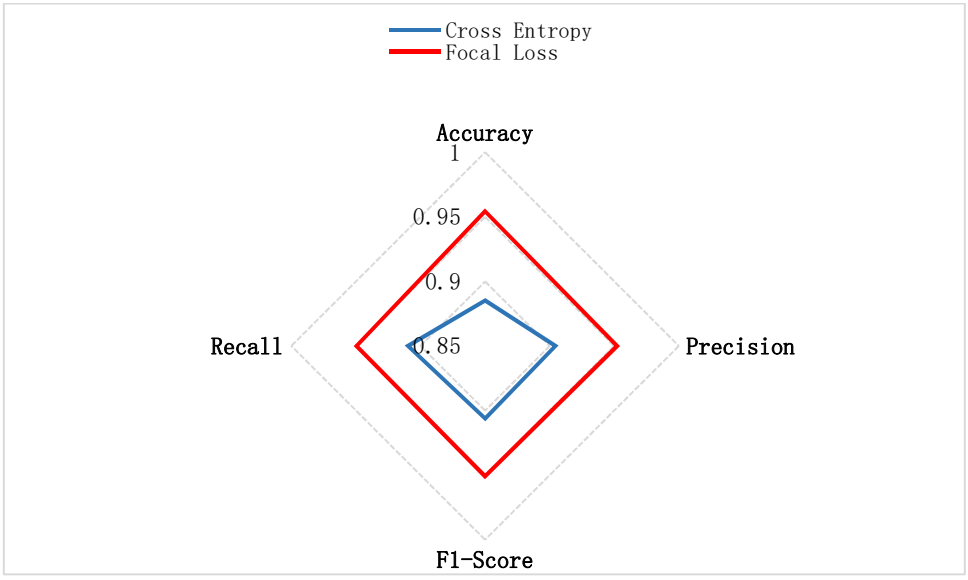


Figure 14. Performance Comparison of Cross-Entropy and Focal Loss Functions.

Table 5. Cross entropy and Focal Loss function performance.

	Cross Entropy	Focal Loss (λ=2)
Accuracy	0.8852	0.9543
Precision	0.9045	0.9522
F1-Score	0.9062	0.9509
Recall	0.9097	0.9496

In this paper, we optimize the parameters to change the descriptive power of the model, to reduce the error of the model and increase the identification rate of the model for VPN-encrypted traffic. The parameters min_child_weight, gamma, max_depth, Eta, subsample, and colsample_bytree are optimized to improve the identification rate of the FL_XGB model for encrypted traffic. The optimal combination of parameter values for the FL-XGB model is searched using Bayesian optimization, and the set of parameters of the FL-XGB model and its optimal search range are shown in Table 6.

Table 6. Parameters and their value ranges.

Parameter Name	Value Ranges
Max_depth	(1, 15)
Eta	(0, 1)
Min_child_weight	(0.1, 20)
Gamma	(0, 20)
Subsample	(0, 1)
colsample_bytree	(0, 1)

By Bayesian optimization algorithm, the maximum depth of the tree Max_depth is 10, the learning rate Eta is 0.3, the minimum leaf weight Min_child_weight is 4, the minimum loss function descent value Gamma is 7.7, Subsample is 0.5, and the feature sampling ratio colsample_bytree is 0.2 after the optimization. The optimal combination of parameters after optimization is shown in Table 7.

Table 7. The optimal combination of parameters.

Parameter Name	Value
Max_depth	10
Eta	0.3
Min_child_weight	4
Gamma	7.7
Subsample	0.5
colsample_bytree	0.2

The optimal set of parameters obtained from Bayesian optimization is input into the FL-XGB model to obtain the ROC curve of the model. The area of the ROC curve indicates its characteristics. When the area enclosed by the ROC curve is 0.5 for random classification, the identification ability of the model is 0. The closer the area is to 1 indicates that the identification ability of the FL-XGB model is stronger. From Figure 15, the area of the ROC curve is close to 1, indicating that the FL-XGB model has a strong identification effect on VPN-encrypted traffic.

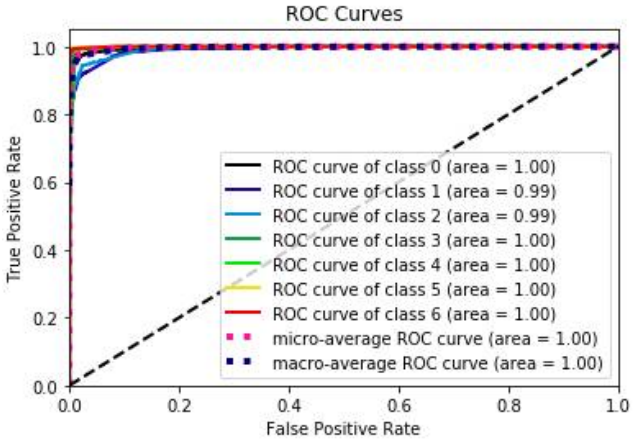


Figure 15. ROC curve.

3.2.3. Performance Analysis and Comparison of FL-XGB-Encrypted Traffic Identification Model

We use Support Vector Machine (SVM), Gradient Boosting Decision Tree (GDBT), Naive_Bayes (NB), Logistic Regression (LR), K-nearest neighbor (KNN), Adaptive Boosting (Adaboost), Linear Discriminant Analysis (LDA), FL_XGB a total of 8 algorithms to identify the VPN traffic data, and then the four metrics Accuracy, Precision, Recall, F1-Score are output. The identification results are shown in Table 8.

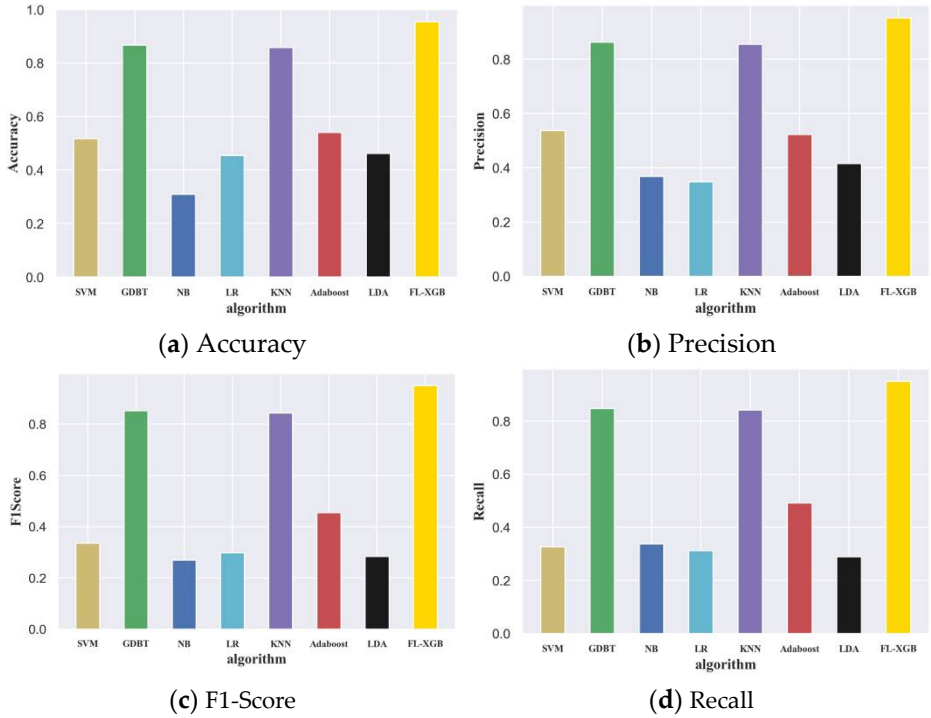
Table 8. Identification results of different algorithms.

Algorithm	ACC	Precision	F1-Score	Recall
SVM	0.5169	0.5374	0.3361	0.3271
GDBT	0.8666	0.8626	0.8521	0.8475
NB	0.3093	0.3685	0.2691	0.3379
LR	0.4548	0.3489	0.2977	0.3126
KNN	0.8575	0.8553	0.8433	0.8416
Adaboost	0.5408	0.5224	0.4545	0.4913
LDA	0.4623	0.4153	0.2841	0.2892
FL-XGB	0.9743	0.9722	0.9703	0.9796

The performance comparison results of each algorithm under the indicators of Accuracy, Precision, F1-Score, and Recall are shown in Figure 16.

From Figure 16, it can be seen that the FL-XGB-VPN-encrypted traffic identification method based on Time-Related mRMR feature selection, FL-XGB traffic identification model construction, and optimization proposed in this paper achieves an identification rate of more than 97%.

In VPN-encrypted traffic identification, the model improves the accuracy by 8.21% over the traditional Xgboost identification model. In addition, the feature selection stage of the FL-XGB-VPN-encrypted traffic identification model preserves the attributes of the original features while ensuring the dimensionality reduction effect. The method maintains key features of the data and provides a basis for analyzing key features of VPN-encrypted traffic. In the parameter optimization stage of the identification model, our approach avoids large-scale searches while allowing the model to converge quickly to the optimal solution. It can be seen that the FL-XGB-VPN-encrypted traffic identification model has high identification ability and promotion ability.



(**a**) Accuracy (**b**) Precision

(**c**) F1-Score (**d**) Recall

Figure 16. Performance comparison chart of different algorithms.

Based on the above results, to verify the universality of the FL-XGB-encrypted traffic identification model, another dataset is used for verification. On the ISCXTor2016 [36] dataset, compare the identification performance of the FL-XGB-encrypted traffic identification model with a total of 8 algorithms including SVM, GDBT, Naive_Bayes, LR, KNN, Adaboost, LDA, and C4.5 for Tor-encrypted traffic data. The four indicators of Accuracy, Precision, Recall, and F1-Score of the 10S feature data of the dataset Scenario-B file are shown in Table 9.

Table 9. Identification results of different algorithms on the ISCXTor2016 dataset.

Algorithm	Accuracy	Precision	F1-Score	Recall
SVM	0.6824	0.5169	0.3671	0.5333
GDBT	0.739	0.6278	0.644	0.6949
NB	0.6911	0.5228	0.5441	0.6001
LR	0.5966	0.4398	0.4498	0.4667
KNN	0.7023	0.6775	0.6962	0.6618
Adaboost	0.7079	0.6911	0.5884	0.6321
LDA	0.6824	0.6778	0.69	0.6333
C4.5	0.762	0.7827	0.7969	0.8333
FL-XGB	0.996	0.9959	0.9963	0.9967

The performance comparison of the FL-XGB-encrypted traffic identification model with other eight algorithms for four metrics, namely Accuracy, Precision, F1-Score, and Recall, is shown in Figure 17. Each performance of the FL-XGB-encrypted traffic identification model reaches above 0.995. As shown in Table 9, the average Accuracy value of other machine learning algorithms is 0.6955, the average Precision value is 0.6171, the average F1-Score value is 0.5971, and the average Recall value is 0.6319. The FL-XGB-encrypted traffic identification model has an average improvement of 30.05% in Accuracy value, 37.88% in Precision value, 39.92% in F1-Score value, and 36.47% in Recall value compared to the other 8 algorithms. Precision, F1-Score, and Recall are important indicators of the precision, false positives, and stability of the model performance. Therefore, the FL-XGB-encrypted traffic identification model has higher accuracy and more stable model performance in

identifying Tor-encrypted traffic data compared to the other eight algorithms. In summary, the FL-XGB-encrypted traffic identification model is also applicable to the encrypted traffic identification of the dataset ISCXTor2016. The FL-XGB-encrypted traffic identification model has high universality for the identification of encrypted traffic and can effectively solve the problem of low accuracy of VPN-encrypted traffic identification.

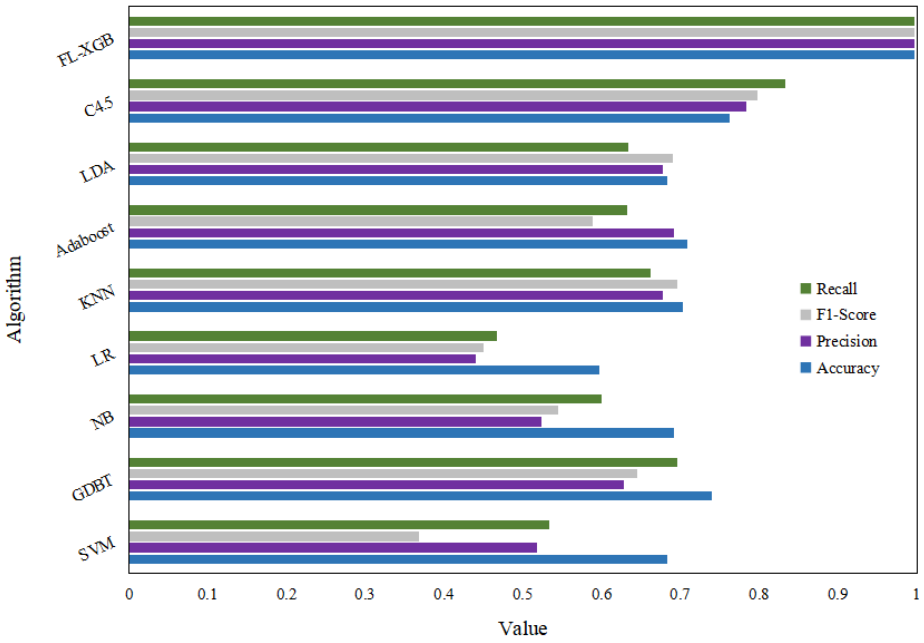


Figure 17. Comparison of different algorithms on the ISCXTor2016 dataset.

4. Conclusions

This paper proposes a VPN-encrypted traffic identification method based on ensemble learning to achieve accurate identification of encrypted traffic. Firstly, a VPN-encrypted traffic feature selection method is proposed. We perform VPN-encrypted traffic mRMR feature selection for the redundant features that exist in Time-Related VPN-encrypted traffic to obtain the optimal Time-Related VPN-encrypted traffic feature set. Secondly, a VPN-encrypted traffic identification model based on ensemble learning is proposed to address the imbalance of data class. Finally, anoptimization method of the VPN-encrypted traffic identification model is proposed to solve the problem of the low identification rate of the previous VPN-encrypted traffic ensemble learning model. In summary, the average identification rate of the FL-XGB-VPN-encrypted traffic identification method proposed in this paper reaches more than 97% for VPN-encrypted traffic. Meanwhile, we also validate the FL-XGB-encrypted traffic identification method proposed in this paper on other encrypted traffic data. The experimental results show that the FL-XGB-encrypted traffic identification method proposed in this paper also has a high identification rate for other encrypted traffic data. This indicates that the set of methods proposed in this paper has high universality in encrypted traffic identification. Therefore, the FL-XGB-VPN-encrypted traffic identification model proposed in this paper can efficiently identify malicious network encryption. The method helps to improve the network service quality and also has certain practical value to maintain network security.

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

Funding: This work was supported by the Science and Technology Development Plan projects of Jilin Province, No. 20210201134GX.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: This study was based entirely on theoretical basic research and did not involve humans.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

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

References

1. Shao, B.; Li, X.; Bian, G. A Survey of Research Hotspots and Frontier Trends of Recommendation Systems from the Perspective of Knowledge Graph. *Expert Syst. Appl.* **2020**, *165*, 113764. [CrossRef]
2. Nisar, K.; Jimson, E.R.; Hijazi, M.H.A.; Welch, I.; Hassan, R.; Aman, A.H.M.; Sodhro, A.H.; Pirbhulal, S.; Khan, S. A Survey on the Architecture, Application, and Security of Software Defined Networking. *Internet Things* **2020**, *12*, 100289. [CrossRef]
3. Gualtieri, L.; Rauch, E.; Vidoni, R. Emerging research fields in safety and ergonomics in industrial collaborative robotics: A systematic literature review. *Robot. Comput.-Ensemble Manuf.* **2020**, *67*, 101998. [CrossRef]
4. Fuentes-García, M.; Camacho, J.; Maciá-Fernández, G. Present and Future of Network Security Monitoring. *IEEE Access* **2021**, *9*, 112744–112760. [CrossRef]
5. Sengupta, S.; Chowdhary, A.; Sabur, A.; Alshamrani, A.; Huang, D.; Kambhampati, S. A survey of moving target defenses for network security. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 1909–1941. [CrossRef]
6. Tahaei, H.; Afifi, F.; Asemi, A.; Zaki, F.; Anuar, N.B. The rise of traffic classification in IoT networks: A survey. *J. Netw. Comput. Appl.* **2020**, *154*, 102538. [CrossRef]
7. Pacheco, F.; Exposito, E.; Gineste, M.; Baudoin, C.; Aguilar, J. Towards the deployment of machine learning solutions in network traffic classification: A systematic survey. *IEEE Commun. Surv. Tutor.* **2018**, *21*, 1988–2014. [CrossRef]
8. Masdari, M.; Khezri, H. A survey and taxonomy of the fuzzy signature-based Intrusion Detection Systems. *Appl. Soft Comput.* **2020**, *92*, 106301. [CrossRef]
9. Khatouni, A.S.; Heywood, N.Z. How much training data is enough to move a ML-based classifier to a different network? *Procedia Comput. Sci.* **2019**, *155*, 378–385. [CrossRef]
10. Juma, M.; Monem, A.A.; Shaalan, K. Hybrid end-to-end VPN security approach for smart IoT objects. *J. Netw. Comput. Appl.* **2020**, *158*, 102598. [CrossRef]
11. Aceto, G.; Ciuonzo, D.; Montieri, A.; Pescapé, A. Toward effective mobile encrypted traffic classification through deep learning. *Neurocomputing* **2020**, *409*, 306–315. [CrossRef]
12. Bu, Z.; Zhou, B.; Cheng, P.; Zhang, K.; Ling, Z.-H. Encrypted Network Traffic Classification Using Deep and Parallel Network-in-Network Models. *IEEE Access* **2020**, *8*, 132950–132959. [CrossRef]
13. Cao, Z.; Xiong, G.; Zhao, Y.; Li, Z.; Guo, L. *A Survey on Encrypted Traffic Classification*; International Conference on Applications and Techniques in Information Security; Springer: Berlin/Heidelberg, Germany, 2014; pp. 73–81.
14. Aceto, G.; Ciuonzo, D.; Montieri, A.; Pescape, A. Mobile encrypted traffic classification using deep learning: Experimental evaluation, lessons learned, and challenges. *IEEE Trans. Netw. Serv. Manag.* **2019**, *16*, 445–458. [CrossRef]
15. Rezaei, S.; Liu, X. Deep learning for encrypted traffic classification: An overview. *IEEE Commun. Mag.* **2019**, *57*, 76–81. [CrossRef]
16. Handa, A.; Sharma, A.; Shukla, S.K. Machine learning in cybersecurity: A review. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1306. [CrossRef]
17. Ribeiro, V.H.A.; Reynoso-Meza, G. Ensemble learning by means of a multi-objective optimization design approach for dealing with imbalanced data sets. *Expert Syst. Appl.* **2020**, *147*, 113232. [CrossRef]
18. Meng, F.; Cheng, W.; Wang, J. Semi-supervised Software Defect Prediction Model Based on Tri-training. *KSII Trans. Internet Inf. Syst. (TIIS)* **2021**, *15*, 4028–4042. [CrossRef]
19. Xibin, D.; Zhiwen, Y.; Wenming, C.; Yifan, S.; Qianli, M. A survey on ensemble learning. *Front. Comput. Sci.* **2020**, *14*, 241–258. [CrossRef]
20. Paxson, V. Empirically derived analytic models of wide-area TCP connections. *IEEE/ACM Trans. Netw.* **1994**, *2*, 316–336. [CrossRef]
21. Sen, S.; Spatscheck, O.; Wang, D. Accurate, scalable in-network identification of p2p traffic using application signatures. In Proceedings of the 13th International Conference on World Wide Web, New York, NY, USA, 17 May 2004; pp. 512–521.
22. Lotfollahi, M.; Siavoshani, M.J.; Zade, R.S.H.; Saberian, M. Deep packet: A novel approach for encrypted traffic classification using deep learning. *Soft Comput.* **2020**, *24*, 1999–2012. [CrossRef]
23. Dutt, I.; Borah, S.; Maitra, I.K. Multiple Immune-based Approaches for Network Traffic Analysis. *Procedia Comput. Sci.* **2020**, *167*, 2111–2123. [CrossRef]
24. Yao, Z.; Ge, J.; Wu, Y.; Lin, X.; He, R.; Ma, Y. Encrypted traffic classification based on Gaussian mixture models and Hidden Markov Models. *J. Netw. Comput. Appl.* **2020**, *166*, 102711. [CrossRef]

25. Chang, L.; Zigang, C.; Gang, X.; Gaopeng, G.; Siu-Ming, Y.; Longtao, H. MaMPF: Encrypted Traffic Classification Based on Multi-Attribute Markov Probability Fingerprints. In Proceedings of the 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), Banff, AB, Canada, 4–6 June 2018; pp. 1–10.
26. Gijon, C.; Toril, M.; Solera, M.; Luna-Ramirez, S.; Jimenez, L.R. Encrypted Traffic Classification Based on Unsupervised Learning in Cellular Radio Access Networks. *IEEE Access* **2020**, *8*, 167252–167263. [CrossRef]
27. Draper-Gil, G.; Habibi Lashkari, A.; Mamun, M.S.; 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. Available online: https://www.unb.ca/cic/datasets/vpn.html (accessed on 1 June 2022).
28. Raikar, M.M.; Meena, M.S.; Mulla, M.M.; Shetti, N.S.; Karanandi, M. Data Traffic Classification in Software Defined Networks (SDN) using supervised-learning. *Procedia Comput. Sci.* **2020**, *171*, 2750–2759. [CrossRef]
29. Dias, K.; Pongelupe, M.A.; Caminhas, W.M.; de Errico, L. An innovative approach for real-time network traffic classification. *Comput. Netw.* **2019**, *158*, 143–157. [CrossRef]
30. Shekhawat, A.S.; Di Troia, F.; Stamp, M. Feature analysis of encrypted malicious traffic. *Expert Syst. Appl.* **2019**, *125*, 130–141. [CrossRef]
31. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
32. Sagi, O.; Rokach, L. Ensemble learning: A survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1249. [CrossRef]
33. Bolón-Canedo, V.; Alonso-Betanzos, A. Ensembles for feature selection: A review and future trends. *Inf. Fusion* **2019**, *52*, 1–12. [CrossRef]
34. Takeda, K.; Kabashima, Y. Multi-Label Feature Selection Algorithm Based on Information Entropy. *J. Comput. Res. Dev.* **2013**, *50*, 1177–1184.
35. Berk, R.A. Classification and Regression Trees (CART). In *Statistical Learning from a Regression Perspective*; Springer Series in Statistics; Springer: New York, NY, USA, 2008.
36. Lashkari, A.H.; Gil, G.D.; Mamun, M.; Ghorbani, A.A. Characterization of Tor Traffic using Time based Features. In Proceedings of the International Conference on Information Systems Security & Privacy, Porto, Portugal, 1 January 2017.