Neural Processing Letters (2023) 55:3185–3203
https://doi.org/10.1007/s11063-022-11005-2

A Block Cipher Algorithm Identification Scheme Based on Hybrid Random Forest and Logistic Regression Model

Ke Yuan[1,2] · Yabing Huang[1] · Jiabao Li[1] · Chunfu Jia[3] · Daoming Yu[1]

Accepted: 9 August 2022 / Published online: 20 August 2022

Abstract
Cryptographic algorithm identification is aimed to analyze the potential feature information in ciphertext data when the ciphertext is known, which belongs to the category of cryptanalysis. This paper takes block cipher algorithm as the research object, and proposes a block cipher algorithm identification scheme based on hybrid random forest and logistic regression (HRFLR) model with the idea of ensemble learning. Based on the NIST randomness test feature extraction method, five block ciphers, AES, 3DES, Blowfish, CAST and RC2, are selected as the research object of cryptographic algorithm identification to carry out the ciphertext classification tasks. The experimental results show that, compared with the existing methods, the cryptographic algorithm identification scheme based on HRFLR proposed in this paper has higher accuracy and stability on binary classification and multi-class classification tasks. In the binary classification tasks of AES and 3DES, the identification accuracy of our proposed cryptographic algorithm identification scheme based on HRFLR can reach up to 74%, and the highest identification accuracy of the five classification tasks is 38%. Compared with the 54% and 28.8% accuracies of random forest-based identification scheme, the accuracy is increased by 37.04% and 18.06%, respectively. This result is significantly better than the 50% and 20% accuracies of random guessing scheme.

Keywords Cryptographic algorithm identification · Ensemble learning · Random forest · Logistic regression

1 Introduction

The cryptanalysis technology is the original technology of code breaking and is used to measure the security of the cryptographic algorithm. Kerckhoffs' basic hypothesis of cryptanalysis explains all the details of cryptographic algorithms and implementations known to

✉ Jiabao Li
iiththan@163.com

[1] School of Computer and Information Engineering, Henan University, Kaifeng 475004, China

[2] Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng 475004, China

[3] College of Cybersecurity, Nankai University, Tianjin 300350, China

cryptanalysts. Most of the existing cryptanalysis techniques are based on the assumption that the cryptographic algorithm to which the ciphertext belongs is known. However, in the actual situation, when facing the ciphertext data obtained, researchers usually do not know the cryptographic algorithm to which it belongs. Recognizing the encryption algorithm used in the ciphertext is a necessary prerequisite for further analysis of the ciphertext, but various cryptographic algorithms are emerging in an endless stream. In order to develop the solution of ciphertext data analysis, it has become the primary task for ciphertext data analysts to recognize the cryptographic algorithm to which the ciphertext data belongs. Therefore, it has important theoretical significance and practical application value to carry out the research on the identification of cryptographic algorithm.

1.1 Related Work

The initial research object of cryptographic algorithm identification is classical cryptography. With the rapid development of modern cryptographic technology, the traditional cryptographic algorithm identification technology based on statistical method is gradually ineffective. Researchers have proposed a series of cryptographic algorithm identification based on machine learning methods. In 2006, Dileep and Sekhar[1]proposed a method to identify block cipher encryption algorithms using support vector machines, and did experiments to distinguish the ciphertext of AES, DES, TDES, RC5 and Bblowfish cipher algorithms under ECB and CBC modes. The results show that the support vector machine model using Gaussian kernel function has the best performance, and ECB mode has better identification accuracy than CBC mode. In 2011, Manjula and Anitha proposed a cryptographic algorithm identification scheme based on C4.5 Decision Tree (DT)[2]. This scheme extracts eight ciphertext features and recognizes eleven cryptographic algorithms, and the identification rate obtained is 70% to 75%. In 2012, Chou et al. proposed to use Support Vector Machine (SVM) to identify cryptosystems[3]. The solution extracts 12 ciphertext features, and distinguishes between the ciphertexts of the two cryptographic algorithms of AES and DES in ECB and CBC modes. It is believed that ECB mode has better performance on some data sets than CBC. In 2018, Mello and Xexeo[4] analyzed ciphertext files encrypted by seven cryptographic algorithms, Arc4, Blowfish, DES, Rijdael, RSA, Serpent and Twofish, in ECB and CBC modes. The ECB can achieve full identification on almost all cryptographic algorithms. In addition, the application of machine learning in homomorphic encryption has become an important research area[5, 6]. As deep learning enters the explosive stage, related research has made great progress, and the application scenarios of deep learning technology are constantly enriched[7, 8]. The identification of cryptographic algorithms combining pattern identification and deep learning has also attracted more and more researchers' attention. In 2013, Mishra et al. proposed a block cipher and stream cipher identification scheme combining pattern identification and decision tree[9]. The identification rate of cryptographic algorithm is greatly improved through block length detection, recurrence analysis, and decision tree based. In 2014, Willam et al. proposed a neural network combined with linguistics and information retrieval methods to extract ciphertext sets from ciphertext files encrypted by five cryptographic algorithms, and then submit the obtained ciphertext sets to the "clustering process"[10]. Finally, the clustering result is processed by the classifier to realize the identification of the cryptographic algorithms. In 2015, Wu et al. proposed a hierarchical identification scheme based on K-means clustering, which has a identification rate of about 90% for typical block ciphers[11]. In addition to the single-layer identification scheme that directly identifies specific cryptographic algorithms, Huang et al. proposed a random forest-based cryptographic algorithms hierarchical identification scheme in 2018, and introduced three clustering methods: CM-cluster, CSN-cluster points and CSBP-cluster points[12]. The results show that the identification effect of adding a layered identification scheme is better than that of a single-layer cryptographic algorithm. In 2019, Zhao et al. used randomness testing to extract ciphertext features, and proposed a identification scheme for basis and randomness testing[13]. In this scheme, six block ciphers are identified in pairs, and their identification rate is different under different features, and the identification rate can reach up to more than 80% under some features. In the same year, Arvind and Ram[14] proposed a method using bit-plane image features and fuzzy decision criteria to realize the identification and isolation of images encrypted with the same key.

1.2 Problem Statement

In summary, most of the existing identification schemses are mainly developed around block ciphers, which are based on statistical methods to collect ciphertext features, and then use classic machine learning classification algorithms for classification. Due to the differences in the size of the collected ciphertext files, file types, ciphertext feature extraction methods, and the choice of classifiers, there is a large gap in the accuracy and stability of the cryptographic algorithm identification results. Therefore, this paper uses ensemble learning ideas to improve the classic classification algorithm and proposes a hybrid random forest and logistic regression (HRFLR) model, and applies it to the cryptographic algorithm identification task of ciphertext files of different file sizes to improve the accuracy and stability of the cryptographic algorithm identification scheme[15–17]. At the level of method composition, the scheme can be divided into two main parts: ciphertext feature extraction and the construction of a cryptographic algorithm identification classifier. The ensemble learning is used to construct the classifier, and the extracted various ciphertext features data are used as the input of the classifier. After the training and testing of the classification model, the identification task of the cryptographic algorithm is finally completed. The main contributions of this paper are summarized as follows.

- According to the features of random forest and logistic regression, this paper proposes a hybrid random forest and logistic regression (HRFLR) model based on the idea of ensemble learning. The model integrates several weakly supervised models in order to obtain a more comprehensive strong supervised model and achieve better classification accuracy.
- The experimental results show that, compared with the existing identification schemes, the proposed scheme has higher identification accuracy in block cipher binary classification and five-class classification tasks under the same ciphertext files size. At the same time, as the size of the ciphertext files changes, the identification accuracy fluctuates. The scheme proposed in this paper has the smallest fluctuation range, the smallest degree of influence and the highest stability.
- Compared with the traditional cryptographic algorithm identification scheme, the scheme proposed in this paper has stronger stability, higher identification efficiency, and more flexible parameter setting in classification problems. The scheme can improve the recognition accuracy and stability problems caused by the increase of the number of cryptographic algorithms, the complexity of ciphertext data and the increase of interference between data to a certain extent.

The rest of this paper is organized as follows. Section 2 discusses the relevant principles of cryptographic algorithm identification and gives definitions. In Sect. 3, an overview of ciphertext feature extraction based on randomness test is given. Section 4 discusses the random forest algorithm and the HRFLR model, and gives the block cipher algorithm identification scheme based on hybrid random forest and logistic regression model. Section 5 presents the experimental design scheme and the criteria for evaluating the experimental results . Section 6 shows the experimental results and provides evaluation analysis and comparison. Finally, it summarizes the current work and makes some suggestions for future work.

2 Cryptographic Algorithm Identification

Statistical methods and machine learning methods are the two main methods to design cryptographic algorithm identification schemes. Different cryptographic algorithms have different design concepts, configurations of cryptographic components, modes of operation, and key setting methods, as well as factors including plaintext types, which may cause differences in the spatial distribution of the ciphertexts. The core of the cryptographic algorithm identification task is to distinguish the small differences in these ciphertext data, so as to achieve the goal of identifying the cryptographic algorithms to which the ciphertext belongs.

The identification scheme based on machine learning methods treats features as a set of attributes that reflecting ciphertext information, and transforms the identification task into a supervised learning task that machine learning is good at. Firstly, learn and train a classifier model on the training data set, and then use the classifier model to classify the input ciphertext data. However, it is difficult to carry out in-depth discussion on the particularity of cryptographic algorithms and ciphertext data by putting the problem of cryptographic algorithm identification into the general classification learning framework. This is a problem in the identification of cryptographic algorithms with machine learning technology as the core. Based on the above considerations, this section discusses and standardizes cryptographic algorithm identification related issues, further improves the definitions of the basic elements of cryptographic algorithm identification issues such as ciphertext, ciphertext features, cryptographic algorithm identification, and cryptographic algorithm identification scheme, and gives formalized definition descriptions. Later, in the fourth section, the cryptographic algorithm identification scheme based on the hybrid random forest and logistic regression model will be explained in detail.

Definition 1 *(Ciphertext)* Set up a collection of cryptographic algorithms as in (1).

$$A = \{a_1, a_2, ..., a_n\}. \tag{1}$$

Where n is the number of cryptographic algorithms. For any given cryptographic algorithm a_i, there is a ciphertext file c_j generated by encrypting plaintext in *mod* mode (2).

$$c_j = \{b_1, b_2, ..., b_s\}. \tag{2}$$

Where b_i is the i-th character of the ciphertext file and *mod* denotes a certain mode of operation.

Definition 2 *(Ciphertext features)* Extract features from the ciphertext file c_j, and obtain a feature set with dimension d.

$$fea = \{x_1, x_2, ..., x_d\}. \tag{3}$$

The ciphertext extraction process can be expressed as a process of mapping the ciphertext file c_j into a feature set fea. $Extr(c_j) \rightarrow fea$, where *Extr* represents the calculation method of feature extraction, also called processing function. In the cryptographic algorithm identification scheme, the ciphertext feature of ciphertext c_j can be expressed as a three tuple as in (4)

$$Fea = (C, Extr, d). \tag{4}$$

Where *C* represents a ciphertext data set composed of *n* ciphertext files, $C = \{c_1, c_2, ..., c_n\}$, *Extr* represents a processing function that maps *C* to ciphertext features, and *d* represents the dimension of ciphertext features.

Definition 3 *(Cryptographic algorithms identification)* For the cryptographic algorithms set *A* and the ciphertext set *C*, there is an identification scheme *I*. In the ciphertext only scenario, the cryptographic algorithm *A* to which the ciphertext set *C* belongs is identified with an accuracy of P^I. This process is called cryptographic algorithm identification. As a three tuple in (5)

$$\delta = (A, I, P^I). \tag{5}$$

Definition 4 *(Cryptographic algorithms identification scheme)* In the identification of cryptographic algorithms, suppose that *SLRP* is the workflow for directly identifying specific cryptographic algorithms, *fea* is the ciphertext feature extracted from ciphertext *C*, and *CA* is the classification algorithm used for identification, then the cryptographic algorithm identification scheme can be described as a three tuple in (6)

$$O = (SLRP, fea, CA). \tag{6}$$

Figure 1 is the working flow chart of cryptographic algorithm identification.

Identification scheme is one of the main research contents of cryptographic algorithm identification. The ciphertext feature fea in Definition 4 is expressed as a triplet in Definition 3. The content in the next section will further describe the triples based on this definition.

3 Ciphertext Feature Extraction Method Based on Randomness Testing

Randomness testing of the output sequence is an important means to evaluate the security of any cryptographic algorithm. The randomness testing usually uses probability statistics to check whether the detected sequence satisfies certain features of the random sequence (such as periodicity, correlation, and distribution features) to determine whether it is random [18].

3.1 Theoretical Basis of Random Detection

When determining the randomness of the ciphertext generated by the encryption algorithm encryption, relevant researchers usually use hypothesis testing to check whether the ciphertext is truly random [19]. When we examine the randomness of the output binary sequence of the cryptographic algorithm, we should first put forward a null hypothesis to be tested, denoted as H_0. Correspondingly, the hypothesis opposite to the original hypothesis is called the alternative hypothesis, which is denoted as H_1. For example, in the randomness testing,

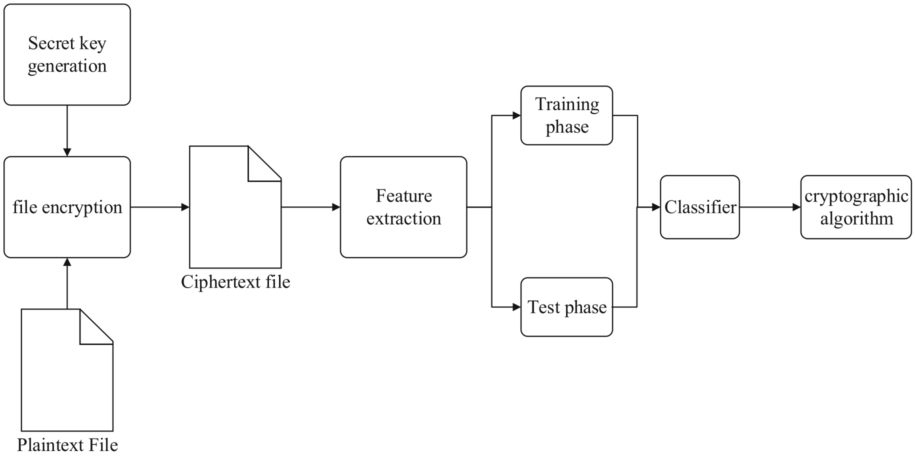

Fig. 1 Flow chart of cryptographic algorithm identification

the null hypothesis H_0 is generally set as: the sequence is random; the alternative hypothesis H_1: the sequence is non-random. If the result of the test supports the null hypothesis, the sequence is considered random, otherwise the sequence is not considered random. The specific inspection steps are as follows:

Steps of hypothesis testing

a. Set the null hypothesis H_0 and alternative hypothesis H_1;.
b. Construct a suitable sample statistic X according to H_0 and determine its distribution.
c. According to the pre-set significance level α, find the critical value in the quantile table of the corresponding distribution of the statistics, and give the rejection domain.
d. Compare the value of X calculated from the sample with the critical value in the third step. If the value of X falls into the rejection domain, reject H_0, otherwise accept H_0.

Table 1 shows the two error types and probabilities that may occur in hypothesis testing.

In practical applications, the P-value method is often used to determine the validity of the null hypothesis. We believe that the sample statistic X obeys the chi-square (χ^2) distribution, and the probability density curve of the distribution is as shown in Fig. 2. First calculate the statistic X, then find the integral from X to positive infinity, and finally compare the integral result P with α. If the P-value is 1, it represents a completely random sequence. If the P-value is 0, it represents a completely non-random sequence. Therefore, if $P \geq \alpha$, H_0 is accepted, otherwise H_0 is rejected. Usually the value range of α is [0.001,0.01], as shown in Fig. 2.

3.2 Ciphertext Feature Extraction Method

The SP 800-22 standard developed by the National Institute of Standards and Technology (NIST) has a wide range of coverage for the detection of sequence randomness, and is aimed

Table 1 Types of hypothesis testing errors

H_0	Calculation result of sample statistic X	Probability of error
True	Accept the null hypothesis	0
True	Reject the null hypothesis	α (First type error)
False	Accept the null hypothesis	β (Second type error)
False	Reject the null hypothesis	0

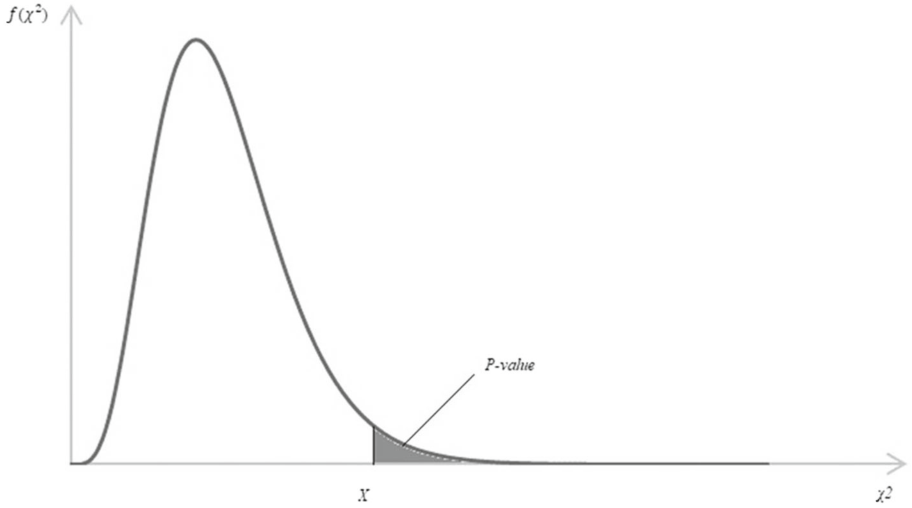

Fig. 2 The probability density curve of the chi-square distribution and its P-value

at the global or partial randomness of the sequence testing[20]. In accordance with existing research and NIST randomness testing requirements, the triplet $fea = (C, Extr, d)$ in Definition 2 is extended to Definition 5.

Definition 5 *(Extract ciphertext features based on randomness testing)* The ciphertext feature can be expressed as a four-tuple in (7)

$$fea = (C, orga, NIST, d). \tag{7}$$

$$NIST = \{nist_1, nist_1, ..., nist_{15}\}. \tag{8}$$

NIST is 15 different randomness testing programs. The $orga$ is the ciphertext data organization form required by $nist_i$,$i = (1, 2, ..., 15)$.

In this section, improvements and parameter adjustments are made on the basis of the open source tool sp 800_22_tests-master[21] written by python to achieve ciphertext feature extraction. The main program is responsible for reading the ciphertext file from the disk, and each randomness testing is encapsulated into a sub-module. The sub-modules are independent of each other and do not interfere with each other. Each module is executed in parallel when features are extracted. Based on the data set in this article, the above 10 meaningful random detections are selected to carry out ciphertext feature extraction, and 10 sets of return values are obtained as the classification basis for cryptographic algorithm identification where for random detections that generates more than one return value, the smallest value in the list

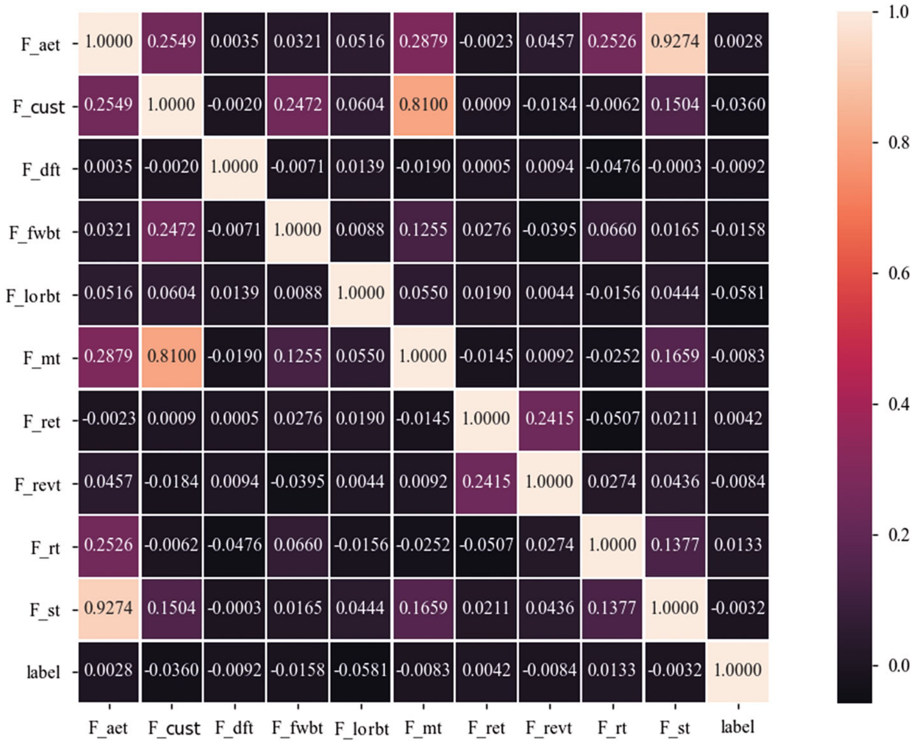

Fig. 3 Heat map of the relationship between ciphertext features

of return values is taken. Figure 3 is a heat map of the relationship between the ciphertext features extracted by random detection. It can be seen from the heat map that the correlation between the ciphertext features extracted by random detection is low and the features are independent of each other.

4 Block Cipher Algorithm Identification Scheme Based on Hybrid Random Forest and Logistic Regression Model

Random forest algorithm [22, 23] is a simple and effective ensemble learning classification algorithm. It is one of the most commonly used classification algorithms in the field of machine learning with high classification accuracy. The basic unit of random forest is a decision tree. When working, the classifier constructs multiple decision trees and integrates them to obtain the best decision result. For tree learning, ensemble learning mainly uses bootstrap, aggregating or bagging three methods [24]. When the sample to be classified is inputted, the classification result of the random forest output is determined by the majority of the classification results of each decision tree. Each decision tree is parallel and does not need pruning, so its speed is no less than that of ordinary CART decision trees, so it is suitable for the classification task in the identification of cryptographic algorithms.

This section carries out the ciphertext classification task based on the features extracted in the third section to verify the effectiveness of the ciphertext data classification task based

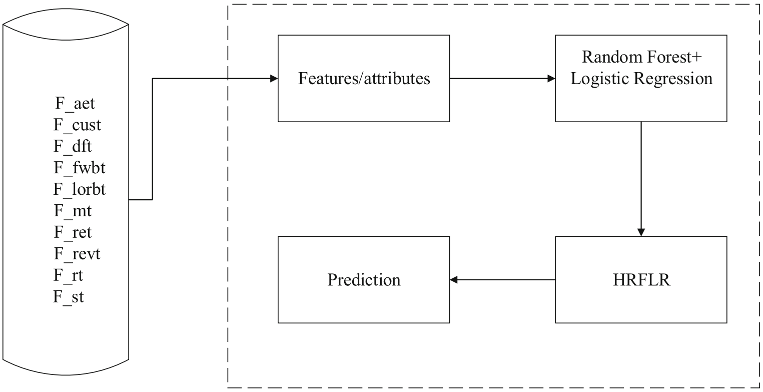

Fig. 4 Cryptographic algorithm identification based on HRFLR algorithm

on random features. We take five kinds of block ciphers, including AES, 3DES, Blowfish, CAST and RC2, as the research object of cryptographic algorithm identification. At the level of method composition, it can be divided into two main parts: ciphertext feature extraction and the construction of cryptographic algorithm identification classifiers. This paper uses ensemble learning to construct a classifier, and uses all kinds of ciphertext feature data extracted as the input of the classifier. After the classifier is trained and tested, the identification of the cryptographic algorithm is realized.

4.1 Hybrid Random Forest and Logistic Regression Classification Model

For the combination of stacking model, besides the advantages of traditional integration, it also has its own unique advantages. According to the principle of statistics, a variety of base learners can explore feature space and fit different features from different angles because of different training principles, which makes the learned features more comprehensive and makes full use of the differences between algorithms. In addition, stacking model uses the learner to conduct secondary generalization learning on the basic training results, which can reduce the effect of bias and variance, and reduce the impact of data interference.

Compared with the traditional classification algorithm model, although the random forest model greatly improves the classification accuracy and has strong stability, this method is prone to overfitting on some sample sets with relatively large noise. In addition, features with more value divisions are likely to have a greater impact on RF decision-making, thereby affecting the effect of the fitted model. Based on this problem, we proposed the HRFLR model.

Figure 4 shows the prediction method of HRFLR. The main idea is the superposition method in ensemble learning, which distributes multiple base classifiers on multiple levels, and the prediction results of the first layer of classifiers are used as the input of the next layer, and so on, achieving integration through multi-layer training, thus obtaining the final classification prediction result [25].

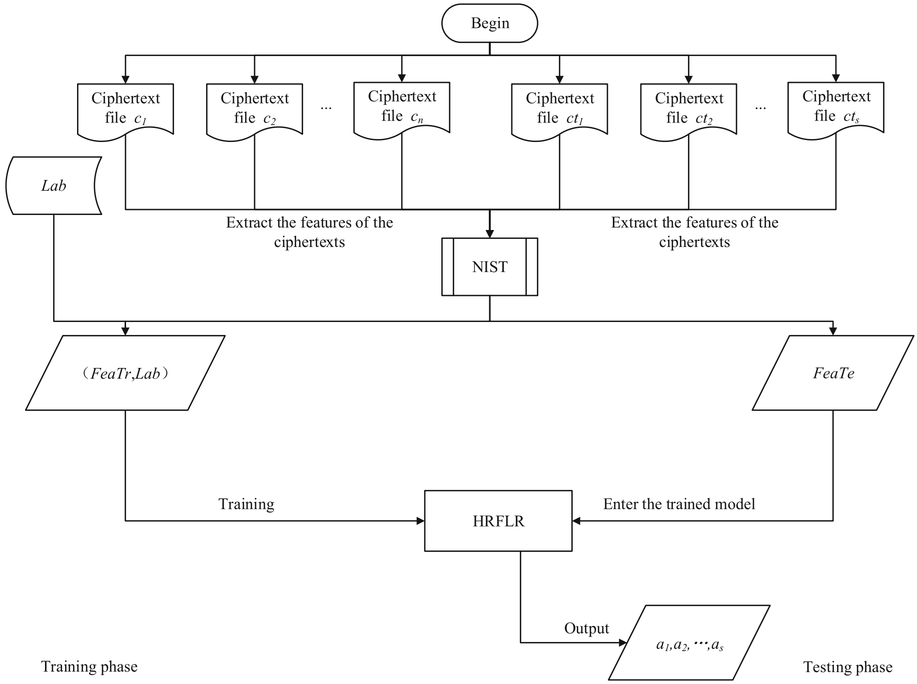

Fig. 5 Flow chart of block cipher algorithm identification scheme based on HRFLR

Algorithm 1 HRFLR algorithm.

```
Require:
    $DTr$ - Training data set with corresponding label of sample;
    $DTe$ - Testing data set without corresponding label of sample.
Ensure:
    Classification labels corresponding to samples in $DTe$ dataset $C(DTe) = \{c(dte_1), c(dte_2), ..., c(dte_d)\}$
    Where $d$ is the number of testing data set.
1: A set of decision tree base classifiers $\xi_{dt} = \{t_1, t_2, ..., t_k\}$ is constructed according to the training set $D$. By voting the prediction of each base classifier, the random forest model $\xi_{rf}$ was obtained, and then the random forest model is used to predict the testing set, and the prediction results $P_{rf} = \{p_1, p_2, ..., p_d\}$ are taken as the new features of the testing set;
2: One-hot coding for the testing set with the new feature;
3: Logical regression model $\xi_{lr}$ is used to classify the testing set.
4: return $C(DTe) = \{c(dte_1), c(dte_2), ..., c(dte_d)\}$.
```

We propose the HRFLR algorithm for cryptographic algorithm identification shown in algorithm 1. In our HRFLR algorithm, firstly, use the original function to train the random forest model, then use the tree constructed by the random forest model to learn the new features, then make a hot coding new features, then add the new coding features to the original features as input to train the logistic regression model, and finally output the predicted results. The main idea of Stacking is to train the model to learn the predicted results using the underlying learner. The HRFLR model integrates different algorithms, makes full use of different algorithms to observe different data from different data space and data structure perspectives, learns from each other and optimizes the results. Therefore, the integrated model has excellent performance, much better than the traditional machine learning model.

In addition, the HRFLR model adopts logistic regression model in the second layer, which can effectively resist the problem of over-fitting in the process of training and then affect the final results of the model.

4.2 Block Cipher Algorithm Identification Scheme Based on HRFLR

The adopted cryptographic algorithm identification scheme is described as in (9) according to Definition 4

$$O = (SLRP, fea, HRFLR). \tag{9}$$

where $SLRP$ is the workflow of the identification scheme. The fea is the ciphertext features extracted from ciphertext data, see definition 5 for details. $HRFLR$ is the hybrid random forest and logistic regression classification model proposed in the previous section.

The program flow chart is shown in Fig. 5. The program is mainly composed of training phase and testing phase.

Training phase

Input: A set of ciphertext files with labels $C = \{c_1, c_2, \ldots, c_n\}$.

a. According to the randomness testing in Sect. 3, extract features from the ciphertext files C and obtain a set of feature sets $FeaTr = \{feaTr_i^j | i = 1, 2, \ldots, n, j = 1, 2, \ldots, d\}$ where $feaTr_i$ is the eigenvector of the ith ciphertext, and d is the feature dimension.
b. Take the cryptographic algorithm of n ciphertext files as labels, and record it as an n-dimensional set $Lab = \{lab_1, lab_2, \ldots, lab_n\}$, and then record the two-tuple composed of feature set $FeaTr$ and label set Lab as the original data set T.
c. Input feature set T, each ciphertext file represents a sample, there are n samples in total, each sample has d features, and the integer k represents the number of trees in the random forest.
d. Sampling with replacement is adopted, and M samples are randomly selected from T to form a Bootstrap sample set T^* as the sample at the root node of the decision tree.
e. Randomly select t attributes from d features as candidate attributes on the T^* data set, and calculate the best split attribute.
f. Based on each value type of the best split attribute, split the current data set horizontally to obtain $P_1, P_2, \ldots, P_m$.
g. For each sub-data set in $P_1, P_2, \ldots, P_m$, randomly select t attributes, and select the best split attribute from them, and split the current data set horizontally here.
h. Repeat step 2 to step 5 to build k decision trees and construct a random forest to obtain a regular classification data set $D(R(P_1), R(P_2)......R(P_m))$.
i. Use the learned tree for feature extraction to construct a new feature $F(P_1, P_2, \ldots, P_m)$.
j. Perform one-hot encoding on the new feature and normalize it.
k. Apply a logistic regression classifier to the extracted features for classification training.

Output: The trained ensemble classifier HRFLR and classification results.

Table 2 List of specific parameters of 5 block cipher algorithms

Mark	Structure	Key	Modes of operation	Parameter scale	Realization method
AES	SP	Selected	ECB	Fixed parameter	Crypto
3DES	Feistel	Selected	ECB	Fixed parameter	Crypto
Blowfish	Feistel	Selected	ECB	Fixed parameter	Crypto
CAST	Feistel	Selected	ECB	Fixed parameter	Crypto
RC2	Feistel	Selected	ECB	Fixed parameter	Crypto

Testing phase

Input: A set of ciphertext files without labels $CT = \{ct_1, ct_2, \ldots, ct_s\}$.

a. Perform feature extraction on the content of the ciphertext file $CT = \{ct_1, ct_2, \ldots, ct_s\}$ to be identified to obtain the ciphertext feature $FeaTe = \{feaTe_i^j | i = 1, 2, \ldots, s, j = 1, 2, \ldots, d\}$.

b. Input the ciphertext feature $FeaTe = \{feaTe_i^j | i = 1, 2, \ldots, s, j = 1, 2, \ldots, d\}$ into the trained classification model.

Output: The cryptographic algorithms label a_1, a_2,..., a_s corresponding to the ciphertext to be tested.

5 Experimental Environment

5.1 Data Preparation

Table 2 shows five cryptographic algorithms for ciphertext data collection. We choose to encrypt the ciphertexts by the Crypto algorithm library of python. The plaintext used in the experiment is random data generated by Python's Crypto cryptographic module using the Fortuna Accumulator method[26]. The plaintext includes a total of 500 files with sizes of 1 KB, 8 KB, 64 KB, 256 KB and 512 KB. The key is a fixed secret key generated by the Cipher module of Crypto, and the ciphertext is generated by the ECB mode of the five block cipher algorithms AES, 3DES, Blowfish, CAST and RC2. During the experiment, the ciphertext samples encrypted by each cipher algorithm are 500 copies and only full rounds of each cipher algorithm are considered. Use the ciphertext feature extraction method introduced in this article to calculate the feature values of all ciphertexts, and each ciphertext sample corresponds to a set of features, and save these values. We conducted repeated random subsampling verification on the number of experiments[27], in which 75% of random sampling was used as the training set and the remaining 25% as the test set, and conducted binary classification and multi-classification experiments on the five cryptographic algorithms.

5.2 Evaluation Criteria for Classification Results

In classification problems, the commonly used evaluation methods include accuracy, precision, recall, F1, ROC, AUC, cost-sensitive error rate and cost curve, etc. The confusion matrix is used to evaluate the model. The confusion matrix produces four results namely

TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative).The following indicators are used to calculate accuracy, precision, and recall.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{10}$$

$$Precision = \frac{TP}{TP + FP}. \tag{11}$$

$$Recall = \frac{TP}{TP + FN}. \tag{12}$$

In order to balance the precision and recall, we added F_1 score to evaluate the performance of the classifier as in 13

$$F_1 = \frac{2TP}{2TP + FP + PN}. \tag{13}$$

In the identification tasks of cryptographic algorithms we studied, we pay more attention to the accuracy of the classification of all cryptographic algorithms. Therefore, we use the accuracy and precision as the standard to evaluate the performance of the classifier.

6 Evaluation and Comparison of Experimental Results

Based on the 10 features extracted by the feature extraction method mentioned above, the prediction model is established and the accuracy of the model is calculated. The eight classification algorithms and their classification results are shown in Tables 3 and 4, where accuracy, precision, recall rate and F_1 scores are compared. The results show that the HRFLR algorithm has the highest accuracy compared with the existing classification algorithms.

6.1 Binary Classifications Identification of Cryptographic Algorithms

The classification algorithms of SVM, Gaussian Naive Bayes (GNB), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), K-NearestNeighbor (KNN), AdaBoost and HRFLR are used to perform binary classification on ciphertext files of different sizes encrypted by AES and 3DES. The results are shown in Table 3.

The first column of Table 3 shows the evaluation index of the identification results of the cryptographic algorithm. The second column is the size of the ciphertext file. It can be seen from the table that the average accuracy of SVM, NB, RF, LR, DT, KNN, Adaboost and HRFLR classifiers of different ciphertext file sizes are 0.552, 0.556, 0.536, 0.516, 0.552, 0.492, 0.512 and 0.712. In addition, the identification accuracy of the HRFLR model is affected by the size of the ciphertext. And it is as high as 0.740 for 512 KB ciphertext files and no less than 0.700 for other ciphertext files.

By comparing the F1 scores of the eight classification models, it can be found that the classification scheme constructed based on HRFLR has the highest F1 scores for all five ciphertext size cases, indicating that the HRFLR classifier has the best classification performance.

In Fig. 6, figures (a) to (e) show the receiver operating characteristic curve (ROC) of eight classification models for AES and 3DES cryptographic algorithm identification on 1 KB, 8 KB, 64 KB, 256 KB and 512 KB files. As shown in the figure, in the binary classification, due to the different size of the ciphertext file, the performance of different classification models is different.

Table 3 The results of binary classification identification based on eight classification algorithms

Evaluating Indicator	File size	Classifiers							
		SVM	GNB	RF	LR	DT	KNN	AdaBoost	HRFLR
Accuracy	512 KB	0.580	0.600	0.540	0.540	0.500	0.440	0.540	0.740
	256 KB	0.580	0.620	0.540	0.540	0.540	0.500	0.500	0.700
	64 KB	0.660	0.600	0.560	0.600	0.500	0.540	0.460	0.700
	8 KB	0.520	0.520	0.480	0.500	0.620	0.520	0.580	0.720
	1 KB	0.420	0.440	0.560	0.400	0.600	0.460	0.480	0.700
Precision	512 KB	0.599	0.626	0.524	0.543	0.500	0.433	0.540	0.750
	256 KB	0.581	0.638	0.543	0.543	0.560	0.500	0.500	0.701
	64 KB	0.725	0.615	0.576	0.625	0.537	0.552	0.502	0.733
	8 KB	0.533	0.539	0.505	0.523	0.637	0.514	0.589	0.744
	1 KB	0.420	0.440	0.562	0.399	0.601	0. 458	0.480	0.703
Recall	512 KB	0.580	0.600	0.540	0.540	0.500	0.440	0.540	0.740
	256 KB	0.580	0.620	0.540	0.540	0.540	0.500	0.500	0.700
	64 KB	0.660	0.600	0.560	0.600	0.500	0.540	0.460	0.700
	8 KB	0.520	0.520	0.480	0.500	0.620	0.520	0.580	0.720
	1 KB	0.420	0.440	0.560	0.400	0.600	0.460	0.480	0.700
F_1-score	512 KB	0.580	0.600	0.540	0.540	0.500	0.440	0.540	0.740
	256 KB	0.580	0.620	0.540	0.540	0.540	0.500	0.500	0.700
	64 KB	0.660	0.600	0.560	0.600	0.500	0.540	0.460	0.700
	8 KB	0.520	0.520	0.480	0.500	0.620	0.520	0.580	0.720
	1 KB	0.420	0.440	0.560	0.400	0.600	0.460	0.480	0.700

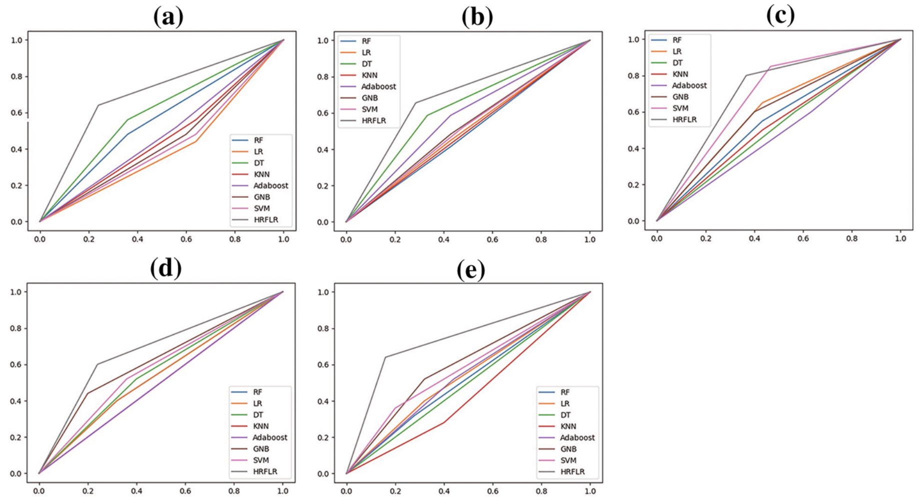

Fig. 6 ROC curves of eight classification models in AES and 3DES cryptographic algorithms identification

In the cryptographic classification task of the different ciphertext file size, the ROC curve of HRFLR model is significantly close to the upper left corner, and the Area Under Characteristic (AUC) value of HRFLR model is the highest, which indicates that HRFLR classification model has the best classification effect.

Figure 7 shows the binary classification accuracy of eight classifiers for ciphertext files with a size of 1 KB to 512 KB. It can be seen from the figure that the classification accuracy of the classic classification algorithm is unstable and fluctuates between 0.4 and 0.66. The HRFLR algorithm we proposed is significantly higher than the other seven, with a stable accuracy above 0.7 .

In addition, compared with the other seven classification models, the HRFLR classification model has the least fluctuation, and its classification accuracy is basically not affected by the size of ciphertext documents, which means that the HRFLR classification model has the strongest stability.

6.2 Multi-classification of Cryptographic Algorithms

In this section, eight classification algorithms are used to classify ciphertext files encrypted by AES, 3DES, Blowfish, CAST and RC2 cryptographic algorithms based on the ten-fold repeated random subsampling verification. The experimental results are shown in Table 4.

As can be seen from table 5, for the multi-classification tasks of AES, 3DES, Blowfish, CAST and RC2, the highest identification rate of HRFLR can reach 38% and the lowest is not less than 30%, which are higher than the identification accuracy of the other seven models and significantly higher than the random classification accuracy of 20%. In addition, the F1 score has a similar pattern to the identification accuracy, indicating that the HRFLR classifier has the best classification results in the majority of cases.

The Fig. 8 shows the multi-classification accuracy of eight classifiers with ciphertext file size from 1 KB to 512 KB. As can be seen from the figure, the classification accuracy of the cryptographic algorithm identification scheme based on a single classification algorithm is low, fluctuating around 0.2, and the lowest is the accuracy rate of AdaBoost classifier on

Table 4 The result of multi-classification identification based on eight classification algorithms

Evaluating Indicator	File size	Classifiers							
		SVM	GNB	RF	LR	DT	KNN	AdaBoost	HRFLR
Accuracy	512 KB	0.310	0.170	0.240	0.190	0.200	0.170	0.210	0.328
	256 KB	0.184	0.136	0.192	0.168	0.224	0.216	0.216	0.300
	64 KB	0.152	0.176	0.168	0.160	0.224	0.240	0.120	0.310
	8 KB	0.176	0.208	0.224	0.184	0.200	0.192	0.152	0.340
	1 KB	0.176	0.264	0.288	0.256	0.256	0.192	0.256	0.380
Precision	512 KB	0.315	0.176	0.265	0.199	0.257	0.196	0.200	0.350
	256 KB	0.184	0.134	0.208	0.190	0.251	0.210	0.247	0.348
	64 KB	0.144	0.193	0.175	0.147	0.280	0.264	0.134	0.294
	8 KB	0.200	0.244	0.286	0.240	0.216	0.262	0.212	0.339
	1 KB	0.139	0.249	0.281	0.211	0.289	0.162	0.301	0.358
Recall	512 KB	0.310	0.170	0.240	0.190	0.200	0.170	0.210	0.328
	256 KB	0.184	0.136	0.192	0.168	0.224	0.216	0.216	0.300
	64 KB	0.152	0.176	0.168	0.160	0.224	0.240	0.120	0.310
	8 KB	0.176	0.208	0.224	0.184	0.200	0.192	0.152	0.340
	1 KB	0.176	0.264	0.288	0.256	0.256	0.192	0.256	0.380
F_1-score	512 KB	0.310	0.170	0.240	0.190	0.200	0.170	0.210	0.328
	256 KB	0.184	0.136	0.192	0.168	0.224	0.216	0.216	0.300
	64 KB	0.152	0.176	0.168	0.160	0.224	0.240	0.120	0.310
	8 KB	0.176	0.208	0.224	0.184	0.200	0.192	0.152	0.340
	1 KB	0.176	0.264	0.288	0.256	0.256	0.192	0.256	0.380

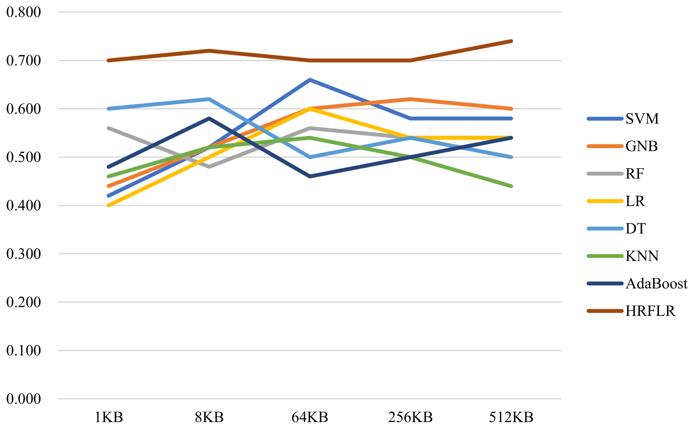

Fig. 7 The binary classification accuracy of the eight classifiers in ciphertext files with sizes from 1 KB to 512 KB

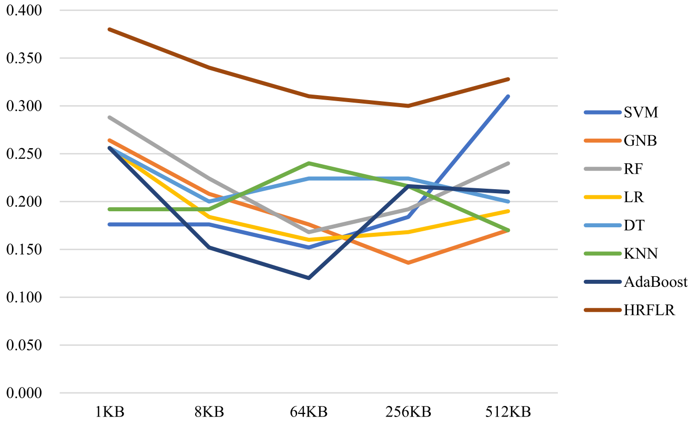

Fig. 8 The multi-classifications accuracy of the eight classifiers with ciphertext file size from 1 KB to 512 KB

64 KB ciphertext file, which is 0.12; The highest identification accuracy is 0.31 for SVM classifier on 512 KB. The identification accuracy of the cipher algorithm based on HRFLR algorithm is significantly higher than that of the other seven classical algorithms, which is more than 0.3. In addition, the identification accuracy of eight models are affected by the ciphertext size, the HRFLR model has the least fluctuation and the highest model stability.

7 Conclusions

According to the features of random forest and logistic regression, this paper proposes a hybrid random forest and logistic regression (HRFLR) model based on the idea of ensemble learning. The model integrates several weakly supervised models in order to obtain a more comprehensive strong supervised model and achieve better classification accuracy. On this basis, a block cipher algorithm identification scheme based on the idea of ensemble learning is proposed. Under the condition that the ciphertext is known, the five typical block ciphers, AES, 3DES, Blowfish, CAST and RC2, are used as the identification objects, and the ciphertext features are extracted by the randomness test method, and the features are used as the classification basis for the subsequent identification tasks for cryptographic algorithm identification. The experimental results show that HRFLR algorithm not only has higher accuracy than single random forest and logistic regression algorithm but also higher accuracy than random guess when dealing with binary classification and multiple classification problems. As a new idea, block cipher algorithm identification scheme based on ensemble learning is worthy of further exploration, it has certain positive significance for future research in cipher algorithm identification.

Acknowledgements This work was supported by the National Key R&D Program of China under Grant 2018YFA0704703; the National Natural Science Foundation of China under Grant 61802111; the Key Research and Promotion Projects of Henan Province under Grant 222102210062; the Basic Research Plan of Key Scientific Research Projects in Colleges and Universities of Henan Province under Grant 22A413004; the National Innovation Training Program for College Students under Grant 202110475072.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Dileep AD, Sekhar CC (2006) Identification of block ciphers using support vector machines. In: The 2006 IEEE International Joint Conference on Neural Network Proceedings, pages 2696–2701. IEEE. https://doi.org/10.1109/IJCNN.2006.247172
2. Manjula R, Anitha R (2011) Identification of encryption algorithm using decision tree. In: Communications in Computer and Information Science, volume 133, pages 237–246. Springer. https://doi.org/10.1007/978-3-642-17881-8_23
3. Chou JW, Lin SD, Cheng CM (2012) On the effectiveness of using state-of-the-art machine learning techniques to launch cryptographic distinguishing attacks. In: Acm Workshop on Security and Artificial Intelligence, pages 105–110. https://doi.org/10.1145/2381896.2381912
4. Sharif SO, Kuncheva LI, Mansoor SP (2010) Classifying encryption algorithms using pattern recognition techniques. In: 2010 IEEE International Conference on Information Theory and Information Security, pages 1168–1172, https://doi.org/10.1109/ICITIS.2010.5689769
5. Sun X, Zhang P, Liu JK, Jianping Yu, Xie W (2020) Private machine learning classification based on fully homomorphic encryption. IEEE Trans Emerg Top Comput 8(2):352–364. https://doi.org/10.1109/TETC.2018.2794611
6. Li J, Kuang X, Lin S, Ma X, Tang Y (2020) Privacy preservation for machine learning training and classification based on homomorphic encryption schemes. Inf Sci 526:166–179. https://doi.org/10.1016/j.ins.2020.03.041
7. Xiao C, Han D, Ma Y, Qin Z (2019) Csigan: Robust channel state information-based activity recognition with gans. IEEE Internet Things J 6(6):10191–10204. https://doi.org/10.1109/jiot.2019.2936580
8. Cheng L, Jiang F, Wang Z, Li J (2021) Multiconstrained real-time entry guidance using deep neural networks. IEEE Trans Aerosp Electron Syst 57(1):325–340. https://doi.org/10.1109/TAES.2020.3015321
9. Mishra S, Bhattacharjya A (2013) Pattern analysis of cipher text: A combined approach. In: 2013 International Conference on Recent Trends in Information Technology (ICRTIT), pages 393–398. https://doi.org/10.1109/ICRTIT.2013.6844236
10. De Souza WAR, Tomlinson A (2013) A distinguishing attack with a neural network. In: 2013 IEEE 13th International Conference on Data Mining Workshops, pages 154–161. IEEE. https://doi.org/10.1109/ICDMW.2013.116
11. Yang W, Tao W, Jindong L (2015) Research on a new method of statistical detection of block cipher algorithm ciphertext. Journal of Ordnance Engineering College 000(003):58–64. https://doi.org/10.3969/j.issn.1008-2956.2015.03.011

12. Liangtao H, Zhicheng Z, Yaqun Z (2018) Hierarchical recognition scheme of cryptosystem based on random forest. Journal of Computer 41(002):382–399. https://doi.org/10.11897/SP.J.1016.2018.00382
13. Zhicheng Z, Yaqun Z, Fengmei L (2019) Recognition scheme of block cipher system based on randomness test. Journal of Cryptography 6(2):177–190. https://doi.org/10.13868/j.cnki.jcr.000293
14. Arvind Ratan R (2020) Identifying traffic of same keys in cryptographic communications using fuzzy decision criteria and bit-plane measures. International Journal of System Assurance Engineering and Management, 11(2):466–480. https://doi.org/10.1007/s13198-019-00878-7
15. Baccour L (2018) Amended fused topsis-vikor for classification (atovic) applied to some uci data sets. Expert Syst Appl 99:115–125. https://doi.org/10.1016/j.eswa.2018.01.025
16. Esfahani HA, Ghazanfari M (2017) Cardiovascular disease detection using a new ensemble classifier. In: 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), pages 1011–1014. IEEE. https://doi.org/10.1109/KBEI.2017.8324946
17. Amin MS, Chiam YK, Varathan KD (2019) Identification of significant features and data mining techniques in predicting heart disease. Telematics Inform 36:82–93. https://doi.org/10.1016/j.tele.2018.11.007
18. Yongqiang Z, Li Shunbo Q, Kailin SL, Chan L, Xiaoru X (2014) Nist randomness test method and application. Computer knowledge and technology 000(026):6064–6066
19. Shisong M, Jinglong W, Xiaolong P (2006) Higher mathematical statistics. Higher Education Press, Beijing
20. J Nechvatal E Barker S Leigh M Levenson D Banks A Heckert J Dray S Vo A Rukhin, J Soto. Statistical test suite for random and pseudorandom number generators for cryptographic applications, nist special publication. *National Institute of Standards and Technology*, 2010
21. David Johnston. sp800_22_tests. https://github.com/dj-on-github/sp800_22_tests, 2019
22. Liaw A, Wiener M, Liaw A (2002) Classification and regression with random forest. R News, 23(23)
23. Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88(11):2783–2792
24. Zhiyong S, Chong S, Yang Z, Zhiwei S (2018) A stochastic forest algorithm for unbalanced classification. Computer and modernization 280(12):60-64+70
25. David H (1992) Wolpert. Stacked generalization. Neural Netw 5(2):241–259. https://doi.org/10.1016/S0893-6080(05)80023-1
26. Ferguson N, Schneier B, Kohno T (2012) Cryptography Engineering: Design Principles and Practical Applications. Wiley Publishing, Newyork
27. Xizhi W (2013) Statistics: From data to conclusion (fourth edition). China Statistics, (6):2

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.