ATLANTIS PRESS International Journal of Networked and Distributed Computing, Vol. 6, No. 3 (July 2018) 155 - 173

One to One Identification of Cryptosystem Using Fisher's Discriminant Analysis

Xinyi Hu [1] , Yaqun Zhao [2]

[12] *State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China*

[1] *E-mail: huxinyi.1994@foxmail.com*

[2] *E-mail: yaqunzhao@163.com*

Abstract

Distinguishing analysis is an important part of cryptanalysis. It is an important content of discriminating analysis that how to identify ciphertext is encrypted by which cryptosystems when it knows only ciphertext. In this paper, Fisher's discriminant analysis (FDA), which is based on statistical method and machine learning, is used to identify 4 stream ciphers and 7 block ciphers one to one by extracting 9 different features. The results show that the accuracy rate of the FDA can reach 80% when identifying files that are encrypted by the stream cipher and the block cipher in ECB mode respectively, and files encrypted by the block cipher in ECB mode and CBC mode respectively. The average one to one identification accuracy rates of stream ciphers RC4, Grain, Sosemanuk are more than 55%. The maximum accuracy rate can reach 60% when identifying SMS4 from block ciphers in CBC mode one to one. The identification accuracy rate of entropy-based features is apparently higher than the probability-based features.

Keywords: Fisher's Discriminant Analysis, One to One Identification, Cryptosystem, Block Cipher, Stream Cipher, Feature Extraction.

1. Introduction

The main purpose of cryptanalysis is to study the deciphering of encrypted messages or the forgery of messages[1]. Specifically, it means using various methods to try to get all or part information of plaintext by ciphertext, under the condition of not knowing or not fully knowing the details of the decryption key and the cryptosystem adopted by the communicator[2]. Therefore, cryptosystem identification is an important part of cryptanalysis[3].

At present, the statistical method and the machine learning have been used to identify cryptosystem from existing ciphertext[4]. The principle of the cryptosystem identification scheme based on statistical methods is to design the identification index first, and then calculate the index value based on the extracted features, and finally identification result is determined by the size of the index value[5]. Based on machine learning, the identification scheme of cryptosystem is regarded the identification task of the cryptosystem as the pattern recognition task[3467]. And because of its simple design and stable results, it has attracted the attention of many researchers.

In 2006, Dileep et al.[8] proposed a identification scheme based on support vector machine(SVM) for AES, DES, 3DES, Blowfish and RC5. In 2011, Manjula et al.[9] proposed a cryptosystem identification scheme based on decision tree, it identified 11 kinds of cryptosystems including classical ciphers,

stream ciphers, block ciphers and public key ciphers. In 2005, Dunham et al.[6] proposed an identification scheme based neural network, it has an identification accuracy rate of 91.3% of different types of plaintexts. In 2010, Sharif et al.[7] used the method of pattern recognition to compare the 8 classification techniques for DES, IDEA, AES and RC2 in ECB mode. The results show that the Rotation Forest (RoFo) classifier has the highest classification accuracy.

Fisher's discriminant analysis (FDA) is a method based on the combination of statistics and machine learning. It has been used in many areas such as image processing and pattern recognition[12], machine intelligence[11], in the classification of speech or music[10]. But few researchers applied FDA into cryptanalysis, only Ray et al. in 2017[13] applied FDA in ciphertext classification. 5 kinds of stream ciphers and 5 kinds of block ciphers were classified by 3 kinds of extracted features. The plaintext file is the ASCII value corresponding to the English text. The classification result shows that it is superior to the random classification. The cryptosystems except the MARS algorithm is a common algorithm, the other 5 kinds of stream ciphers and 4 kinds of block ciphers are simplified to the existing algorithms without safety certification and randomness test, which is a distance away from identifying the practical cryptosystems.

In order to make up for this deficiency, this paper aims at the improvement of the FDA-based classification technique proposed in[13], and selects practical 18 algorithms which are 4 kinds of stream ciphers and 7 kinds of block ciphers in ECB mode and CBC mode. After analysis, 9 features completely different from the 3 features in Ref. 13 were extracted. According to the results of the features extraction, FDA technique was used to identify the cryptosystems one to one. The one to one identification accuracy rate of the stream cipher RC4 and Grain can reach more than 62%. The one to one identification accuracy rates of the block ciphers in the CBC mode can reach more than 58%, the ECB mode are about 70%. The one to one identification accuracy rates between the stream ciphers and the block ciphers in CBC mode are about 59%, and between the stream ciphers and block ciphers in ECB mode are up to 84%, between block ciphers in ECB mode and in CBC mode are more than 80%.

The remainder of the paper is organized as follows. Section 2 describes the principle of Fisher's discriminant analysis and extracts 9 new features; Section 3 establishes the FDA model according to the identification requirements; Section 4 is the experiments and the results; Section 5 is the tests of model; Section 6 is dedicated to some concluding remarks.

2. Prepare knowledge

2.1. Discriminant Analysis

Discriminant analysis is an effective method for multivariate data analysis, which can scientifically determine what type of sample belonging to. It can reveal the internal laws in the numerous data, and enable people to make a correct judgement of the research problems[14]. Discriminant analysis, which was produced in the 1930s, has been widely applied in many areas such as natural science, sociology and economic management in recent years. The point of discriminant analysis is summarized the regularity and principles of the classification based on the data information of several samples of the existing categories, and found the discriminant formulas and the discriminant criterions. The new unknown samples can be classified according to the discriminant formulas and discriminant criterions.

From the perspective of statistical data analysis, the model of discriminant analysis[14] is as follows: The k populations $G_1, G_2, ..., G_k$ have p variants data, the quantitative index is $X = (X_1, X_2, ..., X_p)^T$. Set the distribution function of G_i is $F_i(x) = F_i(x_1, x_2, ..., x_p), i = 1, 2, ..., k$, G_i usually is continuous population. So the probability density function of G_i is $f_i(x) = f_i(x_1, x_2, ..., x_p)$ (if G_i is discrete population, probability function will be used here). For any new sample data $x = (x_1, x_2, ..., x_p)^T$, we should determine which G_i it belongs to. The commonly used discriminant methods for discriminant analysis include distance discriminant, Bayes discriminant, stepwise discriminant and typical discriminant. The following is the Fisher's discrimi-

nant analysis in typical discriminant analysis.

2.2. Fisher's Discrimination of Two Populations

Fisher's discriminant is a discriminant method based on the idea of variance analysis, which can discriminate the different populations well and does not require the distribution of populations. Fisher's discriminant analysis can work properly as long as there are suitable digital feature vectors with different statistical distributions under different categories. Its basic idea is projection (dimensionality reduction), projecting the m data of k populations in one direction, so that the projection can be separated from the populations as much as possible [?].

Definition 1. [15] The covariance matrix between two populations G_1, G_2 is defined as

$$\begin{aligned} B &= n_1(\overline{x}_1-\overline{x})(\overline{x}_1-\overline{x})^T + n_2(\overline{x}_2-\overline{x})(\overline{x}_2-\overline{x})^T \\ &= \frac{n_1 n_2}{n_1+n_2}(\overline{x}_1-\overline{x}_2)(\overline{x}_1-\overline{x}_2)^T \end{aligned}$$

where n_1, n_2 are the number of samples in G_1, G_2, $\overline{x}_1, \overline{x}_2$ are the mean of G_1, G_2, and the total mean of G_1 and G_2 is

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1+n_2}.$$

Definition 2. [15] The deviation matrixes E_1, E_2 of two populations G_1, G_2 are defined as

$$E_j = \sum_{i=1}^{n_j}(x_j^{(i)}-\overline{x}_j)(x_j^{(i)}-\overline{x}_j)^T, j = 1,2,$$

the combination of two deviation matrixes is

$$E = E_1 + E_2,$$

where n_1, n_2 are the number of samples in G_1, G_2, $\overline{x}_1, \overline{x}_2$ are the mean of G_1, G_2.

Theorem 1. [15] For two populations G_1, G_2, n_1, n_2 are the number of samples in two populations, $\overline{x}_1, \overline{x}_2$ are the mean of the two populations, B is the covariance matrix between two populations, E is the combination of two deviation matrixes of two populations. Suppose that E is reversible ($n_1+n_2-2 \geqslant p$), from the properties of eigenvalues, we know that $E^{-1}B$ has unique nonzero eigenvalue λ. Let $\mathbf{a}$ be the corresponding eigenvector, it satisfies $(B-\lambda E)\mathbf{a} = \mathbf{0}$, then

$$\begin{aligned} &\frac{n_1 n_2}{n_1+n_2}(\overline{x}_1-\overline{x}_2)(\overline{x}_1-\overline{x}_2)^T\mathbf{a} \\ &= \frac{n_1 n_2}{n_1+n_2}(\overline{x}_1-\overline{x}_2)^T E^{-1}(\overline{x}_1-\overline{x}_2)E\mathbf{a} \end{aligned}$$

$S_p^{-1}(\overline{x}_1-\overline{x}_2)$ satisfies the above equation, where the joint covariance matrix of two populations is

$$S_p = \frac{1}{n_1+n_2-2}E.$$

Theorem 2. [15] For two populations G_1, G_2, the notations and the conditions are the same as Theorem 1, the Fisher's discriminant formula is

$$y = (\overline{x}_1-\overline{x}_2)^T S_p^{-1}x.$$

Discrimination criterions are

$$\begin{cases} x \in G_1 & \text{if } |y-\overline{y}_1| \leqslant |y-\overline{y}_2| \\ x \in G_2 & \text{if } |y-\overline{y}_1| \geqslant |y-\overline{y}_2| \end{cases}$$

where $y_1 = (\overline{x}_1-\overline{x}_2)^T S_p^{-1}\overline{x}_1$, $y_2 = (\overline{x}_1-\overline{x}_2)^T S_p^{-1}\overline{x}_2$.

2.3. Feature Extraction of Research Objectives

Feature extraction refers to the linear transformation or non-linear transformation of the original feature variables to obtain a smaller set of feature variables with better properties [16]. The main purpose of feature extraction is to reduce the overlapped parts of the original feature variables as much as possible to eliminate the possible correlation between features, so that the new features are more favorable for classification [17].

Due to the classification results in [4], RC2 of different lengths are classified to different categories, so the length-related features may play an important role in identifying the block ciphers. And there is no concept of block length in the stream ciphers (Or the block length in the stream cipher is 1). Thus attempt to divide the stream ciphers, then the length-related features may be important to identify the stream ciphers. In this paper, three different block length of

56bits, 128 bits and 192 bits are used for ciphertexts. According to the uncertainty and randomness of information, three kinds of features based on entropy and probability are designed as follow:

(1) The ciphertexts are divided into 3 different block lengths, and the entropy of a fixed bit in each block is calculated to form the same dimension as the block length. The form is

$$(f_1^n, f_2^n, ..., f_n^n),$$

where

$$f_i^n = -p_i^n \log p_i^n - (1-p_i^n)\log(1-p_i^n),$$

$$p_i^n = \frac{\sum_{j=1}^{\lfloor \frac{l_B}{n} \rfloor} b_{i+n(j-1)}}{\lfloor \frac{l_B}{n} \rfloor},$$

n is the block length, and b_i is the i-th bit in the ciphertexts bit string $(b_1, b_2, ..., b_{l_B})$, l_B is the length of ciphertext file bit string, and p_i is the probability that the i-th fixed bit value is $j-1$ in the block, $i = 1, 2, ..., n$.

(2) The ciphertexts are divided into 3 different block lengths, and the entropy of a fixed byte in each block is calculated to form the dimension as the 1/8 block length. The form is

$$(f_1^{\frac{n}{8}}, f_2^{\frac{n}{8}}, ..., f_{\frac{n}{8}}^{\frac{n}{8}}),$$

where

$$f_i^{\frac{n}{8}} = -\sum_{j=1}^{256} p_{i,j}^{n/8*256} \log p_{i,j}^{n/8*256},$$

n is the block length, $p_{i,j}^{n/8*256}$ is the probability that the i-th fixed byte value is $j-1$ in the block, $i = 1, 2, ..., \frac{n}{8}$.

(3) The ciphertexts are divided into 3 different block lengths, and the probability of all bytes in each block is calculated to form the same dimension as the block length. The form is

$$(f_1^n, f_2^n, ..., f_n^n),$$

where

$$f_i^n = \frac{N_i^n}{l_C},$$

n is the block length, N_i^n is the frequency that the value is $i-1$ in the block, and l_C is the length of the ciphertext file byte string $(c_1, c_2, ..., c_{l_C})$, $i = 1, 2, ..., n$.

The above 9 kinds of features were implemented in the VS2013 software, the notations are shown in Table 1.

Table 1. Features and symbols.

Block Length	Ideas	Features Mark	Dimensions
56bits	(1)	F_{56E}	56
56bits	(2)	$F_{56cut7E}$	7
56bits	(3)	F_{56P}	56
128bits	(1)	F_{128E}	128
128bits	(2)	$F_{128cut16E}$	16
128bits	(3)	F_{128P}	128
192bits	(1)	F_{192E}	192
192bits	(2)	$F_{192cut24E}$	24
192bits	(3)	F_{192P}	192

3. Establishment of System Model

This section proposes a scheme that identifying cryptosystem one to one based on the FDA. The main training and testing phases as follows.

Training phase:

(1) The ciphertext file $F_1, F_2, ..., F_n$ of the known categories were given, where n is the number of files.

(2) Extract the features of ciphertext, and obtain a set of features Fea=$\{f_1^d, f_2^d, ..., f_n^d\}$, where f_i^d is a feature vector of dimension d, $i = 1, 2, ..., n$.

(3) Label the 2 dimensional vector labels Lab=$\{l_1$ or $l_2\}$ of n ciphertext files of known categories, and obtain a set of tagged data (Fea, Lab).

(4) Put (Fea, Lab) into the FDA classifier and train the classification model.

Test phase:

(1) Extract the feature f^d of the ciphertext file F.

(2) Put f^d as an input of FDA classification model and get the result l_1 or l_2.

The flow-process diagram is shown in Fig. 1.

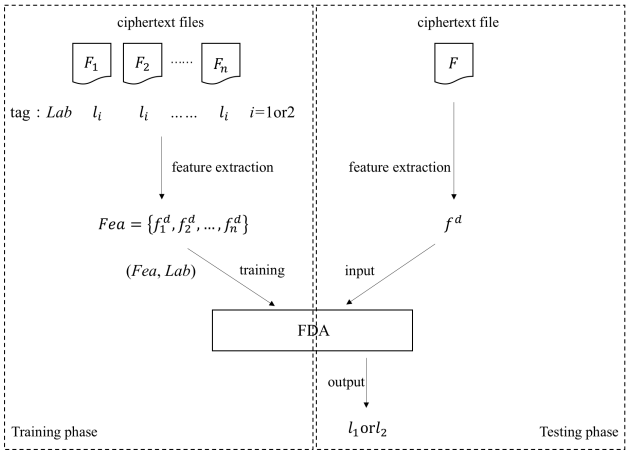


Fig. 1. Flow-process diagram of one to one identification scheme based on Fisher's discriminant analysis.

The following **Algorithm 1** is the algorithm of cryptosystem one to one identification based on FDA.

Algorithm 1: Cryptosystem one to one identification based on FDA

Input: Ciphertext file $F_1, F_2, ..., F_n$ of known classes, and the number of files n_1, n_2, where $F_i \in l_1$ or l_2.
Input: Pending ciphertext file F.
Output: Discriminant result l_1 or l_2.
1: Extract features $Fea \leftarrow f_1^d, f_2^d, ..., f_n^d$ from ciphertext files $F_1, F_2, ..., F_n$.
2: Use (Fea, Lab) to train the model, $x_1 \leftarrow (Fea, l_1)$.
3: The covariance matrix between two populations $B \leftarrow n_1(\overline{x}_1 - \overline{x})(\overline{x}_1 - \overline{x})^T + n_2(\overline{x}_2 - \overline{x})(\overline{x}_2 - \overline{x})^T$.
4: The combination of two deviation matrixes E.
5: The joint covariance matrix of two populations $S_p \leftarrow E/(n_1 + n_2 - 2)$.
6: Extract feature $f^d \leftarrow F$, $x \leftarrow (f^d, Lab)$.
7: Use (f^d, Lab) to test the model, the Fisher's discriminant formula $y \leftarrow (\overline{x}_1 - \overline{x}_2)^T S_p^{-1} x$
8: if $|y - \overline{y}_1| \leqslant |y - \overline{y}_2|$
return $x \in l_1$; or
return $x \in l_2$.

4. Experiments and Results

This section focuses on the identification model built on the above, and the identification algorithm designed on the above. The experiments were carried out as follow. The experimental environment is shown in Table 2.

Table 2. Experimental environment.

Host Model	MacBook Apple MNYN2CH/A
Processor	Inter(R) Core(TM) m5-6Y54 CPU @ 1.10GHz 1.20GHz
Memory	8.00GB
Operating System	Windows 10 Enterprise Edition 2015 (64bits)

18 kinds of cryptosystems are investigated in the experiment: 4 kinds of stream ciphers–RC4, Grain, Sosemanuk, Trivium, 7 kinds of block ciphers–AES-128, Blowfish, Camellia-128, DES, 3DES, IDEA, SMS4 in ECB mode and CBC mode. Algorithms and notations are shown in Table 3. In the data acquisition phase, we select images from the Caltech-256 dataset of California Institute of Technology[18], and made up 1000 files of 512KB size as the plaintext. 1000 files encrypted by the above 18 kinds of cryptosystems, and 1000 ciphertext files were obtained. Then the ciphertext files were truncated, the size of files was still 512KB. The total 18000 ciphertext files were obtained. The stream ciphers were implemented by Java platform program, and the block ciphers were implemented by open source tool OpenSSL.

Table 3. Algorithms and notations.

Algorithms	Notations
RC4	R
Grain	G
Trivium	T
Sosemanuk	S
AES	A
Blowfish	B
Camellia	C
DES	D
3DES	3
IDEA	I
SMS4	4

The accuracy rate of one to one identification was investigated under 9 different features. In the experiments, the ten-fold cross validation method

was implemented as follows:

(1) The data is divided into 10 equal partitions.
(2) Then 9/10 of the data is used for training and 1/10 for testing.
(3) The whole process is repeated 10 times, the overall error rate is equal to the average of error rates of each partition.

The accuracy rate of the 18 cryptosystems for one to one identification under 9 different features are shown in Fig. 2-Fig. 7, and the detailed results are shown in Appendix A. The five-fold cross validation method and twenty-fold cross validation method are also implemented. The results show that the difference of identification accuracy rate is not significant, and the detailed results are shown in Appendix B.

Fig. 2(a) shows that, the one to one identification accuracy rates in 4 kinds of stream ciphers are between 50%-60%. One to one identification accuracy rate of RC4, Grain and Sosemanuk can reach 55%. The accuracy rate of RC4 and Grain is 63%. The identification accuracy rate of RC4 and Sosemanuk is more than 60%. Fig. 2(b) shows that, 9 kinds of features have little differences in the identification rates of stream ciphers, and probability-based feature F_{128P} and F_{192P} are slightly better than other features, the accuracy rate is 60%.

Fig. 3(a) shows that, the identification accuracy rates in ECB mode are more than 60%, and accuracy rates of AES and other cryptosystems can reach 70%. The identification accuracy rates of Camellia with DES, 3DES, IDEA are more than 70%. Fig. 3(b) shows that, the identification accuracy rates of entropy-based features are obviously higher than the probability-based features, especially $F_{56cut7E}$ can reach about 70%, better than the others.

Fig. 4(a) shows that, the identification accuracy rates in CBC mode are between 50%-60%. The average identification accuracy rate of SMS4 with Blowfish, Camellia, DES, 3DES, IDEA, and Blowfish with DES are more than 57%, SMS4 with 3DES can reach 60%. Fig. 4(b) shows that, some of the identification accuracy rates in CBC mode can exceed 55% based on some features, which are close to 60% under F_{128P} and F_{192E}. The result is better than the existing results of CBC block ciphers identification.

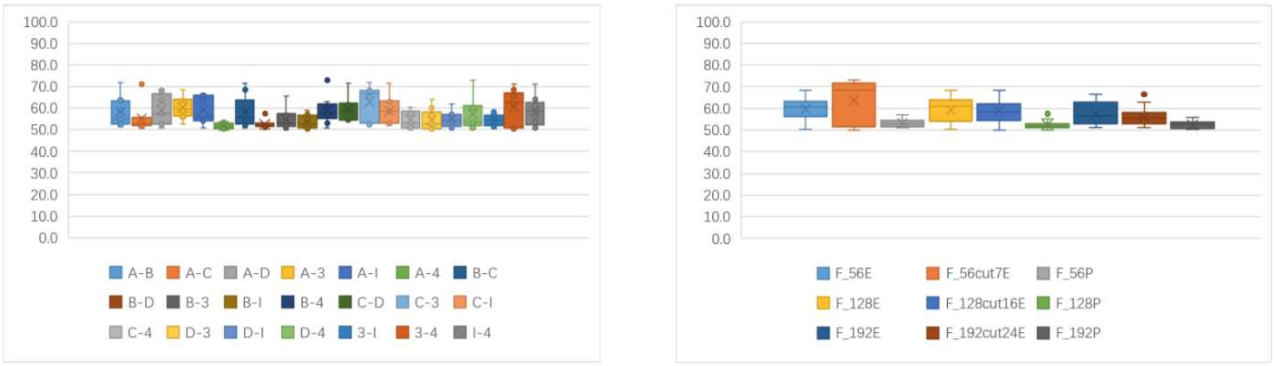
Fig. 5(a) shows that, the one to one identification of stream ciphers and block ciphers in ECB mode, the average accuracy rate is above 70% except Grain and DES. Fig. 5(b) shows that, the identification rates of entropy-based features are much higher than probability-based. F_{56E}, F_{128E} and F_{192E} can reach more than 80%, while the identification rate of probability-based F_{56P} is only about 55%. Therefore, when identifying the stream cipher and block cipher in ECB mode, the entropy-based feature can be extracted according to the requirement, and the high identification accuracy rate can be used to classify whether the ciphertext is encrypted by stream cipher or block cipher in ECB mode.

Fig. 6(a) shows that, the accuracy rates of stream ciphers and block ciphers in CBC mode are obviously lower than that of stream ciphers and block ciphers in ECB mode. The average identification rates are mostly between 50%-60%. The identification rate of Grain and AES is about 65%, Sosemanuk and Camellia can reach 58%. The identification accuracy rates of RC4 with 3DES, IDEA are more than 60%. The identification rates of DES with 4 kinds of stream ciphers are above 55%. Fig. 6(b) shows that, the 9 features have little differences in the identification rates of block ciphers in CBC mode. The entropy-based feature $F_{192cut24E}$ and the probability-based feature F_{192P} are slightly better than other features, and the partial identification rates can exceed 60%.

Fig. 7(a) shows that, the average identification rates of the same algorithm in different patterns are more than 70%, and the average identification accuracy rate of AES algorithm is the highest, more than 72%. Fig. 7(b) shows that, the identification rates of entropy-based features F_{56E}, F_{128E} and F_{192E} are over 80%, $F_{56cut7E}$, $F_{128cut16E}$ and $F_{192cut24E}$ are also close to 80%. The identification rates of probability-based feature F_{56P} is between 50%-60%, F_{128P} and F_{192P} are also about 60%. Thus the identification rates of probability-based features are obviously lower than the entropy-based features.

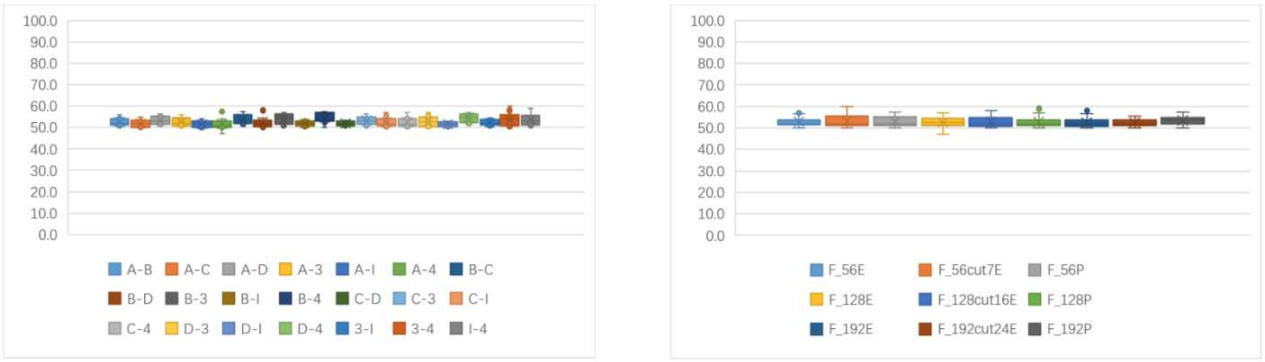
(a) (b)

Fig. 2. (a)One to one identification accuracy rates of 4 kinds of stream ciphers (unit: %). (b)One to one identification accuracy rates of 9 kinds of features based 4 kinds of stream cipher (unit: %).



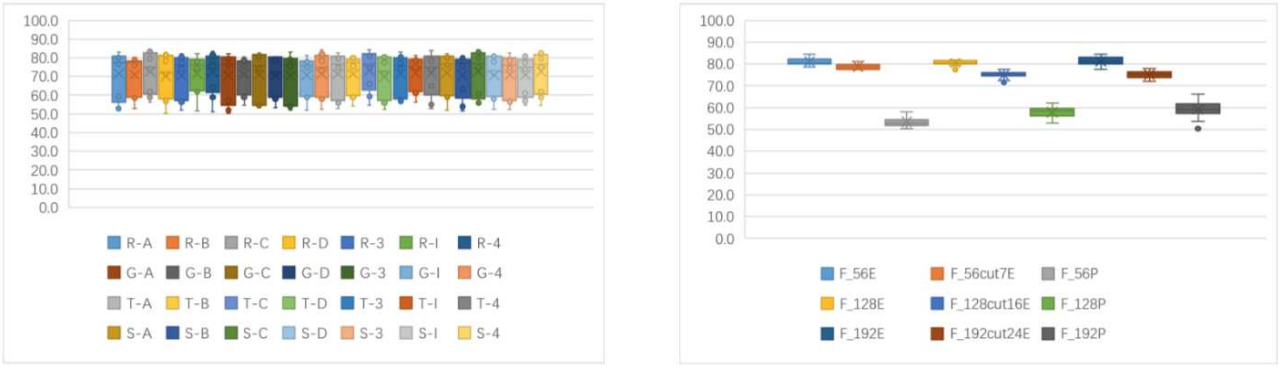
(a) (b)

Fig. 3. (a)One to one identification accuracy rates of 7 kinds of block ciphers in ECB mode (unit: %). (b)One to one identification accuracy rates of 9 kinds of features based 7 kinds of block ciphers in ECB mode (unit: %).



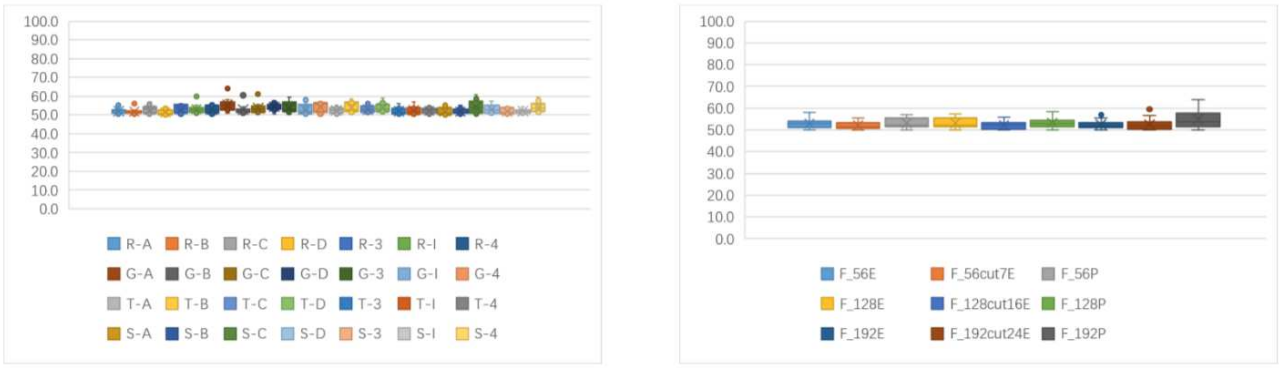
(a) **(b)**

Fig. 4. (a)One to one identification accuracy rates of 7 kinds of block ciphers in CBC mode (unit: %). (b)One to one identification accuracy rates of 9 kinds of features based 7 kinds of block ciphers in CBC mode (unit: %).



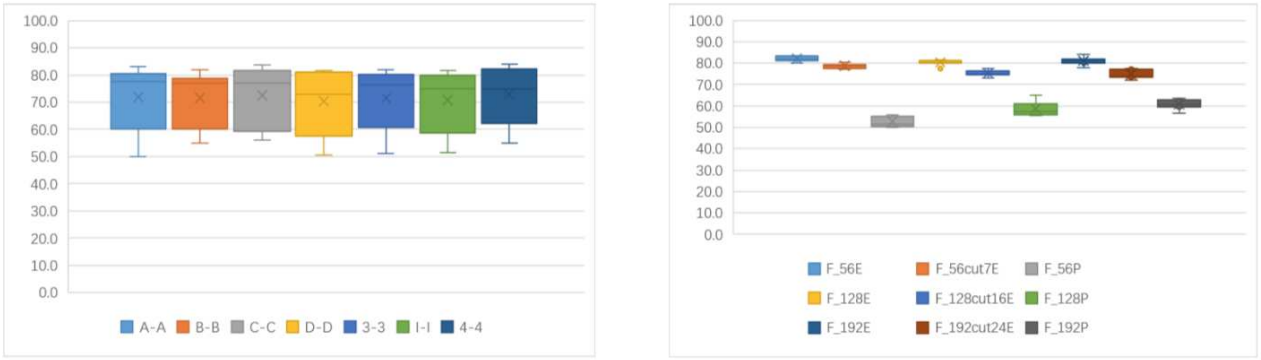
(a) **(b)**

Fig. 5. (a)One to one identification accuracy rates of stream ciphers and block ciphers in ECB mode (unit: %). (b)One to one identification accuracy rates of 9 kinds of features based stream ciphers and block ciphers in ECB mode (unit: %).



(a) **(b)**

Fig. 6. (a)One to one identification accuracy rates of stream ciphers and block ciphers in CBC mode (unit: %). (b)One to one identification accuracy rates of 9 kinds of features based stream ciphers and block ciphers in CBC mode (unit: %).



(a) **(b)**

Fig. 7. (a)One to one identification accuracy rates of block ciphers in ECB mode and CBC mode (unit: %). (b)One to one identification accuracy rates of 9 kinds of features based block ciphers in ECB mode and CBC mode (unit: %).

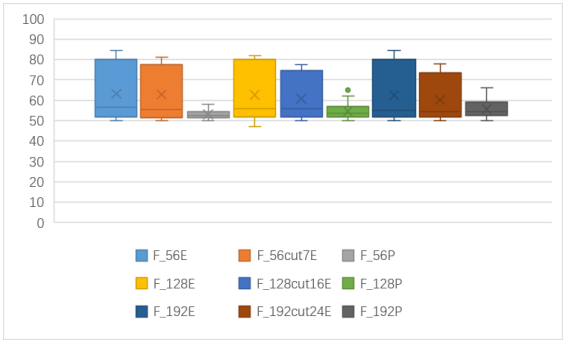


Fig. 8. One to one identification accuracy rates of 9 kinds of features based 18 kinds of cryptosystems (unit: %).

Fig. 8 shows that, the average identification rates of the 6 entropy-based features are about 60%, while the average identification rates of the 3 probability-based features are about only 55%. The identification rates of entropy-based features can exceed 80%, while the identification rates of probability-based features are not more than 68%. Therefore, the identification rates of entropy-based features are significantly higher than the probability-based features.

5. Model Test

If the average identification accuracy rate of each feature is greater than 50%, then we believe that one to one identification scheme based on FDA is better than random classification. For the one to one identification results of each feature, we carried out the t-test and non-parametric test, and tested whether the identification accuracy rate was greater than 50%.

For the t-test, it is assumed that the accuracy rates of identification have a normal distribution with a mean value of μ, and the corresponding hypothesis testing problem is

$$H_0 : \mu = 50\% \Leftrightarrow H_1 : \mu \geqslant 50\%$$

The results show that the p values of the 9 features of the FDA model respectively are 0.0013, 0.0025, 0.003, 0.0010, 0.0020, 0.0023, 0.0011, 0.0018, 0.0019. Therefore, the original hypothesis can be rejected at a significance level of 0.005. T-test using C program.

For non-parametric tests, the hypothesis testing problem is

$$H_0 : q = 50\% \Leftrightarrow H_1 : q \geqslant 50\%$$

Where q is the one to one identification accuracy rate, and the p values of the 9 features of the FDA model respectively are 0.0026, 0.0026, 0.0093, 0.003, 0.0031, 0.0064, 0.0028, 0.0029, 0.0077. Therefore, the original hypothesis can be rejected at a significance level of 0.01. Non-parametric test is implemented in SPSS software.

For each of the above cases, statistical significance is very significant, and the results show that the FDA based method is superior than random classification.

In addition, the sensitivity test of FDA model was also implemented. We select 'lsqr'(least square QR-factorization) in the FDA model solver, it can perform the classification, and support the use of shrinkage to improve the estimation of the covariance matrix. The shrinkage parameter is in the range of [0,1], 0 corresponds to no contraction, the model will make the empirical covariance matrix; 1 corresponds to complete contraction, diagonal covariance matrix will estimate covariance matrix. In the FDA model, the shrinkage parameter was selected as 'auto', which indicated that the appropriate parameters can be selected automatically based on the size of the input data. In the sensitivity test, we set the shrinkage parameters for 0,0.1,0.2, ..., 0.8,0.9,1 total 11 cases respectively. This test only selects the Grain and RC4 of the stream ciphers, and the result is shown in Fig. 9. Because the model remains the same and other cases are similar to this, the results are no longer listed.

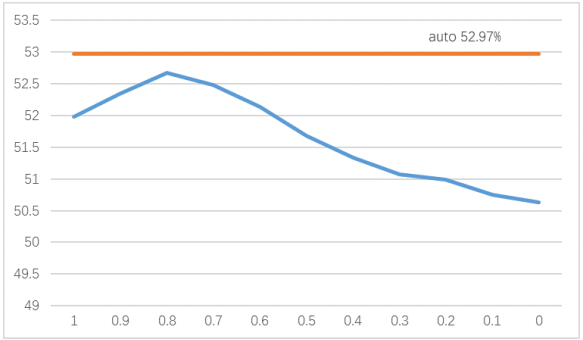


Fig. 9. Identification accuracy rates of Grain and RC4 of different shrinkage parameters in FDA model (unit: %).

Fig. 9 shows that the identification result of our selected parameter 'auto' is better than the result of manual setting of parameters. And with the change

of the shrinkage parameters, the identification accuracy rates had also changed significantly. Therefore, our FDA model selected the correct parameters and had good sensitivity.

6. Conclusion

In this paper, we discuss 4 kinds of stream ciphers and 7 kinds of block ciphers in ECB and CBC modes, and identify the ciphertext generated by 18 different cryptosystems one to one. Aiming at the research objectives, 9 features are extracted and FDA is used to identify the ciphertext of 18 cryptosystems extracted from different features. FDA is a typical discriminant method, and the identification accuracy rates of block ciphers in ECB mode and stream ciphers, and block ciphers in two modes can reach 80%. In the FDA identification results, we find that the accuracy rate of SMS4 with other 6 algorithms of block ciphers in CBC mode can reach 55%-59%, which is superior than the existing papers. In the future, we will continue to study the identification of SMS4 algorithm with other cryptosystems.

Acknowledgments

This work is supported by National Key Research and Development Project 2016-2018 (2016 YFE0100600); State Key Laboratory of Information Assurance Technology Open Fund Project (KJ-15-008).

Appendix A Detailed Results

The detailed results of accuracy rate of the 18 cryptosystems for one to one identification under 9 different features are shown in Appendix A.

Appendix B Validation Methods

The experiment results of five-fold cross validation method and twenty-fold cross validation method are shown in Appendix B.

References

References are to be listed in the order cited in the text. Use the style shown in the following examples. For journal names, use the standard abbreviations. Typeset references in 9 pt Times Roman.

1. D. Feng and D. Pei, *Cryptology guidance*, (Science Press, Beijing, China, 1994).
2. B.Xi, *Cryptology*, (Information Engineer University, 2003).
3. C. Tan and Q. Ji, *An approach to identifying cryptographic algorithm from ciphertext*, (IEEE International Conference on Communication Software and Networks, 8, 2016), pp. 19–23.
4. S. O. Sharif and et.al, *Classifying encryption algorithms using pattern recognition techniques*, (IEEE International Conference on Information Theory and Information Security, 2010), pp. 1168–1172.
5. Y. Wu and et.al, *Block ciphers identification scheme based on the distribution character of randomness test values of ciphertext*, (Journal on Communications, 36, 2015), pp. 146–155.
6. J. G. Dunham and et.al, *Classifying file type of stream ciphers in depth using neural networks*, (IEEE International Conference on Computer Systems and Applications, 2005), pp. 97–103.
7. S. Mishra and A. Bhattacharjya, *Pattern analysis of cipher text: a combined approach*, (International Conference on Recent Trends in Information Technology, 2013), pp. 393–398.
8. A. D. Dileep and C. C. Sekhar, *Identification of block ciphers using support vector machines*, (International Joint Conference on Neural Networks, 2006), pp. 2696–2701.
9. R. Manjula and R. Anitha, *Identification of encryption algorithm using decision tree*, (First International Conference on Computer Science and Information Technology, 2011), pp. 237–246.
10. A. E. Cortizo and et.al, *Application of Fisher linear discriminant analysis to speech/music classification*, (The International Conference on IEEE, 2, 2005), pp. 1666–1669.
11. T. Cooke, *Two variations on Fisher's linear discriminant for pattern recognition*, (IEEE Trans. Pattern Anal. Mach. Intelligence, 24, 2002), pp. 268–273.
12. Z. Liang and et.al, *Two-dimensional Fisher discriminant analysis and its application to face recognition*, (In Computer Vision ACCV Springer, 2006), pp. 130–139.
13. P. K. Ray and et.al, *Classification of encryption algorithms using Fisher's discriminant analysis*, (Defence Science Journal, 67, 2017), pp. 59–65.
14. J. Fan and C. Mei, *Data Analysis*, (Science Press, Bei-

jing, China, 2010).
15. H. Gao, *Application of Multivariate Statistical Analysis*, (Peking University Press, Beijing, China, 2005).
16. A. R. Webb and K. D. Copsey, *Statistical Pattern Recognition*, (Electronic Industry Press, Beijing, China, 2015).
17. X. Zhang, *Pattern Recognition*, (Tsinghua University Press, Beijing, China, 2010).
18. G. Griffin and et.al, *Caltech-256 Object Category Dataset*, (Califonia Institute of Technology, 2007).

Appendix A. The success rate of the 18 cryptosystems for one to one identification under 9 different features

Table 1. One to one identification success rates of 4 kinds of stream ciphers (unit: %)

	Random	F_{56E}	$F_{56cut7E}$	F_{56P}	F_{128E}	$F_{128cut16E}$	F_{128P}	F_{192E}	$F_{192cut24E}$	F_{192P}
R-G	50	53	50.5	53.5	51	56	50	51	51	55
R-T	50	51.5	51	52.5	51	52.5	54	53	53	56
R-S	50	51.5	52.5	56.5	52.5	51	60.5	53	50.5	51
G-T	50	53.5	50.5	54.5	53.5	56	53.5	54	50.5	62.5
G-S	50	51.5	52	54	56.5	50.5	51.5	53	52.5	57
T-S	50	51.5	53.5	50.5	52	53	53	50.5	52.5	54

Table 2. One to one identification success rates of 7 kinds of block ciphers in ECB mode (unit: %)

	Random	F_{56E}	$F_{56cut7E}$	F_{56P}	F_{128E}	$F_{128cut16E}$	F_{128P}	F_{192E}	$F_{192cut24E}$	F_{192P}
A-B	50	63.5	72	51.5	63.5	53	55	59	52.5	53.5
A-C	50	55.5	71	51.5	56	52.5	52	52.5	51	54
A-D	50	68	68.5	53	65.5	57.5	51	63.5	53	52.5
A-3	50	59.5	68.5	57	64	64	52.5	62	59	56
A-I	50	66	66.5	51	65.6	55	54	63.5	54.5	56
A-4	50	52	50	51.5	53.5	50.5	51.5	53.5	52.5	50.5
B-C	50	68.5	71.5	54.5	55.5	59	52	56.5	53.5	51
B-D	50	57.5	51.5	51.5	51.5	50	53	53	52	52.5
B-3	50	52	52.5	56.5	53	65.5	50	53.5	58.5	51
B-I	50	57	50.5	54.5	50.5	59	52.5	51	56.5	51.5
B-4	50	59	73	53	63	58	57.5	60.5	57.5	51
C-D	50	63	71.5	54.5	61.5	58.5	59.5	55.5	54.5	54
C-3	50	65	72	53.5	68.5	68	52	66.5	66.5	52.5
C-I	50	60.5	71.5	53	63	58.5	53	63.5	54	52
C-4	50	51	59	52	58	60.5	51	53	57.5	50.5
D-3	50	50.5	50.5	52.5	64	60	52	53	56.5	50.5
D-I	50	62	51	55.5	57.5	54	52	57	51	54.5
D-4	50	61.5	73	52.5	51	60.5	50.5	52.5	60	53.5
3-I	50	58	51.5	55	54.5	58	52.5	53	55.5	50.5
3-4	50	62.5	71	51.5	61.5	68.5	50.5	65.5	63	50.5
I-4	50	61.5	71	51	61	64	50.5	58.5	56	53.5

Table 3. One to one identification success rates of 7 kinds of block ciphers in CBC mode (unit: %)

	Random	F_{56E}	$F_{56cut7E}$	F_{56P}	F_{128E}	$F_{128cut16E}$	F_{128P}	F_{192E}	$F_{192cut24E}$	F_{192P}
A-B	50	52	51	54	52	51.5	56	53	51	54
A-C	50	50	52	55	53	50.5	53.5	51	50	53
A-D	50	53	51	51.5	55.5	53	51.5	53	55	56.5
A-3	50	52	51	56	52.5	51	50.5	52	55.5	53.5
A-I	50	50.5	50.5	52	50.5	50	52	52.5	54	53.5
A-4	50	57.5	50.5	51	53	50	52	51	52	54
B-C	50	53.5	53	55	51	55.5	51.5	56.5	52.5	57.5

B-D	50	52	52	51	54.5	52	50	58	52	50
B-3	50	51	57	56	56.5	50.5	54.5	52	52.5	55
B-I	50	52	50.5	52	50.5	53.5	51	51	53	53
B-4	50	53.5	55.5	57.5	57	57	52.5	50	53.5	53.5
C-D	50	53.5	53	51	50.5	51.5	51	52.5	51.5	53
C-3	50	52	54	51	51	56	52	54	53	56.5
C-I	50	53	55.5	50	51.5	51	53	50.5	51	57
C-4	50	50.5	57	52	54.5	54	51.5	50.5	51.5	51
D-3	50	57	55.5	50	52.5	54.5	53	50.5	52	50.5
D-I	50	52	51	50.5	52.5	50.5	53	51.5	50.5	52.5
D-4	50	56.5	52	56	52.5	55	57	56.5	55	51.5
3-I	50	52	50	54	53.5	52	51	54.5	51.5	52
3-4	50	50.5	60	54	54	58	54	51.5	51	50
I-4	50	56.5	53	53	51	51.5	59	50.5	54.5	54.5

Table 4. One to one identification success rates of stream ciphers and block ciphers in ECB mode (unit: %)

	Random	F_{56E}	$F_{56cut7E}$	F_{56P}	F_{128E}	$F_{128cut16E}$	F_{128P}	F_{192E}	$F_{192cut24E}$	F_{192P}
R-A	50	83	80.5	53	80	77.5	53	81	76.5	59.5
R-B	50	80	77.5	53	77.5	76	58.5	79	77.5	59.5
R-C	50	84	79.5	56.5	81.5	74	62	83.5	75	59
R-D	50	81.5	78.5	50.5	81	71.5	59	81.5	72	57
R-3	50	79.5	77.5	52	80.5	75.7	56.5	82	74	57.5
R-I	50	80	78	51.5	82	74.5	62	78	76	62.5
R-4	50	80.5	78.5	51	81.5	76	59	83.5	75.5	64
G-A	50	82	81	52	79.5	76.5	57	80	76	50.5
G-B	50	78.5	78	54.5	77.5	75.5	61.5	80.5	77.5	59
G-C	50	82.5	80.5	53.5	81.5	74.5	54.5	82	75.5	55
G-D	50	80.5	78.5	58	80.5	71.5	59.5	80	72	53.5
G-3	50	79	79	52	80.5	76	54	83	75.5	54.5
G-I	50	79	77.5	52	81.5	75	59	78	75.5	59.5
G-4	50	82	78.5	52.5	80.5	75	58	84.5	73	59.5
T-A	50	82.5	81	53	79.5	76.5	58.5	80.5	75	56
T-B	50	80.5	78	54	77.5	75.5	60.5	80	77.5	59
T-C	50	84.5	81	54.5	81.5	74.5	59.5	82.5	76.5	66
T-D	50	80.5	77.5	52.5	80.5	71.5	56	80.5	72.5	58
T-3	50	79.5	77.5	57	80.5	75.5	57	83	74	59
T-I	50	80	78.5	56.5	81.5	75	61	77.5	75	63
T-4	50	80.5	78.5	53	81	75	55	84	74	65
S-A	50	82	81	52	80	76	59	81	76.5	60.5
S-B	50	80.5	77.5	51.5	77.5	75.5	54	80	78	63
S-C	50	84.5	80	55.5	82	75	56	83	77	60.5
S-D	50	81	78	52.5	81	72.5	60.5	80.5	72	58
S-3	50	79.5	78.5	52.5	80.5	76	56.5	82.5	74.5	57.5
S-I	50	81.5	77.5	54.5	81	74	57	77.5	75	61
S-4	50	82.5	79.5	54.5	81	76.5	59	83.5	75	62

Table 5. One to one identification success rates of stream ciphers and block ciphers in CBC mode (unit: %)

	Random	F_{56E}	$F_{56cut7E}$	F_{56P}	F_{128E}	$F_{128cut16E}$	F_{128P}	F_{192E}	$F_{192cut24E}$	F_{192P}
R-A	50	50	52.5	55	51.5	52.5	52.5	52.5	50.5	51.5
R-B	50	51	50.5	51.5	56	52.5	52	50	51	51.5
R-C	50	51	50	51	55.5	50.5	53	53	53	57
R-D	50	51	50.5	52	54	50	50.5	50.5	50	53.5
R-3	50	53.5	53.5	51	56	56	55.5	51	50	50.5
R-I	50	51.5	51.5	53	55	52.5	50.5	52.5	60	51
R-4	50	56.5	55.5	52.5	51	55	50.5	52	50.5	52.5
G-A	50	58	52.5	55.5	53	53.5	55.5	55	51	64
G-B	50	54.5	50.5	51.5	52	51	51.5	52	52	60.5
G-C	50	53	51.5	53.5	51.5	55.5	52.5	51	54	61
G-D	50	55	54	57	55	54	54	52.5	50.5	56.5
G-3	50	56.5	52.5	52	51.5	52	54.5	57	50.5	59.5
G-I	50	53.5	50	50.5	50.5	51	58.5	52	53	58
G-4	50	56.5	52	56	56	50.5	57	50.5	53	56.5
T-A	50	52.5	51.5	51.5	50.5	54.5	52.5	55	50.5	53.5
T-B	50	52	54	51	52	52.5	57	51.5	56.5	58.5
T-C	50	50.5	51.5	56	51.5	52.5	51.5	53.5	56.5	53
T-D	50	53	52	54.5	56.5	50.5	52	59	52	54
T-3	50	53.5	51.5	56	51.5	50	53.5	50.5	51.5	50.5
T-I	50	53.5	52	52	57	50.5	54.5	50.5	52	50
T-4	50	53	51	52.5	52.5	50.5	54	52	50.5	54
S-A	50	50	50.5	51	52.5	50.5	50	55	51	55.5
S-B	50	51	55	51.5	50	51.5	53.5	53	50.5	51
S-C	50	54	50.5	56	53	51	58.5	55.5	50	60.6
S-D	50	51	51	56.5	57.5	52.5	51	52.5	54	51.5
S-3	50	54	54.5	50	50.5	50.5	53.5	51.5	52	50.5
S-I	50	52	51	52	51	52.5	51	50.5	54.5	53
S-4	50	51	53.5	54.5	52	53.5	52.5	51.5	59.5	57.5

Table 6. One to one identification success rates of block ciphers in CBC modes and ECB modes (unit: %)

	Random	F_{56E}	$F_{56cut7E}$	F_{56P}	F_{128E}	$F_{128cut16E}$	F_{128P}	F_{192E}	$F_{192cut24E}$	F_{192P}
A-A	50	83	80.5	50	80	77.5	60	80.5	75.5	60.5
B-B	50	82	77	55	77.5	75.5	57.5	80	77.5	63
C-C	50	83.5	79.5	56	81.5	74.5	56	82	77	62.5
D-D	50	81	78.5	50.5	81	73	55.5	81.5	72	59.5
3-3	50	80	77.5	51	80.5	76.5	65	82	74	56.5
I-I	50	81.5	78.5	51.5	81	75	57	78	73.5	60.5
4-4	50	83.5	79.5	55	81	75	61	84	73.5	63.5

Appendix (B, 1). Five-fold cross validation

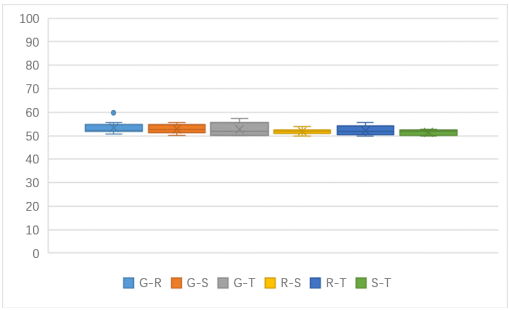
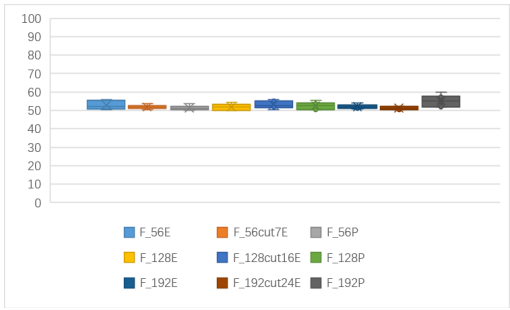


Fig. 1 (a). One to one identification success rates of 4 kinds of stream ciphers (unit: %).



(b) One to one identification success rates of 9 kinds of features based 4 kinds of stream cipher (unit: %).

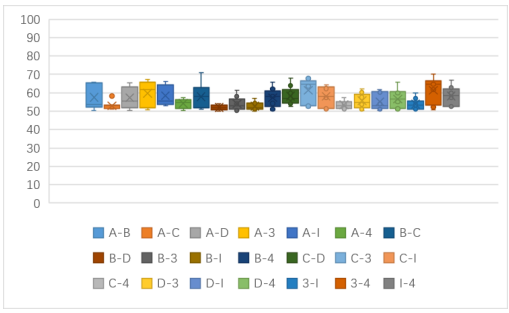
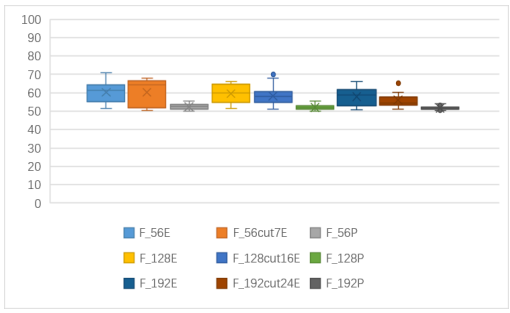


Fig. 2 (a). One to one identification success rates of 7 kinds of block ciphers in ECB mode (unit: %).



(b) One to one identification success rates of 9 kinds of features based 7 kinds of block ciphers in ECB mode (unit: %).

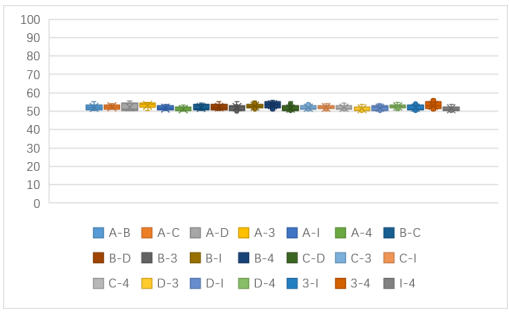
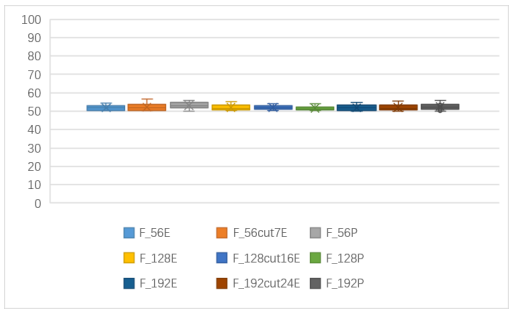


Fig. 3 (a). One to one identification success rates of 7 kinds of block ciphers in CBC mode (unit: %).



(b) One to one identification success rates of 9 kinds of features based 7 kinds of block ciphers in CBC mode (unit: %).

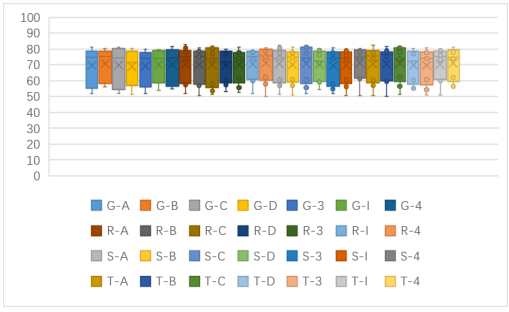
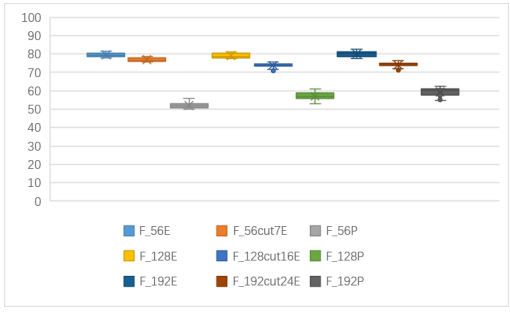


Fig. 4 (a). One to one identification success rates of stream ciphers and block ciphers in ECB mode (unit: %).



(b) One to one identification success rates of 9 kinds of features based stream ciphers and block ciphers in ECB mode (unit: %).

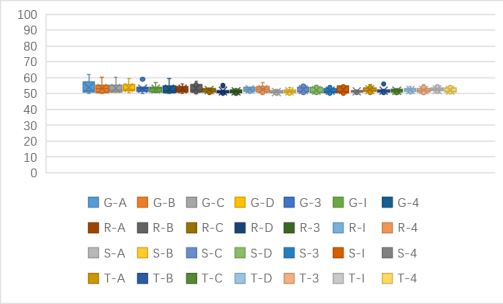
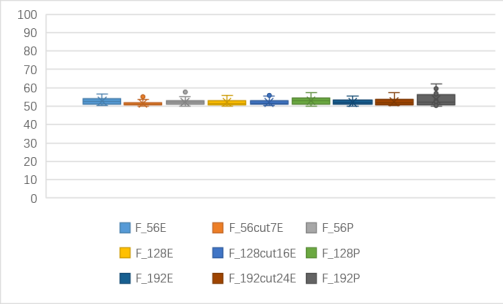


Fig. 5 (a). One to one identification success rates of stream ciphers and block ciphers in CBC mode (unit: %).



(b) One to one identification success rates of 9 kinds of features based stream ciphers and block ciphers in CBC mode (unit: %).

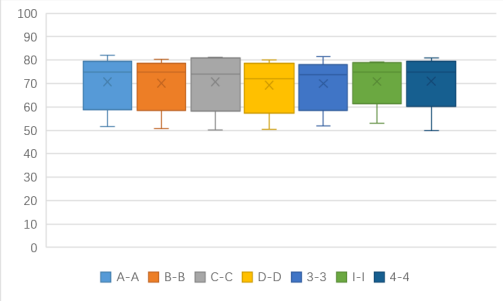
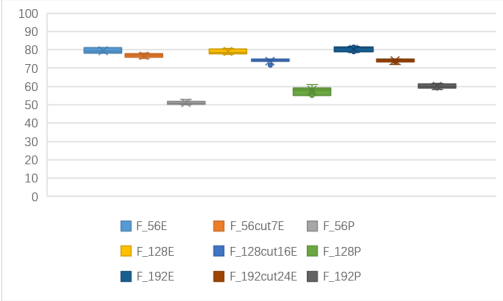


Fig. 6 (a). One to one identification success rates of block ciphers in ECB mode and CBC mode (unit: %).



(b) One to one identification success rates of 9 kinds of features based block ciphers in ECB mode and CBC mode (unit: %).

Fig. 7. One to one identification success rates of 9 kinds of features based 18 kinds of cryptosystems (unit: %).

Appendix (B, 2). Twenty-fold cross validation

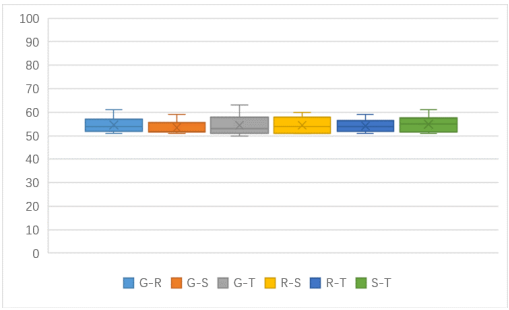
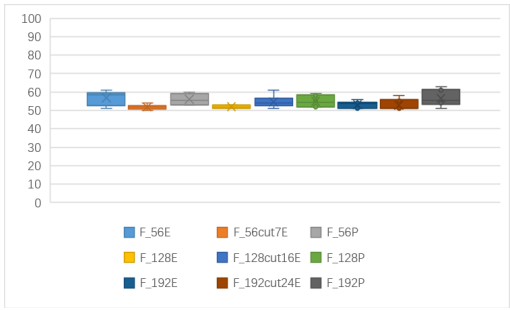


Fig. 1 (a). One to one identification success rates of 4 kinds of stream ciphers (unit: %).



(b) One to one identification success rates of 9 kinds of features based 4 kinds of stream cipher (unit: %).

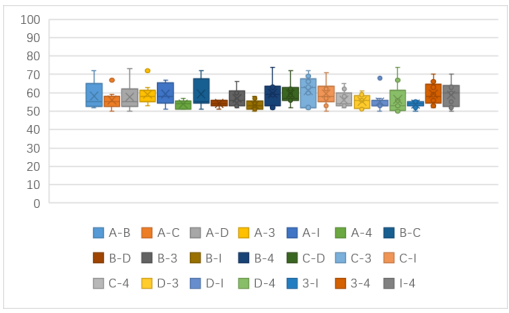
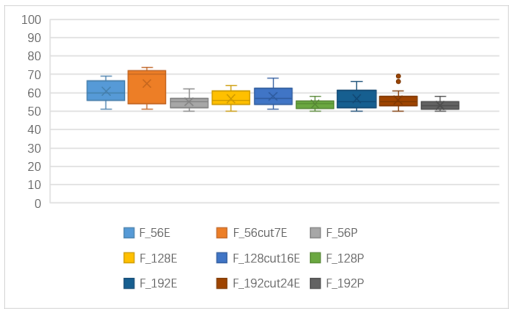


Fig. 2 (a). One to one identification success rates of 7 kinds of block ciphers in ECB mode (unit: %).



(b) One to one identification success rates of 9 kinds of features based 7 kinds of block ciphers in ECB mode (unit: %).

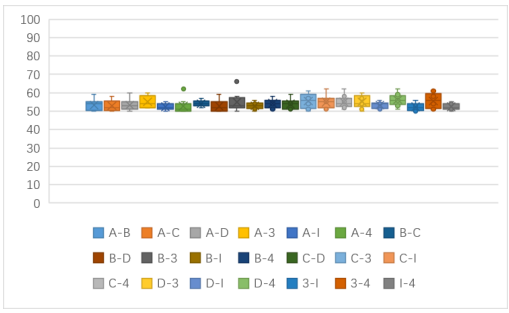


Fig. 3 (a). One to one identification success rates of 7 kinds of block ciphers in CBC mode (unit: %).

(b) One to one identification success rates of 9 kinds of features based 7 kinds of block ciphers in CBC mode (unit: %).

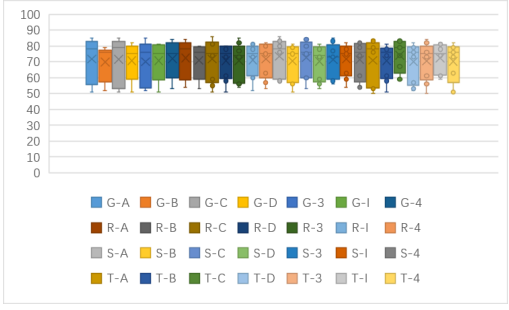
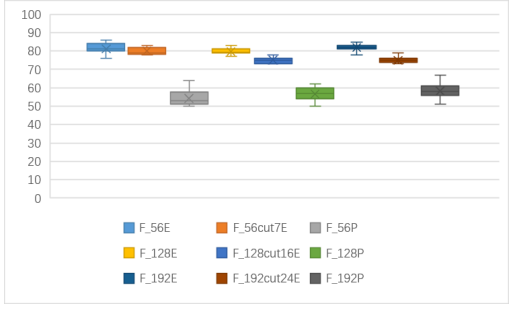


Fig. 4 (a). One to one identification success rates of stream ciphers and block ciphers in ECB mode (unit: %).



(b) One to one identification success rates of 9 kinds

of features based stream ciphers and block ciphers in ECB mode (unit: %).

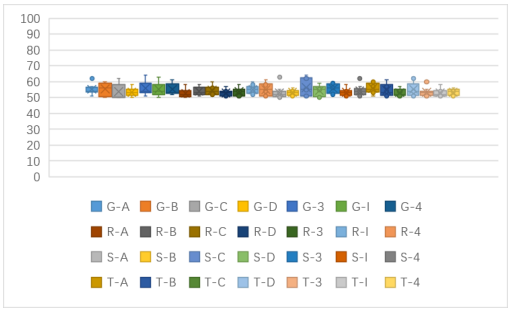


Fig. 5 (a). One to one identification success rates of stream ciphers and block ciphers in CBC mode (unit: %).

(b) One to one identification success rates of 9 kinds of features based stream ciphers and block ciphers in CBC mode (unit: %).

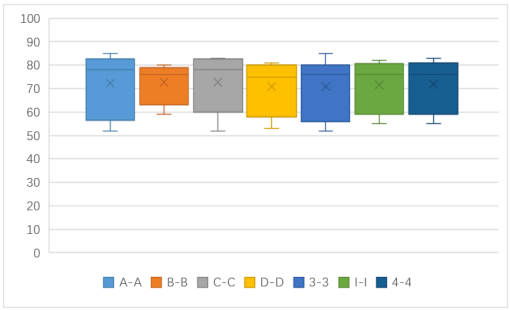
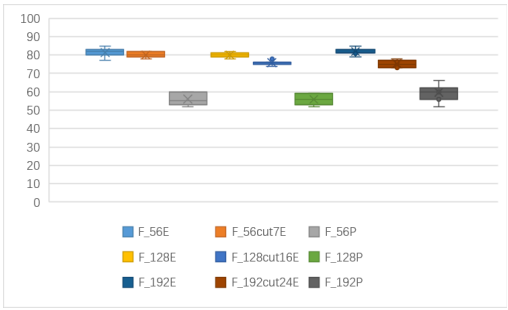


Fig. 6 (a). One to one identification success rates of block ciphers in ECB mode and CBC mode (unit: %).



(b) One to one identification success rates of 9 kinds of features based block ciphers in ECB mode and CBC mode (unit: %).

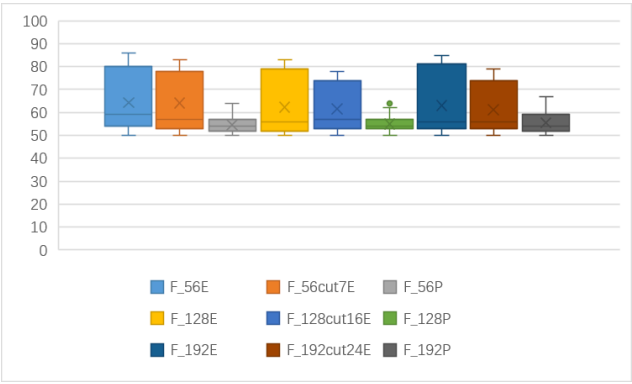


Fig. 7. One to one identification success rates of 9 kinds of features based 18 kinds of cryptosystems (unit: %).