Hindawi
Security and Communication Networks
Volume 2020, Article ID 8875864, 10 pages
https://doi.org/10.1155/2020/8875864

Research Article

Cryptosystem Identification Scheme Based on ASCII Code Statistics

Wenyu Zhang,[1] **Yaqun Zhao**,[2] **and Sijie Fan**[2]

[1]*State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China*
[2]*State Key Laboratory of Cryptography and Science, Beijing, China*

Correspondence should be addressed to Wenyu Zhang; 994709309@qq.com

Received 9 August 2020; Revised 25 November 2020; Accepted 3 December 2020; Published 15 December 2020

Academic Editor: Angel M. Del Rey

In the field of information security, block cipher is widely used in the protection of messages, and its safety naturally attracts people's attention. The identification of the cryptosystem is the premise of encrypted data analysis. It belongs to the category of attack analysis in cryptanalysis and has important theoretical significance and application value. This paper focuses on the extraction of ciphertext features and the construction of cryptosystem identification classifiers. The main contents and innovations of this paper are as follows. Firstly, inspired by language processing, we propose the feature extraction scheme based on ASCII statistics of ciphertexts which decrease the dimension of data preprocessing. Secondly, on the basis of previous work, we increase the types of block ciphers to eight, encrypt plaintext of the same sizes as experimental objects, and recognize the cryptosystem. Thirdly, we use two machine learning classifiers to perform classification experiments including random forest and SVM. The experimental results show that our scheme can not only improve the identification accuracy of 8 typical block cipher algorithms but also shorten the experimental time and reduce the computation load by greatly minimizing the dimension of the feature vector. And the various evaluation indicators obtained by the scheme have been greatly improved compared with the existing published literature.

1. Introduction

With the development of information industry, information security gradually becomes an important part of society. According to the requirement in different complex conditions, cryptologists design a number of encryption algorithms such as DES, AES, and IDEA. The mathematic theory applied in different cryptosystems is different. Therefore, there is no generic satisfactory cryptanalysis method to fix all the problems met in cryptanalysis. Most practical cryptanalysis techniques are designed for certain cryptosystems with specific structure. Therefore, the identification of the cryptosystem becomes a basic task of cryptanalysis which should be solved before the analysis of certain cryptosystems. At the same time, the ability to resist the identification of the cryptosystem can be used as an indicator to measure the security of the cryptosystem, which provides a valuable reference for the design of the cryptosystem [1]. The cryptosystems which can resist distinguishing attack are seen as the algorithm with strong security. The research on the identification of the cryptosystem has played a dual role in promoting the application of cryptanalysis and the development of cryptography [2].

In 2006, Dileep and Sekhar [3] proposed a block cipher recognition scheme based on support vector machine (SVM) with the help of text classification and counting. The author compared the recognition performance of SVM and *K*-means method and adopted a multicryptosystem including fixed-length document vector and variable-length document vector. In 2008, Nagireddy [4] considered cryptosystem recognition as a cryptosystem attack and recognized five cryptosystems. It was found that block ciphers in ECB mode were easier to recognize.

In 2012, Chou et al. [5] proposed a classification scheme based on support vector machine (SVM), which can be recognized and classified as two working modes of block

ciphers (CBC and ECB). The experiment is designed as one-to-one classification distinguishing progress which is implemented in three encryption algorithms (AES, DES, and RC4) with 1000 samples. The experiment result shows the strong classification ability in the cryptographic distinguishing attack of SVM. In 2018, Hu and Zhao [6] designed a distinguishing attack based on Fisher's discriminant analysis (FDA) theory. In this work, the authors extract 9 kinds of statistical data as the feature of ciphertext which is used to distinguish 4 stream ciphers and 7 block ciphers in a one-to-one identification experiment. The experiment result shows the identification accuracy of encrypted files in ECB mode can reach 80%. The identification accuracy of stream ciphers SMS4 from block ciphers in CBC mode can reach 60%.

In 2018, Huang et al. [7] proposed a two-stage cryptosystem recognition scheme based on random forest. In this work, Huang et al. divide the cryptosystem recognition problem into 2 sequential procedures, "cluster recognition" and "single recognition." In the first stage, the scheme recognizes the cluster of cryptosystems, and then, the classifier identifies the type of cryptosystem. Compared with the traditional single-stage scheme, the two-stage scheme outperforms 19.55%, 21.40%, and 22.99% with respect to recognition accuracy in the 3 considered settings, respectively.

In this paper, the main contribution of our work is organized as follows. In Section 2, we give the basic definition of the cryptosystem identification and the system description of the cryptosystem identification scheme. In Section 4, we propose a cryptosystem identification scheme based on the statistical characteristics of the ASCII code. In Section 5, we have considered the influence of our feature extraction scheme in different working modes of block ciphers on the experimental results. Moreover, we increase the number of cryptosystems to eight (AES-128, AES-256, Blowfish-64, Camellia-128, DES, 3DES, IDEA-64, and SMS4-128). In order to show the influence of different machine learning classifiers (support vector machine and random forest), we compare the evaluation of experiments with various experimental indicators, including precision, recall, and $F1$-score.

2. Preliminaries

To guarantee the safety of communication and document transmission, block ciphers are frequently applied by people. Application of block cipher has become a standard in privacy guarantee, which refuse anyone other than the communicator to obtain the information in transmission. We present below the brief introduction of block cipher which is necessary to understand our work.

2.1. Block Cipher. As a core part of cryptosystem, block cipher is widely used to protect the security of information. While making documents secret, block cipher is also applied as the basic function, such as random number generator, hash function, and digital signature [8]. To encrypt the content of the message, plaintext is divided into fixed length, which is named as block.

Before encrypting, block cipher divides message m into a group of fixed length $\{m_1, m_2, m_3, \ldots, m_n\}$. Adding message and key k, encryption algorithm e_k outputs ciphertexts in groups $\{c_1, c_2, c_3, \ldots, c_n\}$ $(c_i = e_k(m_i))$, where

(a) $|m_1| = |m_2| = |m_3| = \cdots = |m_n|$

(b) $|c_1| = |c_2| = |c_3| = \cdots = |c_n|$

(c) $\sum_{i=1}^{n} |m_i| = \sum_{i=1}^{n} |c_i|$

Assuming that the key is the same, the transformation of block cipher to any plaintext block would not be different. Therefore, the research of block cipher only needs to study the transformation law of any group.

The workflow of block cipher is shown in Figure 1.

2.2. Two Operation Modes of Block Cipher. Adapting to different work requirements, there are several operation modes of block ciphers for the user to choose from. In this work, we mainly consider the electronic codebook (ECB) mode and cipher block chaining (CBC) mode.

ECB mode is a concise encryption method, which encrypts each block of plaintext separately. Applying the same operation for all blocks (as shown in Figure 2), the encryption process can be realized by parallel computing, requiring that plaintext bit length is an integral multiple of block [9].

In 1976, IBM designed and proposed the CBC (cipher block chaining) mode which is an improvement of ECB mode on encryption indeterminacy [10]. Instead of encrypting each block directly, the CBC mode adds a random IV (initialization vector) to the plaintext before encryption and sets the previous cipher block as the next IV (as shown in Figure 3).

3. Theoretical Fundamentals for Cryptosystem Identification

3.1. Cryptosystem Identification. Most of the cryptosystem identification tasks based on machine learning classifiers adopt supervised learning mode [11]. The scheme can be summarized into four steps [12]. Firstly, select the object of classification and identification. Secondly, extract the feature vectors of the experimental object. Thirdly, select and train the appropriate machine learning classifier. Finally, perform the cryptosystem identification. However, this description is simply to classify the cryptosystem identification as a pattern identification problem, and it is impossible to conduct an in-depth study on the particularity of cryptosystem identification, for it makes it difficult to make a major breakthrough in the technical level.

For the above reasons, in this section, we give the definition of cryptosystem identification and cryptosystem identification scheme. The results of the scheme are evaluated by the evaluation criteria of machine learning identification classification, namely, precision, recall rate, and $F1$-score.

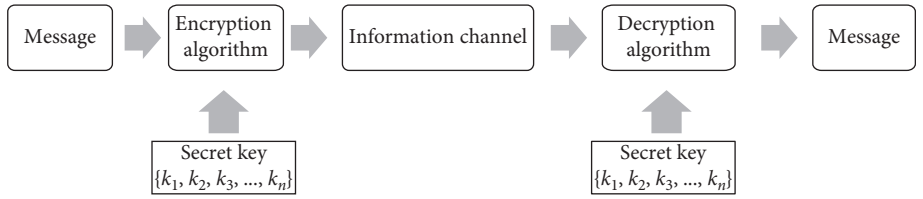

Figure 1: Workflow of block cipher.

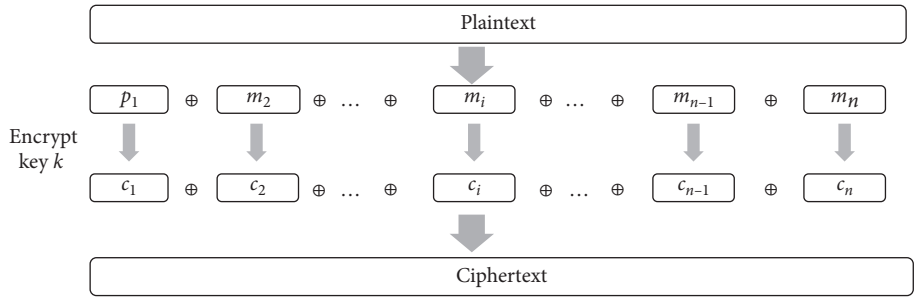

Figure 2: ECB mode.

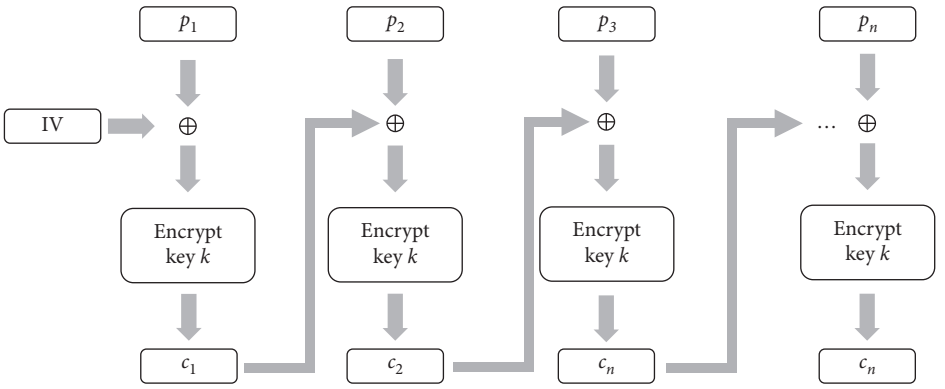

Figure 3: CBC mode.

Definition 1. Consider a cryptosystem set, where n is the number of cryptosystems. C is the ciphertexts generated by the cryptosystem in the cryptosystem set CR. If there is an identification scheme S, the cryptosystem of C can be recognized with a certain evaluation indicator A in the case of its cryptosystem unknown. This process is called cryptosystem identification.

The experimental evaluation indicator A in the above definition usually refers to the machine learning classification evaluation indicator, precision, recall rate, $F1$-score, and accuracy, which is slightly different from the identification accuracy under pattern identification [13].

Definition 2. The WP is the workflow for cryptosystem identification. Fea is the feature extracted from ciphertexts, and RA is the classification classifier applied, and then the triple (WP, Fea, RA) is noted as cryptosystem identification scheme.

We present the workflow of cryptosystem identification in Algorithm 1.

3.2. Random Forest. The random forest uses the bootstrap resampling technique to randomly extract k samples from the original training sample set N, generate a new training sample set, and then, generate k classification trees based on the self-service sample set to form a random forest. The classification result of the new data depends on the score formed by the classification tree voting. Each tree in the forest has the same distribution, and the classification error depends on the classification ability of each tree and the correlation between them. Feature selection applies a random method to split each node and then compares the errors generated in different situations and determines the number of features by estimating error, classification ability, and correlation analysis [14].

3.3. Support Vector Machine. The main theory of support vector machine (SVM) is to establish an optimal decision hyperplane so that the distance between the two types of samples on the two sides of the plane closest to the plane is

maximized. For a multidimensional sample set, the model SVM randomly generates a hyperplane and moves continuously to classify the samples until the sample points belonging to different categories in the training sample are located on both sides of the hyperplane. For the same classification problem, there may exist several different hyperplanes that could separate the dataset with satisfactory accuracy. Therefore, the learning model SVM contains good generalization capabilities in the classification problem. While ensuring the classification accuracy, the SVM finds such a hyperplane to maximize the white space on both sides of the hyperplane, so that it can achieve the optimal classification of linear separable samples [15].

4. A Cryptosystem Identification Scheme Based on ASCII Code Statistics

The encryption process of image data is to convert the image into an array and store it in the database after base64 encryption. However, there are a large number of pixels with the same color in the image, and the distribution of pixels with the same color in different images is different, which may lead to differences in the ASCII code distribution of the encrypted image data. Therefore, based on the difference of the statistical value distribution of the ASCII code in ciphertexts, combined with the definition of the previous cryptosystem identification scheme, we design the following cryptosystem identification scheme. The program consists of two stages: training stage and testing stage (as shown in Figure 4).

4.1. Training Stage

(1) Collect a set of ciphertexts files with known $F_1, F_2, \ldots, F_{m-1}, F_m$ encryption algorithm.

(2) Calculate the frequency of all ASCII codes in the ciphertexts file, and create a dictionary for each ciphertext file. All ASCII codes and their occurrence frequency are the keys and values of the dictionary.

(3) All the dictionary values obtained from each ciphertext are extracted as the feature vector of ciphertexts. Since there are 256 extended ASCII codes, the feature vectors extracted from each ciphertext are 256-dimensional. Then, we get a set of eigenvectors $\mathrm{FEA} = \{\mathrm{fea}_1, \mathrm{fea}_2, \ldots, \mathrm{fea}_m\}$, which are 256-dimensional vectors.

(4) The cryptosystem categories of all ciphertext files can be represented by an m-dimensional array $\mathrm{LABLE} = (l_1, l_2, \ldots, l_m)$; (FEA, label) represents the eigenvector with the cryptosystem label. The labeled data (FEA, label) are input into the classifier to train the classification model.

4.2. Testing Stage

(1) Extracting feature vectors fea^* from ciphertexts F to be recognized

(2) Input the feature vector fea^* into the trained classification model, and the model will give the classification results l^* of ciphertexts F

In this work, we apply random forest classification algorithm and MLP as cryptosystem identification classifiers which are simple to implement and have small computation cost. The diversity of its internal basic learners is not only from the sample disturbance but also from the category attribute disturbance, which can improve its generalization ability.

5. Experiment, Result, and Discussion

In this section, we applied the identification model, which was implemented in classifiers random forest and MLP, and the feature extraction method in previous content. The experiment environment is shown in Table 1.

In this work, we applied the Caltech-256 image dataset of California Institute of Technology as a data source which contains 30607 images [16]. The image data are collected from Google images which have been screened out unsuitable samples. After the collection, we implemented the encryption phase.

Before encryption, we divided the dataset into 1000 files of 512 KB size and then encrypted the pieces with eight encryption algorithms (as shown in Table 2) in ECB and CBC modes, which was performed by software OPENSSL. The same random key was applied while the training stage and testing stage are in both ECB mode and CBC mode. The IV utilized in CBC mode is randomly produced by OPENSSL.

5.1. Evaluation Index. We apply four indexes to measure the classification model. TP (true positive) represents the number of right examples which are sentenced to right ones; TN (true negative) represents the number of right examples which are sentenced to wrong ones; FP (false positive) represents the number of wrong examples which are sentenced to right ones; FN (false negative) represents the number of wrong examples which are sentenced to wrong ones, as shown in Table 3 [17].

Definition 3 (precision). Precision means the ratio of TP in the samples which are sentenced to the right ones:

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{1}$$

Definition 4 (recall). The index recall refers to the proportion of TP in samples which are right:

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{2}$$

Definition 5 ($F1$-score). The harmonic average of precision and recall is as follows:

Input: cryptosystems set CR, testing set $C1$, and training set $C2$
Output: cryptosystems cr
(1) Extract the feature (Fea) of ciphertext $C1$
(2) Train the classification algorithm S with training set $C1$
(3) Input the featureof testing set $C2$
(4) Cryptosystems cr

Algorithm 1: Cryptosystem identification.

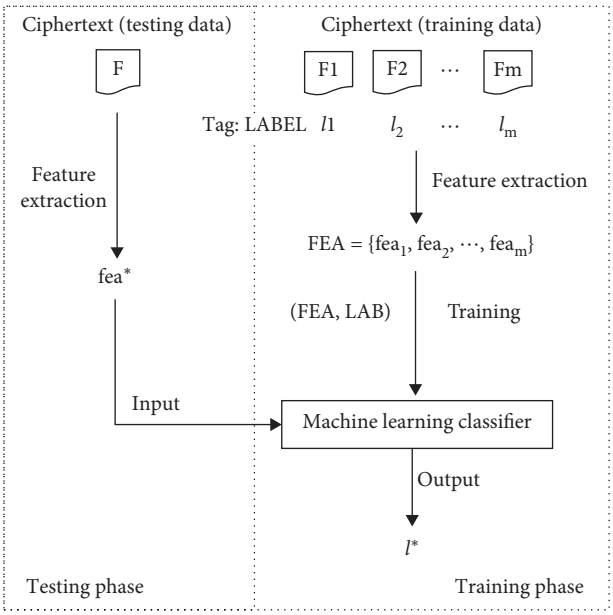

Figure 4: Flowchart of cipher system identification based on machine learning classifier.

Table 1: Experimental environment.

Operation system	Ubuntu16.04LTS
Processor	Intel Core i5-6600K @ 2.8 GHz
Memory	8 GB

Table 2: Cryptosystems and notations.

Cryptosystems	Notation
AES-128	A1
AES-256	A2
Blowfish-64	B
Camellia-128	C
DES	D
3DES	3D
IDEA-64	I
SMS4-128	S

Table 3: Definition of TP, FP, FN, and TN.

Reality	Precision	
	True	False
True	TP	FN
False	FP	TN

$$F1 - \text{score} = \frac{2}{(1/\text{precision}) + (1/\text{recall})} = \frac{2 \times \text{precision} \times \text{recall}}{\text{recall} + \text{precision}}. \tag{3}$$

Definition 6 (accuracy). Accuracy is defined as the proportion of samples which are correctly classified in the whole dataset:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}. \tag{4}$$

We applied the ten-fold cross-validation strategy to calculate the identification accuracy in our experiment. Finally, we got the distribution of identification accuracy and the average values of precision, recall, and $F1$-score.

5.2. Result and Analysis

5.2.1. Evaluation in ECB Mode. To perform the experiment, we divided the database into two parts, 30% collection as a testing set and the other as a training set. Observing the evaluation of classifier (Figures 4–6), the experiment result can be concluded as a successful one.

Figure 5 shows that, in random forest classifier, the identification precision rates in eight encryption algorithms are between 50% and 85%. Identification precision rates of DES and IDEA are higher than 80%. The precision rates of Camellia, 3DES, and SMS4 are between 69% and 77% while the precision rates of AES-256 and Blowfish are 58%, which are obviously lower than the others.

Figure 5 shows that, in the RS classifier, the identification precision rates in eight encryption algorithms are between 78% and 100%. Identification precision rates of AES-128, Camellia-128, DES, 3DES, and IDEA are 100%. For the cryptosystems Blowfish and SMS4, the precision rate also reached 95% and 98%. And for AES-256, the precision rate is 78%.

Precision rate is an index that reflects the true correct proportion of the data predicted correctly. We can conclude that our feature extraction scheme helps the MLP classifier almost make correct prediction of truth. The average of precision in MLP is better than that in random forest.

Figure 6 shows that, for classifier random forest, the recall rates of eight cryptosystems in ECB mode are between 65% and 72% and are evenly distributed. The average recall rates of AES-128 and AES-256 are more than 70%. The recall rates of the rest cryptosystem range from 65% to 68%.

In Figure 6, the recall rate of MLP classifier ranges from 21% to 78%. The recall rate of SMS4 algorithm is the best which reaches 78%. The recall rates of AES-256 and Blowfish are 44% and 40%. The recall rates of AES-128, Camellia, 3DES, and IDEA are between 31% and 24%. The cryptosystem DES has the worst recall rate 21%.

The recall rate is the proportion of the right classification in the right example. We can conclude that our feature extraction method in MLP classifier would miss some right samples, while it has a better performance to find more right samples. And for cryptosystem, SMS4 MLP classifier shows a strong differentiation.

The index $F1$-score comprehensively measures precision and recall which reflect the ability to classify right sample. Figure 7 shows that, for random forest, the $F1$-score rates are between 63% and 74%. The $F1$-score rates of Camellia, DES, 3DES, and IDEA are more than 68%. And the $F1$-score rates of AES-128, AES-256, Blowfish, 3DES, and SMS4 range from 63% to 68% while, for the classifier MLP, the $F1$-score rates are between 34% and 56%.

5.2.2. Evaluation in CBC Mode. Compared to ECB mode, CBC mode is more complex and has higher security. Therefore, it is more difficult for the classifier to identify the cryptosystem with ciphertexts encrypted in CBC mode.

Figure 8 shows that, in CBC mode, the precision rates of random forest classifier in eight encryption algorithms are between 11% and 17%. All in all, identification precision rates of DES and IDEA are higher than 80%. The precision rates of Camellia, 3DES, and SMS4 are between 69% and 77% while the precision rates of AES-256 and Blowfish are 58%, which is obviously lower than the others.

Figure 9 shows that, in the RS classifier, the identification recall rates in eight encryption algorithms are between 10% and 22%. And in the MLP classifier, the identification recall rates in eight encryption algorithms are between 11% and 23%.

In Figure 10, we can conclude that, for the ability to search right samples, there is no obvious difference between MLP classifier and RS classifier for CBC mode. In RS classifier, the identification $F1$-score rates in eight encryption algorithms are between 10% and 19%. In MLP classifier, the identification $F1$-score rates in eight encryption algorithms are between 12% and 24%.

By combining the data, we obtained the average accuracy in two operation modes in Figure 10. In Figure 11, the classification accuracy of random forest classifier in the ECB mode is stable at over 83.0%, and the average value is close to 85.5%. The classification accuracy of CBC mode is lower than that of ECB mode; the average accuracy is 18.0%, but it is still higher than that of random classification by 12.5%.

For MLP classifier, we obtained the average accuracy in two operation modes in Figure 12. In Figure 12, the classification accuracy of random forest classifier in ECB mode is stable at over 50.0%, and the average value is close to 53.0%. The classification accuracy of CBC mode is lower than that of ECB mode; the average accuracy is 13.0%, close to the random classification by 12.5%.

While MLP has better performance in precision rate and $F1$-score, random forest classifier has better performance in identification accuracy. We can conclude that the sample which MLP predicts right is always right, while it would sentence some right sample into wrong ones. Moreover, the indices precision, recall, $F1$-score, and accuracy in CBC mode would be less than that in ECB mode on average.

If the average identification accuracy rate is greater than 12.5%, then we refer the identification progress based on our feature.

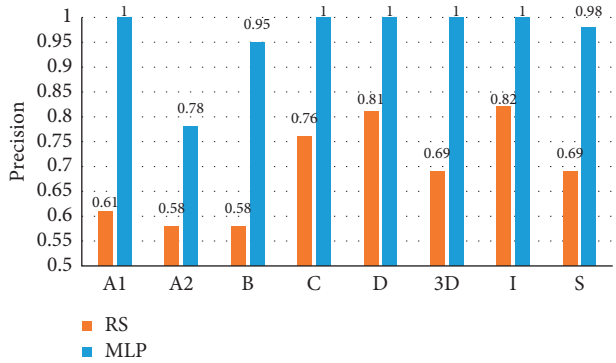

Figure 5: Precision of cryptosystem identification for random forest and MLP in ECB mode.

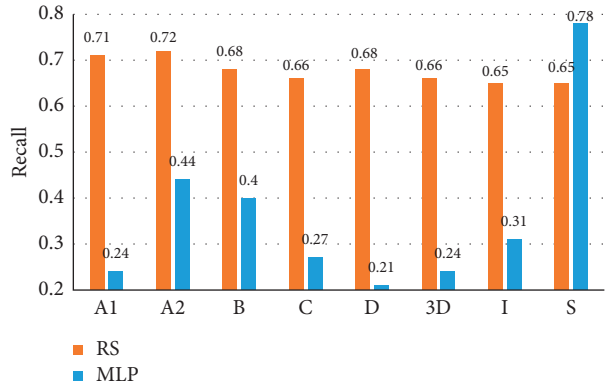

Figure 6: Recall of cryptosystem identification for random forest and MLP in ECB mode.

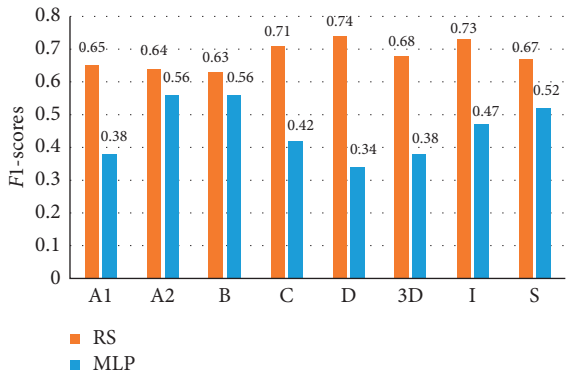

Figure 7: *F*1-scores of cryptosystem identification for random forest and MLP in ECB mode.

5.3. Comparison with Existing Schemes. Table 4 shows the comparison of the results of previous works. From the table, we can see that our cryptosystem identification scheme has higher identification accuracy and can support more kinds of cryptosystem identification tasks. The identification accuracy rate in ECB mode is 84.5% by using random forest classifier, which is higher than the other existing works.

The encryption of operation mode CBC is more difficult for the classifier to recognize than that of ECB mode. The encryption result of ECB mode is only depended on using key and plaintext block, while the encryption result of CBC mode is also affected by the last cipher block. Therefore, the statistical characteristics of encryption in CBC mode are more confused which contributes to the unsatisfactory identification accuracy of 14%.

In addition, we also found that different classifiers have different identification results for different cryptosystems. In particular, for some cryptosystems, there are big differences.

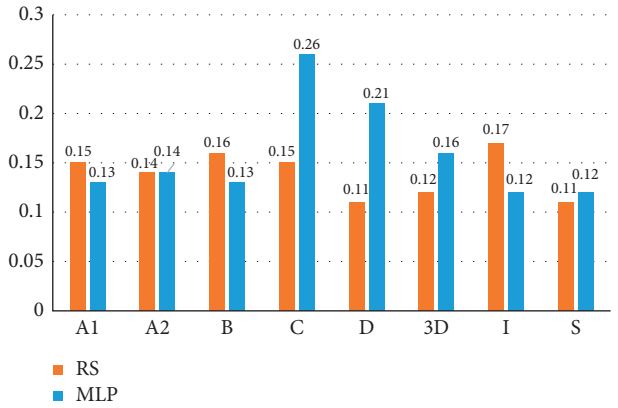

Figure 8: Precision of cryptosystem identification for random forest and MLP in CBC mode.

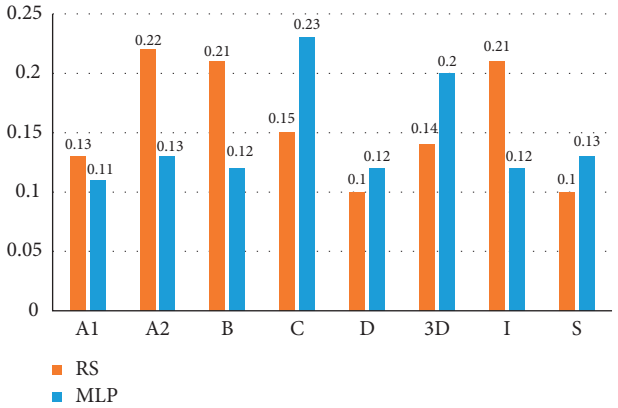

Figure 9: Recall of cryptosystem identification for random forest and MLP in CBC mode.

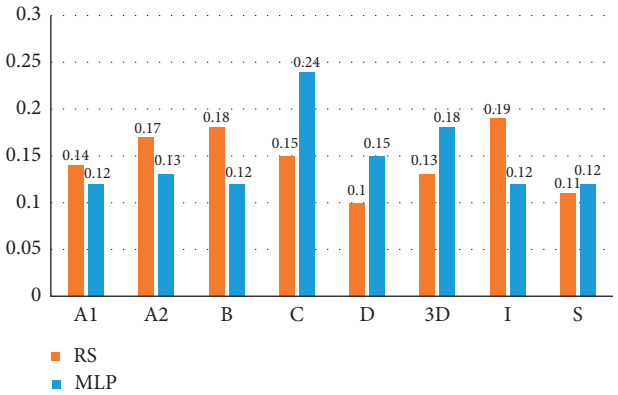

Figure 10: *F*1-score of cryptosystem identification for random forest and MLP in CBC.

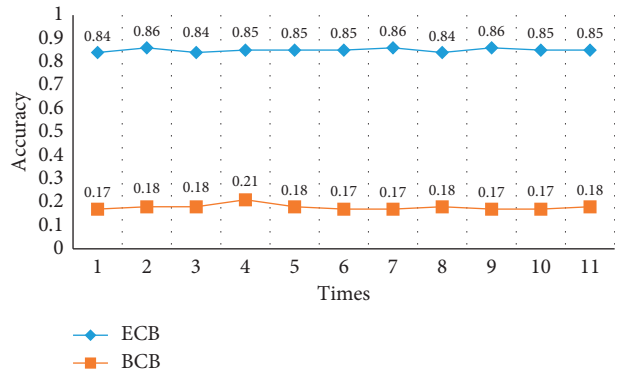

Figure 11: Average accuracy in two operation modes for random forest.

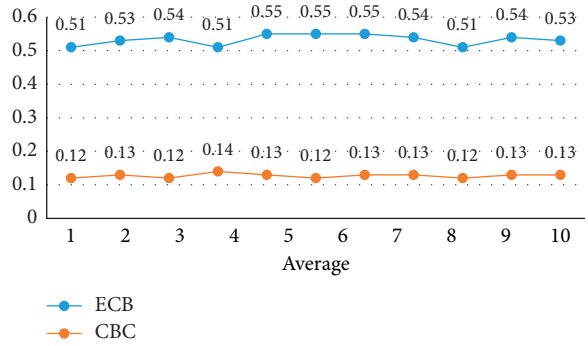

Figure 12: Average accuracy in two operation modes for MLP.

Table 4: Comparison between identification schemes.

Sources of experimental results	This paper	[3]	[7]	[18]	[19]	[6]
Number of cryptosystem types	8	5	5	10	8	5
ECB mode classification accuracy	85.5%	41%	21.5%	36.65%	30.84%	75.5%
CBC mode classification accuracy	14.2%	20%	20%	20%	12.5%	20%

In future research, firstly, we will design better classification features according to different working modes of block cipher encryption algorithm to identify cryptosystems with low identification accuracy, such as CBC mode. Secondly, we can also map stream cipher and public key cipher into the cryptosystem to be recognized to improve the applicability, robustness, and versatility of the scheme. Finally, we will optimize different machine learning classifiers for different cryptosystems to get higher identification accuracy.

6. Conclusion

The cryptosystem identification is an important part of cryptanalysis. This work proposes a novel cryptosystem identification scheme based on the statistical value of ASCII code from ciphertexts with an average accuracy of 84.5%. The classifier distinguished block ciphers: 3DES, AES-128, AES-256, Blowfish, Camellia-128, DES, IDEA, and SMS4 from patterns found on a set of ciphertexts. The result of classification illustrates that the internal mathematic property of the encryptions produces distinguishable features in ciphertexts, which make difference between different encryption.

The experiment shows that the appropriate feature extraction scheme can significantly improve the accuracy of cryptosystem identification, in this case. And as the identification was successful, we can know that the statistical value of ASCII code from ciphertexts reflects some information about encryption. It is a challenge for the rule of cryptography that the attacker cannot obtain any information from ciphertexts.

We concluded that a set of encrypted pictures in ECB mode could be identified by extracting the statistical value of ASCII code, provided that attacker already has some document identified as the training set. Future research may require improvement in feature extraction to classify the ciphertexts more correctly.

Data Availability

The data used to support the findings of this study are included within the article. Data can be used for free by everyone to verify the experimental results.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Key Research and Development Project 2016–2018 (2016YFE0100600), National Natural Science Foundation of China under Grant 61872381, State Key Laboratory of Information Assurance Technology Open Fund Project (KJ-15-008), and State Key Laboratory of Cryptography and Science.

References

[1] M. K. Brown, F. Ca, and G. M. Gutoski, "Using a digital certificate with multiple cryptosystems," *Journal on Communications*, vol. 36, no. 4, pp. 47–69, 2017.

[2] S. Palit, S. N. Sinha, M. A. Molla, A. Khanra, and M. Kule, "A cryptanalytic attack on the knapsack cryptosystem using binary Firefly algorithm," in *Proceedings of the 2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011)*, pp. 428–432, Allahabad, India, September 2011.

[3] A. D. Dileep and C. C. Sekhar, "Identification of block ciphers using support vector machines," in *Proceedings of the International Joint Conference on Neural Networks IEEE*, pp. 2696–2701, Vancouver, BC, Canada, July 2006.

[4] S. Nagireddy, "A pattern identification approach to block cipher identification," http://www.lantana.tenet.res.in/website/files/thesis/MS/sreenivasuluNR/thesis.pdfMaster Science Dissertion, Indian Institute of Technology Madras, Chennai, India, 2008, http://www.lantana.tenet.res.in/website/files/thesis/MS/sreenivasuluNR/thesis.pdfMaster Science Dissertion.

[5] J. W. Chou, S. D. Lin, and C. M. Cheng, "On the effectiveness of using state-of-the-art machine learning techniques to launch cryptographic distinguishing attacks," in *Proceedings of the 5th ACM Workshop on Security and Artificial Intelligence*, p. 105, Raleigh, NC, USA, October 2012.

[6] X. Hu and Y. Zhao, "One to one identification of cryptosystem using Fisher's discriminant analysis," *International Journal of Networked and Distributed Computing*, vol. 6, no. 3, pp. 155–173, 2018.

[7] L.-T. Huang, Z.-C. Zhao, and Y.-Q. Zhao, "A two-stage cryptosystemscheme based on random forest," *Chinese Journal of Computers*, vol. 41, no. 2, pp. 382–399, 2018.

[8] W.-U. Yang, W. Tao, X. Meng, and L. Jin-dong, "Block ciphers identification scheme based on the distribution character of randomness test values of ciphertext," *Journal on Communications*, vol. 36, no. 4, pp. 1–10, 2014.

[9] L. L. Case, C. E. Cannon, M. Sun, and T. E. Tkacik, "Systems and methods for data encryption," U.S. Patent and Trademark Office, Washington, DC, USA, U.S. Patent No. 9954681B2, 2018.

[10] A. Abidi, C. Guyeux, B. Bouallègue, and M. MacHhout, "Conditions to have a well-disordered dynamics in the CBC mode of operation," in *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, pp. 226–231, Hammamet, Tunisia, October 2018.

[11] V. M. Lomte and A. D. Shinde, "Review of a new distinguishing attack using block cipher with a neural network," *International Journal of Science and Research*, vol. 3, no. 8, pp. 2012–2015, 2014.

[12] W. Augusto Rodrigues de Souza, L. Alfredo Vidal de Carvalho, and J. Antonio Moreira Xexeo, "Identification of N block ciphers," *IEEE Latin America Transactions*, vol. 9, no. 2, pp. 184–191, 2011.

[13] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, pp. 452–459, 2015.

[14] E. Scornet, G. Biau, and J. P. Vert, "Consistency of random forests," *The Annals of Statistics*, vol. 43, no. 4, pp. 1716–1741, 2015.

[15] A. Abdiansah and R. Wardoyo, "Time complexity analysis of support vector machines (SVM) in LibSVM," *International Journal of Computer Applications*, vol. 128, no. 3, pp. 28–34, 2015.

[16] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007, https://authors.library.caltech.edu/7694/.

[17] M. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly, "Assessing generative models via precision and recall," in *Advances in Neural Information Processing Systems*, pp. 5228–5237, 2018, https://arxiv.org/abs/1806.00035.

[18] P. K. Ray, S. Kant, B. Roy, and A. Basu, "Classification of encryption algorithms using the hidden markov model," *Calcutta Statistical Association Bulletin*, vol. 64, no. 3-4, pp. 277–290, 2012.

[19] S. O. Sharif, L. I. Kuncheva, and S. P. Mansoor, "Classifying encryption algorithms using pattern identification techniques," in *Proceedings of the 2010 IEEE International Conference on Information Theory and Information Security, ICITIS 2010*, pp. 1168–1172, Beijing, China, December 2010.