ENCRYPTION MODES IDENTIFICATION OF BLOCK CIPHERS BASED ON MACHINE LEARNING

Ruiqi Xia[1], Manman Li[2] and Shaozhen Chen[2]

[1]Department of Cyberspace Security,
Information Engineering University, Zhengzhou, China
[2]State Key Laboratory of Mathematical Engineering and Advanced Computing,
Kexue Avenue, Zhengzhou, China

ABSTRACT

Encryption modes affect the security of block ciphers. This paper proposes a new approach for identifying encryption modes based on machine learning and feature engineering. In the conditions of random keys and initialization vectors, five encryption modes are used for identification. Each mode is encrypted by several block ciphers. By comparing with previous work, we have overcome the shortcomings and improved the accuracy of identification by about 30% to 40%. The experiments improve the existing results and can effectively help cryptanalysts recover the keys.

KEYWORDS

Machine learning, Cryptography, Block ciphers, Feature engineering, Encryption modes, Ciphers identification

1. INTRODUCTION

Cryptography is broadly used in privacy protection and communication security with the development of computer techniques. It is the main target to recover the keys as well as obtain the plaintext for cryptanalysts who acquire the ciphertext through public channels most of the time. Therefore, how to infer the keys by using ciphertext is very important in modern cryptography [1].

Block cipher is one of the most popular topics in cryptanalysis nowadays, which has many advantages such as efficient encryption and good diffusivity. According to Kerckhoffs's assumption [2], cryptanalysts know the cryptography algorithms and the implementation details. For block ciphers, it is equal to say that cryptanalysts ought to know the species of algorithms and the encryption modes to recover the keys. Thus the premise of keys recovery is to identify the algorithms and the encryption modes.

Due to the rapid development of artificial intelligence, machine learning has been successfully combined with cryptography. The approaches to identify block ciphers from ciphertext have been proposed by many scholars based on machine learning. Dileep et al. [3]set up the bag-of-words models to identify DES, KASUMI and other algorithms who achieved the accuracy of over 70%. Subsequently, Manjula and Chou et al. [4,5] tried to distinguish more algorithms by Support Vector Machine(SVM) and decision tree models, which improved the results remarkably. Then, Mishra and Mello et al. [6,7] used the PART, C4.5 algorithms to identify the ciphers. The identification accuracy reached around 100%. Recently, Sandeep et al. [8] has successfully

DOI: 10.5121/ijnsa.2022.14501

applied deep learning techniques in identifying the cryptography algorithms such as AES and Blowfish.

However, compared with the work above, the research about identifying the encryption modes is relatively less to our best knowledge. In 2016, Cheng et.al [9] studied the identification of encryption modes of block ciphers and they gave a perfect experiment which effectively distinguished the encryption modes in some special conditions. However, there are still several problems to be solved according to the previous work.

1. The identification behaved not well in Cheng's work when keys and the initialization vectors were changed.
2. According to the previous work, the identification of OFB mode and CBC mode were relatively unsatisfactory. They did not consider the CTR mode, either.
3. The methods need to be updated due to the development of deep neural networks and ensemble learning.

Thus in this work we propose another creative identification approach to encryption modes of block ciphers. The work concentrates on five modes including Electronic Code Book (ECB), Cipher Block Chaining (CBC), Cipher Feedback (CFB), Output Feedback (OFB) and Counter (CTR). Four machine learning models are set up and the feature program is based on three randomness indices published by NIST (The National Institute of Standards and Technology). Not only do we solve the problems above, but also improve the identification accuracy further. Our accuracy is totally 30% higher than [9]. The block ciphers we select are 3DES, AES-128, ARIA, PRESENT, and SIMON-32.

The first section is the introduction. The second section briefly introduces the encryption modes. Then we propose our scheme of identification in section 3. The results are shown in section 4. The last section is the conclusion.

2. ENCRYPTION MODES

We present a brief introduction of encryption modes and the conditions of our experiments, which is convenient to understand our work. Encryption modes of block ciphers illustrate how the block ciphers operate in a cryptosystem [10]. There are five modes that are commonly used in reality.

Electronic Code Book (ECB). In this mode, the ciphertext is obtained by direct encryption of the block cipher to each plaintext block. Therefore, ECB mode is deterministic and not safe to some extent.

Cipher Block Chaining(CBC). A uniform initialization vector(*IV*) should be chosen to use this mode first. Ciphertext blocks are generated by applying the block cipher to the *XOR* of the current plaintext block and the previous ciphertext block.

Cipher Feedback(CFB). Like Cipher Block Chaining (CBC), CFB mode also makes use of an initialization vector (*IV*) in the blocks. CFB uses a block cipher as a component of a different or random number generator. The previous ciphertext block is encrypted and the output is *XOR*ed with the current plaintext to create the current ciphertext block.

Output Feedback(OFB). Output Feedback generates a pseudorandom stream by encrypting the initialization vector (*IV*). Each block of the plaintext is encrypted by *XOR*ing it with the appropriate block of the stream.

Counter(CTR). To use CTR mode, a uniform value $ctr \in \{0,1\}^n$ is first chosen. Then a pseudorandom stream is generated by encrypting the block ($ctr + i$) mod 2, which i are viewed as integers.

To guarantee the safety of encryption, the cryptosystem often changes the keys and the *IV*s of CBC, CFB, and OFB modes in reality. Therefore we use random keys and initialization vectors(*IV*s) to encrypt in this work, which is huge progress compared with Cheng's work.

3. FUNDAMENTALS FOR ENCRYPTION MODES IDENTIFICATION

The experiments aim to identify five encryption modes of block ciphers using random keys and *IV*s. The ciphertext is encrypted by 3DES, AES-128, ARIA, PRESENT, and SIMON-32. Based on machine learning and neural networks, we construct an identification system including feature program and recognition models.

3.1. Design for the model

Figure 1 shows that the identification system mainly consists of three parts, which are obtaining the original datasets, feature program, and classification models. The whole process is composed of the training stage and testing stage. The output of the recognition models is expressed by the rate of precision, recall and accuracy.

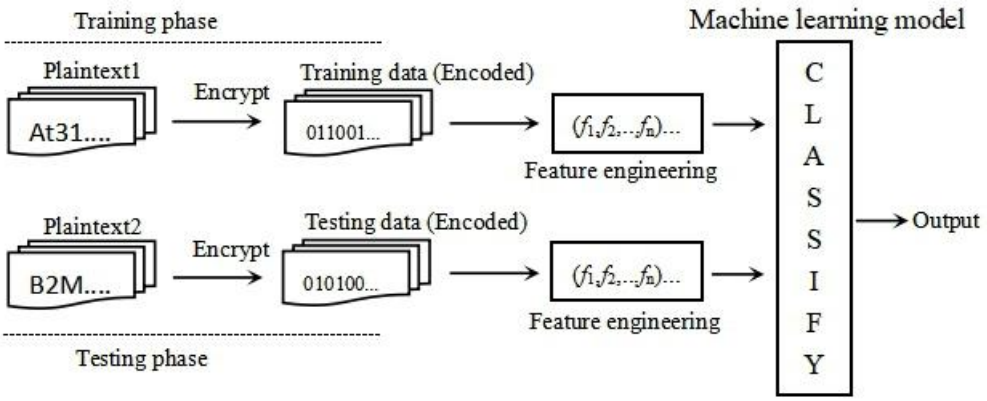

Figure 1. Model of Identification

In our experiments, five block ciphers are encrypted by random keys to get the ciphertexts and each cipher is encrypted in five modes. They are shown in Table 1.

Table 1. The block ciphers used for encryption.

Algorithm	Block length	Structure
3DES	56 bits	Feistel
AES-128	128 bits	SPN
ARIA-128	128 bits	SPN
PRESENT	64 bits	SPN (Lightweight)
SIMON-32	32 bits	Feistel (Lightweight)

3.2. Feature Engineering

The ciphertext seems diffused and random, especially when the keys and *IV*s are unfixed. Feature indices can help find out the characters of the data and help the machine learning models work better. Considering the previous work, many scholars used ASCII codes or bit segments as the feature indices, while we select three statistical indices published by NIST [15]. They are as follows.

Binary Matrix Rank index. The index calculates the ranks of the sub-matrices of the cipher sequence.

Runs index. The index collects the sum of each length of run in the sequence. A run of length k consists of exactly k identical bits and is bounded before and after with a bit of opposite value.

Serial index. The index gets the sum of all sub-sequences of the cipher sequence. A sequence of length m has 2^m sub-sequence. If there is no such sub-sequence, note it with 0.

After encryption, each ciphertext's sequence extracts such three feature indices to make up one feature vector. We save the vectors as the feature files and input them into machine learning models.

3.3. Machine Learning Models

The encryption modes identification is mainly based on machine learning. Machine learning is specifically applied in classification or regression, which has been well applied in cryptography.

Random forest. Random forest is based on the bootstrap resampling technique to extract n samples from the original dataset N. It consists of multi-layers of decision trees which classify the data in order [16].

Fully connected neural network. The network has the input layer, hidden layers and output layer. Such a network connects each neuron between each two successive layers [17].

Feed forward neural network. Such a network has no feedback during the training stage, while the network's structure is similar to deep neural network [18].

One dimensional Convolutional neural network The network consists of four parts named Convolutional layer, Rectified Linear Units layer, Pooling layer and Fully-Connected layer. We revise the input layer for one dimension to fit the format of our feature vectors [19].

4. EXPERIMENTS AND RESULTS

In the experiments we use the Open American National Corpus as the data source of plaintext [20]. The size of plaintext is around 1 GB. Then the plaintext files are divided into 1000 parts, which are about 1.1 MB. We encrypt them with five algorithms in five encryption modes to get the ciphertext. The keys and the *IV*s are changed in each time of encryption.

Subsequently, we extract the three feature indices of the pieces of ciphertext. Combine all of them to form the numerical vectors and save them as feature files. Each file is about 750 KB with 2600 feature vectors or so. Attach each vector with the corresponding labels. We use 0 to 4 to represent the five modes.

Input the feature files into four classification models. Split the datasets into training sets, testing sets and validation sets with the ratio of 4:3:3. Collect the results of each model.

4.1. Results

The results of the classifier are expressed by accuracy, precision and recall[21]. *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 and *FN* (False Negative) represents the number of wrong examples which are sentenced to wrong ones. Hence accuracy means the ratio of all samples which are correctly sentenced in the entire dataset.

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

Precision means the ratio of *TP* in the samples sentenced to be right.

$$precision = \frac{TP}{TP + FP} \tag{2}$$

Recall refers to the proportion of *TP* in the whole right samples.

$$recall = \frac{TP}{TP + FN} \tag{3}$$

The results of four machine learning models are shown in Table 2 to Table 13 as follows.

Table 2. The precision of random forest.

Algorithm(Label)	ECB	CBC	CFB	OFB	CTR
3DES(0)	82%	76%	68%	71%	75%
AES-128(1)	86%	69%	75%	64%	82%
ARIA-128(2)	75%	75%	69%	65%	77%
PRESENT(3)	81%	79%	70%	69%	80%
SIMON-32(4)	82%	70%	76%	74%	73%

Table 3. The recall of random forest.

Algorithm(Label)	ECB	CBC	CFB	OFB	CTR
3DES(0)	68%	49%	41%	51%	40%
AES-128(1)	54%	42%	50%	39%	59%
ARIA-128(2)	66%	44%	48%	42%	55%
PRESENT(3)	59%	51%	49%	47%	60%
SIMON-32(4)	62%	40%	55%	50%	63%

Table 4. The accuracy of random forest.

Algorithm(Label)	ECB	CBC	CFB	OFB	CTR
3DES(0)	80.33%	77.35%	79.51%	74.22%	69.74%
AES-128(1)	83.81%	68.84%	78.52%	52.58%	86.57%
ARIA-128(2)	79.23%	74.06%	80.48%	66.23%	78.95%
PRESENT(3)	78.65%	75.68%	75.13%	60.62%	70.46%
SIMON-32(4)	80.10%	71.29%	81.23%	75.77%	68.13%

Table 2, Table 3 and Table 4 show the properties of random forest model. The precise of each mode of different ciphers ranges from 60% to around 85%, while the recall is between about 40% and 70%. The accuracy of each mode is higher than the average 20%. The results indicate that the random forest can effectively identify different encryption modes in a more strict situation.

Table 5. The precision of fully connected neural network.

Algorithm(Label)	ECB	CBC	CFB	OFB	CTR
3DES(0)	81%	74%	77%	80%	75%
AES-128(1)	83%	78%	79%	82%	76%
ARIA-128(2)	76%	70%	68%	75%	70%
PRESENT(3)	85%	71%	70%	73%	70%
SIMON-32(4)	84%	81%	75%	72%	72%

Table 6. The recall of fully connected neural network.

Algorithm(Label)	ECB	CBC	CFB	OFB	CTR
3DES(0)	69%	59%	47%	58%	56%
AES-128(1)	64%	55%	67%	61%	58%
ARIA-128(2)	51%	47%	60%	55%	49%
PRESENT(3)	65%	62%	59%	68%	55%
SIMON-32(4)	66%	52%	58%	60%	63%

Table 7. The accuracy of fully connected neural network.

Algorithm(Label)	ECB	CBC	CFB	OFB	CTR
3DES(0)	84.24%	74.62%	77.1%	83.24%	80.55%
AES-128(1)	82.05%	85.69%	77.4%	81.18%	82.03%
ARIA-128(2)	70.25%	75.01%	79.24%	65.45%	59.62%
PRESENT(3)	81.42%	79.24%	82.00%	76.32%	70.15%
SIMON-32(4)	86.37%	72.85%	75.66%	70.27%	83.43%

According to Table 5 to Table 7, the fully connected neural network behaves better than random forest. The precision is totally higher than 70% and the recall is also over 50%. The model's accuracy keeps more stable which the gap between the highest and the lowest accuracy is around 10% in each cipher.

Table 8. The precision of feed forward neural network.

Algorithm(Label)	ECB	CBC	CFB	OFB	CTR
3DES(0)	80%	72%	75%	81%	85%
AES-128(1)	85%	81%	82%	71%	79%
ARIA-128(2)	74%	71%	75%	69%	70%
PRESENT(3)	83%	77%	79%	81%	80%
SIMON-32(4)	80%	82%	85%	72%	76%

Table 9. The recall of feed forward neural network.

Algorithm(Label)	ECB	CBC	CFB	OFB	CTR
3DES(0)	65%	66%	61%	62%	66%
AES-128(1)	72%	59%	64%	60%	62%
ARIA-128(2)	64%	64%	60%	59%	62%
PRESENT(3)	68%	62%	67%	58%	74%
SIMON-32(4)	70%	63%	62%	61%	61%

Table 10. The accuracy of feedforward neural network.

Algorithm(Label)	ECB	CBC	CFB	OFB	CTR
3DES(0)	83.52%	81.72%	76.63%	82.57%	81.91%
AES-128(1)	86.42%	75.15%	79.42%	77.28%	80.40%
ARIA-128(2)	81.63%	61.21%	66.56%	59.19%	75.05%
PRESENT(3)	85.67%	74.36%	73.14%	83.32%	70.21%
SIMON-32(4)	88.13%	79.45%	80.34%	73.58%	78.67%

Feed forward neural network improves the precision further. Most of the encryption modes in different ciphers have the precision of over 70% while the recall ranges from 45% to 80%. In addition, the accuracy is mainly between 75% and 85% which is higher compared with the fully connected neural network.

Table 11. The precision of one dimensional convolutional neural network.

Algorithm(Label)	ECB	CBC	CFB	OFB	CTR
3DES(0)	85%	77%	75%	62%	66%
AES-128(1)	83%	71%	74%	70%	76%
ARIA-128(2)	74%	81%	73%	69%	78%
PRESENT(3)	87%	78%	72%	75%	70%
SIMON-32(4)	84%	72%	76%	71%	73%

Table 12. The recall of one dimensional convolutional neural network.

Algorithm(Label)	ECB	CBC	CFB	OFB	CTR
3DES(0)	71%	63%	59%	64%	60%
AES-128(1)	70%	66%	62%	61%	73%
ARIA-128(2)	55%	61%	52%	60%	70%
PRESENT(3)	70%	68%	71%	65%	72%
SIMON-32(4)	72%	62%	56%	61%	69%

Table 13. The accuracy of one dimensional convolutional neural network.

Algorithm(Label)	ECB	CBC	CFB	OFB	CTR
3DES(0)	79.18%	75.74%	76.46%	70.12%	72.34%
AES-128(1)	77.46%	68.26%	72.23%	65.85%	78.84%
ARIA-128(2)	60.34%	62.75%	58.49%	66.50%	70.03%
PRESENT(3)	80.82%	71.37%	77.17%	70.25%	74.34%
SIMON-32(4)	83.10%	66.42%	60.78%	67.94%	72.72%

One dimensional convolutional neural network is less effective compared with other neural networks. Though the precision is mostly higher than 70%, the recall is relatively much lower than other networks. The accuracy decreases to 60% and 80%, which is slightly lower than other models.

4.2. Discussion

Based on the results above, the identification scheme we proposed is verified to be effective and valuable. The classification models can identify the ECB mode better than other modes while OFB and CFB modes are worse. The reason why ECB mode's accuracy is higher may be the low security of ECB mode. It is possible to be distinguished easily.

In addition, the algorithms used for encryption can also influence the results of identification. The precision and recall of ARIA-128 and AES-128 are totally lower than the accuracy of PRESENT and SIMON-32. PRESENT and SIMON-32 belong to the lightweight block ciphers and the structures are simple. This may lead to the high accuracy.

According to the analysis, we suggest that people should use more complex and safer ciphers with the encryption modes except for ECB. Figure 2 shows that our work has remarkably enhanced the accuracy of the identification models compared with the previous work. The accuracy of each encryption mode is higher than 80%, which is higher than Cheng's work for around 30% to 40%. We deduce that the reason may be the feature indices we choose. The three statistical indices extract more information from the ciphertext than the traditional ways.

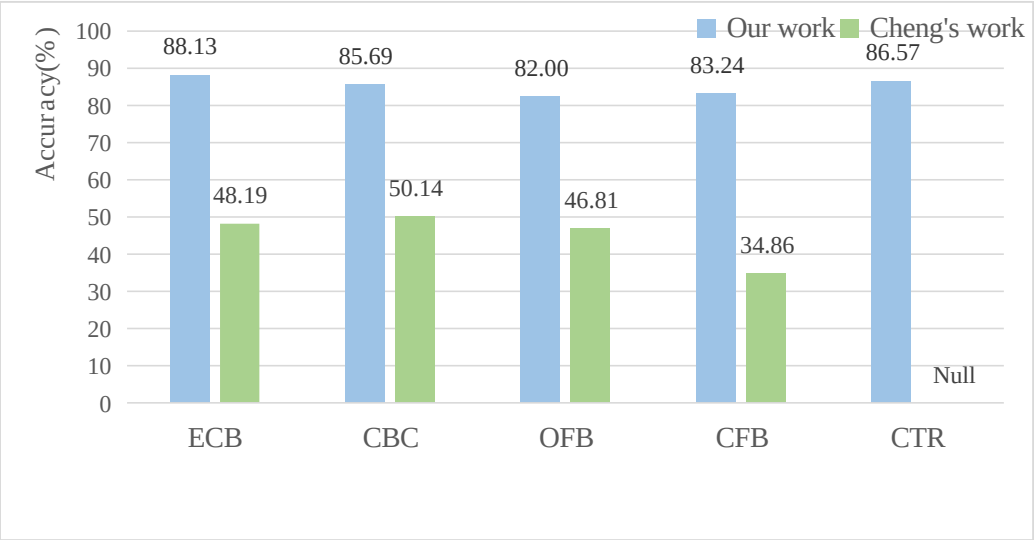

Figure 2. The comparison of identification accuracy.

5. CONCLUSIONS

It is necessary to improve the previous work on encryption modes identification. Cryptanalysts should be aware of block ciphers as well as the encryption modes if they tend to recover the keys from the ciphertext [22]. Our work helps them identify the encryption modes in the conditions of random keys and initialization vectors(*IV*s) to make the keys recovery more effectively. The accuracy of each encryption mode is enhanced by about 30%. We will apply such technique to more general situations in cryptography and attempt to identify more cryptographic objects or modes efficiently.

ACKNOWLEDGEMENT

Thanks my fellows in the State Key Laboratory of Mathematics and Advanced Computing! This paper is supported by Open Fund Project of the State Key Laboratory of Mathematical Engineering and Advanced Computing (No. 2019A08).

REFERENCES

[1] Coron,Mohammed. Muzammil H, Analysis on Contribution of Cryptography and Steganography in Protecting Information in Diverse Environments. *Proceedings of Third International Conference on Communication, Computing and Electronics Systems: ICCCES 2021*. Vol. 844. Springer Nature, 2022.

[2] RS. Mrdovic and B. Perunicic, Kerckhoffs' principle for intrusion detection, *Networks 2008 - The 13th International Telecommunications Network Strategy and Planning Symposium, 2008*, pp. 1-8, doi: 10.1109/NETWKS.2008.6231360.

[3] Dileep, A. D., & Sekhar, C. C. (2006, July). Identification of block ciphers using support vector machines machines. In The *2006 IEEE International Joint Conference on Neural Network Proceedings* (pp. 2696-2701). IEEE.

[4] Manjula, R., & Anitha, R. (2011, January). Identification of encryption algorithm using decision tree. In *International Conference on Computer Science and Information Technology* (pp. 237-246). Springer, Berlin, Heidelberg.

[5] Chou, J. W., Lin, S. D., & Cheng, C. M. (2012, October). 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* (pp. 105-110).

[6] Barbosa, F., Vidal, A., & Mello, F. (2016). Machine learning for cryptographic algorithm identification. *Journal of Information Security and Cryptography (Enigma)*, 3(1), 3-8.

[7] De Mello, F. L., & Xexeo, J. A. M. (2016). Cryptographic algorithm identification using machine learning and massive processing. *IEEE Latin America Transactions*, 14(11), 4585-4590.

[8] Pamidiparthi, S., & Velampalli, S. (2021). Cryptographic algorithm identification using deep learning techniques. In *Evolution in Computational Intelligence* (pp. 785-793). Springer, Singapore.

[9] Tan Cheng, Yifu Li, and Shan Yao. A novel identification Approach to Encryption Mode of Block cipher. *4th International Conference on Sensors, Mechatronics and Automation*, Zhuhai, China. 2016.

[10] AbdRogaway Phillip. Evaluation of some blockcipher modes of operation. *Cryptography Research and Evaluation Committees (CRYPTREC) for the Government of Japan* (2011).

[11] Patil Priyadarshini, et al. A comprehensive evaluation of cryptographic algorithms: DES, 3DES, AES, RSA and Blowfish. *Procedia Computer Science 78* (2016): 617-624.

[12] Li Wei, Dawu Gu, and Juanru Li. Differential fault analysis on the ARIA algorithm. *Information Sciences 178.19* (2008): 3727-3737.

[13] Katagi Masanobu, and Shiho Moriai. Lightweight cryptography for the internet of things. *sony corporation 2008* (2008): 7-10.

[14] Diffi.Ray Beaulieu, Douglas Shors, et al. 2015. The SIMON and SPECK lightweight block ciphers. In *Proceedings of the 52nd Annual Design Automation Conference*. Association for Computing Machinery, New York, NY, USA, Article 175, 1–6. DOI:https://doi.org/10.1145/2744769.2747946.

[15] ElGaPareschi Fabio, Riccardo Rovatti, and Gianluca Setti. On statistical tests for randomness included in the NIST SP800-22 test suite and based on the binomial distribution. *IEEE Transactions on Information Forensics and Security 7.2* (2012): 491-505.

[16] Biau G´erard, and Erwan Scornet. A random forest guided tour. *Test 25.2* (2016): 197-227.

[17] Schwing Alexander G., and Raquel Urtasun. Fully connected deep structured networks. *arXiv preprint* arXiv:1503.02351 (2015).

[18] Cheng Shuhui, et al. TWD-SFNN: Three-way decisions with a single hidden layer feed forward neural network. *Information Sciences 579* (2021): 15-32.

[19] Zhang Ziyan, et al. A Haze Prediction Method Based on One-Dimensional Convolutional Neural Network. *Atmosphere 12.10* (2021): 1327.

[20] Po.Ide Nancy. The American national corpus: Then, now, and tomorrow. *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages, Summerville, MA. Cascadilla Proceedings Project*. 2008.

[21] Jizat Jessnor Arif Mat, et al. Evaluation of the machine learning classifier in wafer defects classification. *ICT Express 7.4* (2021): 535-539.
[22] Blackledge Jonathan, and Napo Mosola. Applications of Artificial Intelligence to Cryptography. (2020).

AUTHORS

Ruiqi Xia, Graduate student at the Institute of Cyberspace Security, Information Engineering University.

Manman Li, Ph.D. of State Key Laboratory of Mathematical Engineering and Advanced Computing.

Shaozhen Chen, Professor of State Key Laboratory of Mathematical Engineering and Advanced Computing.

APPENDIX

The main code and data for this paper is available at
https://github.com/xrq2590/Encryption-modesidentification.