Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 29 (2026), No. 2-B, pp. 769–777
DOI : 10.47974/JDMSC-2527

Cryptanalysis of ciphers using machine learning

Pooja Bagane [§]
Sonali Kothari [†]
Ankur Goyal *
Chahak Sengar [‡]
Sumedh Dongre [@]
Siddharth Prabhakar [#]
Department of Computer Science and Engineering
Symbiosis Institute of Technology
Symbiosis International (Deemed University)
Lavale
Pune 412115
Maharashtra
India

Abstract

This paper focuses on the recognition of cipher encryption keys via machine learning, specifically Vigenère and Advanced Encryption Standard (AES) ciphers. It does this by analyzing pairs of plaintexts with their associated ciphertexts. It is a study that attempts to achieve a classification model for the task of predicting the encryption key between pairs of plaintexts and ciphertext without knowledge of the encryption key. A set of different plaintexts, ciphertext, and keys have been used to train the model. The results had proved the success of machine learning over existing encryption techniques, also illuminating their potential weaknesses and further giving impetus to the field of cryptanalysis.

Subject Classification: *Primary 94A60, Secondary 68T05.*

Keywords: *Cryptanalysis, Advanced encryption standard cipher, Vigenère cipher, Machine learning, Random forest classifier, Key identification.*

[§] *E-mail:* **pooja.bagane@sitpune.edu.in**
[†] *E-mail:* **sonali.kothari@sitpune.edu.in**
* *E-mail:* **ankur_gg5781@yahoo.co.in (Corresponding Author)**
[‡] *E-mail:* **chahak.sengar.btech2021@sitpune.edu.in**
[@] *E-mail:* **sumedh.dongre.btech2021@sitpune.edu.in**
[#] *E-mail:* **siddharth.prabhakar.btech2021@sitpune.edu.in**

© Taru Publications

1. Introduction

As data security assumes prominence in the modern digital world, the classical encryption techniques present an insight to the science of cryptography: namely, the Advanced Encryption Standard (AES) and Vigenère ciphers. However, the earlier techniques remain susceptible to some of the more modern analysis techniques. This paper looks into whether machine learning is useful in classifying and recovering encryption keys from plaintext and ciphertext pairs.

We have generated our existing datasets for the analysis via the Common Passwords List available on Kaggle. It contains 14 million passwords. This choice gives us some advantages: the randomness of passwords, their short length reduces computational complexity, and single-word content minimizes sequence dependency issues. This approach helps in identifying patterns that can be used for key identification without previous knowledge about the keys themselves. This work points to the vulnerabilities of classical ciphers and shows the potential of machine learning to heighten our understanding about the security of cryptography.

2. Literature Review

Zheng, Zhang, and Wang [1] boosted Bayesian Key Search with the Bayes-UCB strategy to the setting of attacking Speck 32/64, and more than doubled the chance of successfully recovering the key. Analysis showed that deep learning outperforms classical cryptanalysis in theory. Shafique et al. [2] introduced a ML-based image encryptor, which only encrypt the high/medium level blocks, which can reach 97.4% accuracy and computational efficiency.

Kumar, Tanwar, and Kumar [3] proposed a deep learning methodology for chosen-plaintext attack on AES and they were able to attain 67% precision, 99% recall and 80% F1-score, however, they proclaimed machine learning is so far not even capable of breaking AES securely. Sarkar et al. [4] summarized the AES and its error diffusion properties and recommend parity and redundancy schemes to diffuse input manipulated on cipher text byte for alleviation against single bit input manipulation. Tolba [5] stressed that encryption was a mechanism to protect data and that cryptanalysis was necessary to quantify a cipher's strength based on a recognizably increasing effort with which mathematical errors could be identified or plaintext recovered.

Shafique et al. [2], introduced a real-time machine learning based image encryption that reaches an accuracy of 97.4% for sub-second speed, which is enough for real-time, safe communications. Naik [6] had a closer look at the role of machine learning in the design of lightweight block ciphers for IoT, concluding that performance and design can benefit, but ends with computational efficiency and security as a challenge.

Jeong, Ahmadzadeh, and Moon [7] analysed five block ciphers using deep learning approach and recommended enhancements like larger keys, sentence level encryption and RNNs. Larger datasets result in a greater success of the attack, but current models are so far inefficient. Singh, Sivangi, and Tentu [8] investigated cryptanalysis with neural networks, using raw encrypted data to identify weaknesses in a cipher, and even to try to extract encryption. They demonstrated that AI has potential in modern cryptanalysis. Kim et al. [9] developed an efficient deep learning cryptanalysis for S-DES and S-AES and claimed better accuracy and a smaller number of parameters. S-AES, was more difficult to master and had an enhanced level of security.

Kim et al. [10] proposed estimating mutual information (MI) using neural networks to detect information leakage in encryption systems, and also found that MI can defeat even computationally secure systems. Tarawneh [11] presented the significance of encryption and cryptanalysis and described cryptanalyst attacks and mathematical algorithms used by cryptanalysts for obtaining plaintext.

Dalton and Stamp [12] treated WWII ciphers with an accuracy higher than 97% by means of ML/DL. Even in some cases, traditional ML could be better than DL. The similarity of the cipher made classification difficult. Ithape and Purushothama [13] introduced ML models (DNN, RF, logistic regression) to recognize encryption algorithms in multiple cipher modes (ECB, CBC, CFB, CTR) with an accuracy greater than 65%, as opposed to classical statistical methods. Saha and Pathak [14] presented the use of ML/DL in cryptanalysis to get rid off classical limits, stating that such techniques are able to lower computational differences and enhance the efficiency of encryption breaking [15][16].

3. Methodology

Data Generation: The datasets have been created by using the Common Passwords List from Kaggle. The original dataset has about 14 million passwords. The size of the original dataset has been reduced in order to manage processing time and maintain meaningful patterns in the

data for this project. This dataset selection provides a simple advantage of balanced randomness and length constraints-appropriate for the needs of cryptographic models without raising efficacy in computation. For each encryption scheme (Vigenère and AES), prepared datasets containing encrypted cipher-texts next to their passwords were constructed. Each dataset has two main attributes: Passwords which are used to encrypt data and Ciphertexts that is the encrypted output of the cipher. The goal is to predict the encryption key used in the process.

Data Preprocessing: Proper preprocessing is one of the critical points for model accuracy and robustness. Key steps include:

- Handling Missing or Unwanted Data: The columns for password and ciphertext were checked for missing (NaN) values and non-string entries.
- Feature Engineering: Text data such as passwords or ciphertext cannot be fed directly into the learning machine.

The following steps were used to convert them into a numerical feature space:

- Character to ASCII Mapping: Each character in a password or ciphertext was mapped to its corresponding ASCII values.
- Fixed-Length Truncation and Padding: A limit of 10 was set. All the text under 10 characters were padded with zeros. Everything over 10 characters was truncated at the first 10 characters.
- Feature Combination: The password and ciphertext features were stacked horizontally to create the feature matrix for the models' input.
- Data Splitting and Normalization: Stratified splits were generated for all datasets with 80:20 divisions so that the presence of each class was adequate.
- Standardization: Feature standardization is performed using the standard scaler for improved convergence and performance of the model.
- Model Selection and Evaluation: Multiple machine learning models were explored, including Random Forest Classifier Logistic Regression, Ada Boost Classifier, Gradient Boosting Classifier, Gaussian Naive Bayes, Ridge Classifier.

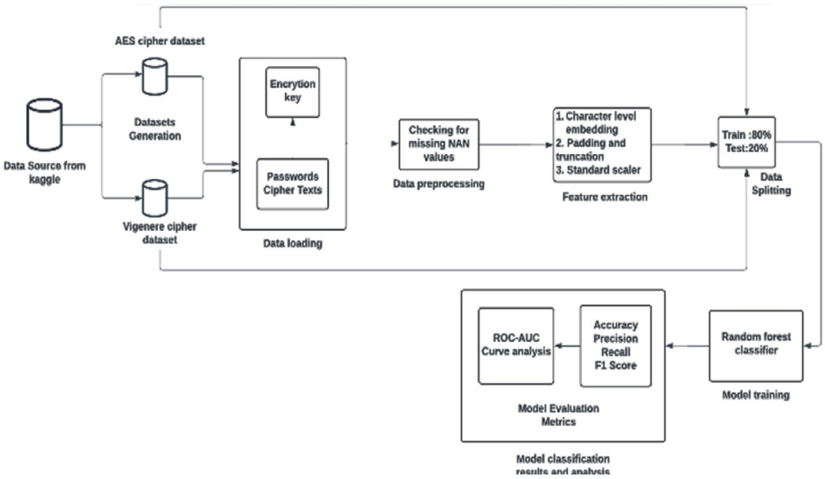

Figure 1

System Architecture

AUC-ROC are metrics used for performance assessment of the proposed models. Multiclass ROC curves were plotted using One-vs-Rest method. A detailed analysis of the existing approach is shown in Figure 1. The system architecture diagram highlights the flow of steps taken to process the dataset and classify the ciphers using random forest classifier.

4. Implementation Details

Data Loading and Preprocessing: Firstly, dataset is imported using pandas and NaN values were replaced with empty strings and converted into a string data type to avoid errors while processing.

Feature Extraction & Data Splitting: ASCII-based feature extraction both on passwords and on the ciphertexts. After application of feature extraction, feature matrix is prepared and then data is split up into corresponding training and test sets followed by standardization.

Model Training and Predictions: Random Forest Classifier was trained in order to make predictions on test data.

Performance Metrics: The model's efficiency was evaluated via different metrics like Accuracy, Precision, Recall, F1 Score.

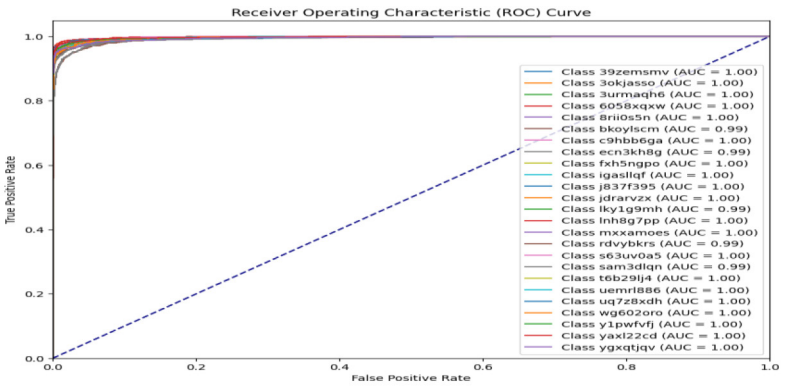

Figure 2

ROC-AUC Curve (Vigenère Cipher)

5. Discussion and Conclusion

A summary of performance comparison between AES and Vigenère ciphers is noted in Table 1. Both the ciphers obtained high precision, recall, F1 scores where the power of the machine learning models was highly effective in predicting the cryptographic keys. Vigenère cipher outperformed the AES model with an accuracy.

All the performance metrics depicts similar trends. Vigenère, however, shows a marginally higher precision, recall, and F1 score over AES. Perhaps the reason again is that the models are picking up the simpler substitution-based encryption better due to easier patterns resulting from Vigenère

In addition to these performance metrics, ROC-AUC curves are drawn for both the ciphers in order to represent their classification abilities. In both of the models, as can be seen from Figure 2 for Vigenère and Figure

Table 1

Comparison for testing cipher's performance

Dataset	AES	VIGENERE
Accuracy	91.31%	96.06%
Precision	91.85%	96.07%
Recall	91.31%	96.06%
F1 Score	91.24%	96.06%

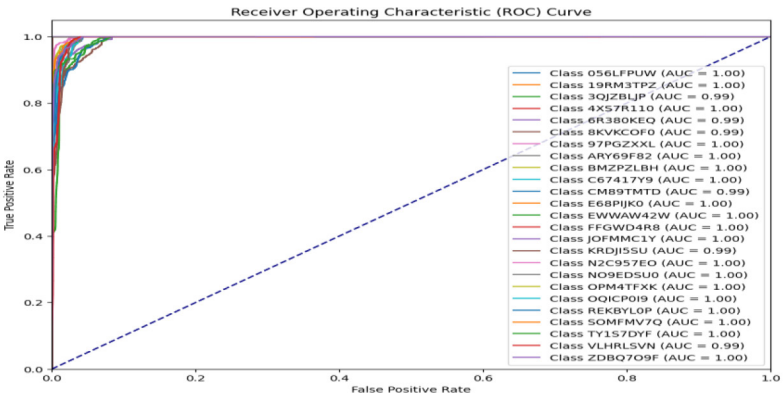

Figure 3

ROC-AUC Curve (Advanced Encryption Standard Cipher)

3 for AES, achieved values were at the boundaries of 0.99 and 1.00, meaning the classification was excellent. Thus, these results have shown that the model distinguished good keys and bad keys with minimal false positives and negatives.

Differences in performance between the two ciphers then suggest that the complexity of encryption indeed lies in predicting accuracy. For this reason, the more complex AES, being a symmetric block cipher with several rounds of transformation, presents more complexity, but the model did well, achieving greater than 91% on all metrics. The Vigenère cipher, being simple and presenting its fixed key-length, periodic substitution pattern, offered more predictable outputs, and so outperformed the others in prediction slightly more.

These results authenticate the efficiency of applying machine learning models in key prediction along different encryption schemes. On the other hand, results indicate that both simple and complex ciphers can be improved by using this predicting model, with room for further improvements like tuning hyperparameters or implementing a sophisticated neural network to further enhance its performance.

This methodology can be extended to other cryptographic schemes like RSA or DES. Additionally, integrating neural network architectures (e.g., LSTM) may improve performance by capturing sequential dependencies in passwords and ciphertexts. Apart from that, Transformer models which have gained significant attention for their state-of-the-art performance on sequence-based tasks-would potentially be useful in handling more expansive datasets and complex cryptographic methods.

References

[1] W. Zheng, L. Zhang, and Z. Wang, "Theoretical explanation and improvement of deep learning-aided cryptanalysis," *IACR Cryptol. ePrint Arch.* (2024).

[2] A. Shafique, A. Mehmood, M. Alawida, M. Elhadef, and M. U. Rehman, "A fusion of machine learning and cryptography for fast data encryption through the encoding of high and moderate plaintext information blocks," *Multimedia Tools Appl.* (2024), doi: 10.1007/s11042-024-18959-6.

[3] K. Kumar, S. Tanwar, and S. Kumar, "Deep-learning-based cryptanalysis through topic modeling," *Eng. Technol. Appl. Sci. Res.*, vol. 14, no. 1, pp. 12524–12529 (2024), doi: 10.48084/etasr.6515.

[4] B. Sarkar, A. Saha, D. Dutta, G. De Sarkar, and K. Karmakar, "A survey on the Advanced Encryption Standard (AES): A pillar of modern cryptography," *Int. J. Comput. Sci. Mobile Comput.*, vol. 13, no. 4, pp. 68–87 (2024), doi: 10.47760/ijcsmc.2024.v13i04.008.

[5] Z. Tolba, "Cryptanalysis and improvement of multimodal data encryption by machine-learning-based system," arXiv preprint arXiv:2402.15779 (2024).

[6] M. Naik, "Machine learning-based lightweight block ciphers for resource-constrained Internet of Things networks: A review," *Int. J. Elect. Comput. Eng.*, vol. 14, no. 3, pp. 2896–2907 (2024), doi: 10.11591/ijece.v14i3.pp2896-2907.

[7] O. Jeong, E. Ahmadzadeh, and I. Moon, "Comprehensive neural cryptanalysis on block ciphers using different encryption methods," *Mathematics*, vol. 12, no. 13, pp. 1936 (2024), doi: 10.3390/math12131936.

[8] A. Singh, K. B. Sivangi, and A. N. Tentu, "Machine learning and cryptanalysis: An in-depth exploration of current practices and future potential," *J. Comput. Theories Appl.*, vol. 2, no. 1, pp. 27–42 (2024), doi: 10.62411/jcta.9851.

[9] H. Kim, S. Lim, Y. Kang, W. Kim, D. Kim, S. Yoon, and H. Seo, "Deep-learning-based cryptanalysis of lightweight block ciphers revisited," *Entropy*, vol. 25, no. 7, pp. 986 (2023), doi: 10.3390/e25070986.

[10] B. D. Kim, V. A. Vasudevan, J. Woo, A. Cohen, , R. G. D'Oliveira, T. Stahlbuhk, and M. Médard, "Crypto-Mine: Cryptanalysis via mutual information neural estimation," in *Proc. IEEE Int. Conf.*

Acoust., Speech Signal Process. (ICASSP) (2024), doi: 10.1109/ICASSP48485.2024.10447256.

[11] M. Tarawneh, "Cryptography: Recent advances and research perspectives," *IntechOpen* (2024), doi: 10.5772/intechopen.111847.

[12] B. Dalton and M. Stamp, "Classifying World War II era ciphers with machine learning," *Cryptologia*, pp. 1–31 (2024), doi: 10.1080/01611194.2024.2304888.

[13] S. Ithape and B. R. Purushothama, "Identification of encryption method for block ciphers using machine learning methods," in *Proc. IEEE India Conf. (INDICON)* (2023), doi: 10.1109/INDICON59947.2023.10440785.

[14] A. Saha and E. Pathak, "A study of machine learning techniques in cryptography for cybersecurity," *Amer. J. Electron. Commun.*, vol. 1, no. 4 (2021), doi: 10.15864/ajec1404.

[15] K. A. Hassoon and H. R. Yassein, "A new multi-dimensional public key cryptosystem via KAH-Octo algebra," *J. Discrete Math. Sci. Cryptogr.*, vol. 28, no. 2, pp. 349–354 (2025).

[16] V. Bhatnagar, A. Jain, R. Gupta, and S. Mehta, "Descriptive analysis of COVID-19 patients in the context of India," *Journal of Interdisciplinary Mathematics*, vol. 24, no. 3, pp. 489–504 (2021).

Received May, 2025