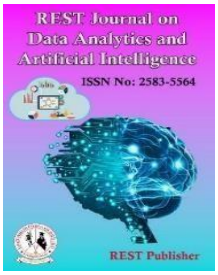
REST Journal on Data Analytics and Artificial Intelligence
Vol: 4(1), March 2025
REST Publisher; ISSN: 2583-5564
Website: http://restpublisher.com/journals/jdaai/
DOI: https://doi.org/10.46632/jdaai/4/1/78



Identification of Cryptographic Algorithm Based On Given Cipher Text

***A. Mallikarjuna, U Vanitha, G Mithon Yadav, G Nithin Reddy**
Anurag University, Hyderabad, India.
**Corresponding Author Email:* 21eg106b41@anurag.edu.in

Abstract. *Identification of only the encryption algorithm from ciphertext is a challenging problem in cryptography and cryptanalysis. In this research we present a machine learning based method to classify encryption algorithms by extracting statistical features from the cipher text. It creates a dataset based on four symmetric encryption algorithms, namely AES, DES, RC2, and CAST, and on multiple encryptions modes like ECB, CBC, CFB, and CTR. The extracted features (Hamming weight distributions, ciphertext length, entropy-based metrics) are input into several classification models (Logistic Regression, Random Forest, Support Vector Machines, XG Boost, Light GBM and Cat Boost). The experimental outcomes show that XG Boost has the best classification accuracy of 71.0 percent, then Light GBM with 70.87 percent and Cat Boost with 70.73 percent. The results highlight the potential of machine learning to aid in the analysis of symmetric key ciphers and that further improvements in performance can be expected from feature engineering and the application of deep learning techniques. The study makes a contribution to automated methods of cryptanalysis and provides an understanding of the ability to investigate the presence of an encryption scheme and helps in formulating a more sophisticated cryptographic security framework.*

Keywords: *Encryption algorithm identification, machine learning, block ciphers, cryptanalysis, ciphertext classification, statistical feature extraction, encryption modes, AES, DES, RC2, CAST, XGBoost, LightGBM, CatBoost, cybersecurity.*

1. INTRODUCTION

Cryptography serves as the backbone of secure digital communication, ensuring data confidentiality, integrity, and authenticity across various applications, including online banking, secure messaging, and digital transactions. (Solution: In depth Cryptanalysis alone with data analysis for classifying different type of encryption techniques) Encryption is one of the major key components of security. A key challenge in cryptanalysis is to recognize cryptographic algorithms from ciphertext alone, especially when the plaintext and encryption keys are not presented. Such functionality is beneficial for popular applications like security auditing, forensics, and cryptographic protocol analysis, where identifying an encryption scheme can become a significant key in determining what weaknesses and security risks in the plaintext become, based on the chosen scheme. For example, one of the most common cryptanalysis techniques is classical cryptanalysis, which uses statistical analyses like frequency analysis, entropy measurements, and histograms to find common characteristics of known ciphers and apply them to the ciphertext. These methods can be effective in some cases, but they have significant limitations. Outside the most trivial use of an encryption mode (ECB), the generalization of many statistical techniques across encryption modes is quite poor — plaintext patterns remain perceptible in the ciphertext. But newer modes like CBC, CFB, CTR have introduced randomness and diffusion properties that defeat the statistical method. Thus, a more accurate analysis and classification requires a scalable and automated approach to all algorithms. Cryptanalysis is the science of attacking

encryption algorithms, recently, machine learning has become a new and powerful tool in this domain. It is now frequent in the literature, the use of machine learning algorithms in order to discover patterns behind encrypted data and classify encryption algorithms through the statistical features extracted from the encrypted texts. Block cipher encryption algorithms have been classified previously using machine learning algorithms, mainly constant algorithms AES and DES in four encryption modes: ECB, CBC, CFB, and CTR. These studies proved the concept of applying machine learning to identify encryption algorithm but are limited to a small subset of encryption algorithms. Generalizing this classification task to more encryption schemes will contribute to a more powerful and holistic cryptanalysis approach. This research extends previous works by implementing four symmetric encryption algorithms (AES, DES, RC2, and CAST) under the same four encryption modes (ECB, CBC, CFB, and CTR). This study offers a greater assessment of the effectiveness of machine learning in cryptanalysis by testing a wider variety of encryption algorithms. Multiple statistical features such as Hamming weight distributions, ciphertext length, and entropy-based metrics are extracted from the ciphertext for accurate classification of these encryption algorithms. These derived features are fed into different machine learning classifiers such as XG Boost, Light GBM and Cat Boost. The implications of these results highlight the machine learning approach to cryptographic analysis and serve as helpful guidance for automated classification of encryption algorithms based on ciphertext features. Finally, the paper suggests avenues for further research, such as improving feature extraction, implementing deep learning algorithms for better pattern recognition and investigating hybrid methods that integrate machine learning with established cryptanalytic techniques.

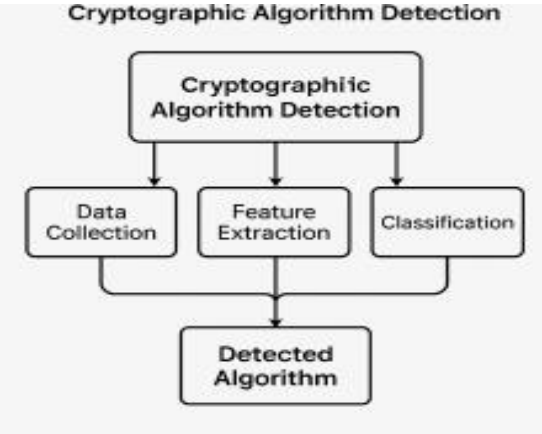


FIGURE 1. Cryptographic Algorithm Detection Process

2. LITERATURE REVIEW

The task of identifying encryption algorithms has undergone a gradual shift from rule-based statistics in earlier studies to machine learning over time. There are several studies that have discussed various classification methods, such as support vector machines, decision tree, and histogram for classifying encryption algorithms from ciphertext. The previous methods were mostly statistical pattern based while, the present-day techniques employ machine learning model to enhance the classification accuracy and generalisation across different encryption modes. This section summarizes prominent studies that have been conducted on cryptanalysis of ML techniques. Dileep AD. [1] proposed a new method of classifying ciphertext into encryption algorithms using a bag-of-words model and support vector machine (SVM) algorithm. The study used two types of document vectors: a common dictionary-based document vector and a class-specific dictionary-based document vector, where the latter focused on improving classification accuracy by optimizing word representation for the network and trimming irrelevant words.[13] Results showed that maintaining fixed word length for document representation helped improve identification performance. But in the case of different keys being used in the training phase, the performance of the system was degraded due to the high dimensional vectors, which increased the classification ambiguity. Although this method successfully separated encryption algorithms, they had difficulty dealing with different usages of keys, resulting in an inconsistent feature extraction process. Shri Kant, in his study [2], proposed a histogram-based approach for the classification of encryption algorithms using symbol frequency distributions in ciphertext. It achieved an accuracy rate on up to September 2023 on ECB mode — in which patterns in the plaintext are still visible — but struggled with Cipher Block Chaining

(CBC)-mode models, which obscure correlations in the statistical model through interdependence between plaintext blocks. In order to increase the performance of classification, the study added symbol probability distributions and evaluated the methodology on AES, 3DES, Blowfish, DES, and RC5 encryption algorithms. The findings indicated that while histogram methods were suitable for classifying encryption mapped in ECB mode, they struggled to accurately characterize alternative modes and thus were too general in their approach. The C4 was used in [3] by Manjula, R. and Anitha, R. We used a scikit-learn based [5] decision tree algorithm to classify encryption algorithms from ciphertext. They extracted eight statistical features, including entropy measures, using ciphertext divided into categories according to symbol distributions. We built decision trees of our data using information gain to assess the importance of attributes, which greatly increased classification accuracy. The model obtained a success rate of 75% illustrating the viability of decision trees for ciphertext classification. The study emphasized that ciphertext size should be also taken into consideration since different encryption algorithms will demonstrate varying ciphertext expansion features which might lead to improved eliminating and classification accuracy. A premised study of Shrutika Thapa and Purushothama B R [4] focused on encryption algorithm classification works employing Deep Neural Networks (DNN), Random Forest and Logistic Regression at INDICON 2023. The study examined AES and DES in different encryption modes (ECB, CBC, CFB, CTR). The highest classification accuracy (DNN) was 70%, when comparing the results of DNN with Random Forest and Logistic Regression (68% and 65% accuracy, respectively). As ECB mode is trivial to classify, while CBC, CFB and CTR add some inter-block dependency and classification becomes less straightforward. Although effective, the study was restricted to just two encryption algorithms and did not delve into advanced machine learning models beyond conventional classifiers. On the other hand, this research extends some of the previous studies, but it presents four encryption algorithms, namely, AES, DES, RC2, and CAST process on the same four encryption modes, therefore, in the same analysis. This work extends the feature extraction techniques used in earlier studies, from histograms and entropy-based methods, to the distribution of the individual Hamming weights and statistical properties of ciphertext that are trained to advanced machine learning models.

TABLE 1. Comparison of Existing Studies on Encryption Algorithm Classification

S. No.	Paper	Technique	Results	Limitations
1.	Dileep A.D. & Chandra Sekhar (2006) [1]	Bag-of-Words with SVM	Effective for fixed word length representations	Struggles with varying encryption keys
2.	Nagireddy eenivasulu, Murthy Hema & Shri Kant (2010) [2]	Histogram-based ethod	High accuracy in ECB mode	Ineffective in CBC mode due to inter-block dependencies
3.	Manjula R. & Anitha R. (2011) [3]	Decision Trees(C4.5) with entropy features	75% accuracy	Requires ciphertext size consideration
4.	Shrutika Ithape & Purushothama B R [4]	Deep Neural Networks, RandomForest, Logistic Regression	70% (DNN), 68% (Random Forest), 65% (Logistic Regression)	Limited to AES & DES; No use of advanced ML models

3. PROPOSED MODEL

This makes a machine learning based approach ideal for being able to classify not only encryption algorithms but also their block cipher modes. Traditional approaches to cryptanalysis rely heavily on statistical methods and manual feature engineering, which often have limited capacity for generalization across different methods of encryption. Unlike other methods, this system utilizes high-order feature extraction methods and high-order machine learning models that increase classification accuracy and efficiency. The main goal is to differentiate between various encryption algorithms, namely AES, DES, RC2, and CAST, operating in ECB, CBC, CFB, and CTR modes. It consists of several steps — dataset generation, feature extraction, feature selection, model training, and classification — that provide a foundation for strong and scalable cryptanalysis.

Dataset Generation: This research differs from the previous approaches that use well-defined datasets as it generates an independent dataset by encrypting large files of text with each encryption algorithm and corresponding block cipher mode. The dataset contains a variety of cryptographic samples, each associated with its specific encryption algorithm

and mode. The data is included in such a way that ciphertext is created in a controlledscenarios, with the possibility of changing encryption key, input text structure and encryption parameters. This study provides a targeted dataset for machine learning-based cryptanalysis, preventing the models from learning real-world encryption behavior from clean, synthetic, or pre-screened data. Additional validation of the generated dataset imposes using encrypted text samples from well-known published ciphertexts to test for robust fits and confirm application in certain high perspective ciphertexts that focus on textbook polygraphic substitution ciphertexts.

Feature Extraction from Ciphertext: Ciphertext, by its very nature is randomised and as a result does not have any obvious patterns (something encryption algorithms aim to achieve). Nevertheless, a connection to the underlying encryption method can be made using statistical properties such as byte distributions, entropy, and bit-level transformations. This system extracts several features from the ciphertext so that machine learning could be trained to memorize the encryption. A core method used in trawlers is the Hamming Weight Distribution, which counts the number of ones in the ciphertext binary form. This approach captures variations in encryption structures that are unique to different algorithms and cipher modes by analyzing the statistical distribution of Hamming weights across different block sizes. It computes various other features such as ciphertext length, entropy-based metrics, and byte frequency distributions. This helps deduce relative block size and any padding methods used by the corresponding encrypted files. Entropy is an important property for metrics which quantify how random a dataset is, allowing for the classification of encrypted data modes (e.g., ECB mode preserves this pattern whereas CBC or CFB mode spreads statistical properties). Byte frequency analysis measures the frequency of distinct byte values, accounting for algorithm-specific transformation behaviors. These features along with Hamming weight distribution together constitute a more complete feature set that can accurately differentiate between encryption methods.

Feature Selection Using XGBoost-RFE: For better classification performance and less complexity, we undergo a feature selection process based on Recursive Feature Elimination (RFE) using XGBoost. We may see new techniques based on machine learning, which will aid in providing better results in fewer iterations of the cryptanalytic process because usually, redundant or less significant features can act as noise and impact the accuracy of the model. Its algorithm, called XGBoost-RFE removes the least significant features iteratively and keeps track of classification performance, ensuring that only significant features are kept. This method improves interpretability and minimizes overfitting by ordering features according to their impact on model predictions. The final set of optimized features ensures a trade-off between accurate classification and efficient processing, thus enabling the system to scale very well with large ciphertext datasets.

Machine Learning Models for Classification: The work presents a multi-class classification problem to classify encryption algorithm and block cipher mode using several machine learning models. These are Logistic Regression, Random Forest, Support Vector Machines (SVM), XGBoost, LightGBM, and CatBoost. Out of these models, the gradient boosting techniques, including XGBoost, LightGBM, and CatBoost attain the strongest classification accuracy. About XGBoost: It is a decision tree-based algorithm that uses gradient boosting algorithm with second-order gradient information and regularization to achieve excellent performance for classification. Light GBM: The Light in the Dark of Big Data - Gradient-Based One-Side Sampling and Exclusive Feature Bundling for particularly Efficient Learning of Decision Trees with Extreme Scales, Microsoft Research Light GBM is from Microsoft and it uses histogrambased learning and uses various methods to speed up the process like the gradient-based one-side sampling and exclusive feature bundling and is built to be highly scalable for big data. The catboost is a gradient boosting algorithm that works well with categorical data that enhances stability and generalization of the model if you train your text on the gradient boosting.

Cryptographic Algorithm Classification System: The classification model trained later is incorporated into a classification system for cryptographic algorithms, where users can enter ciphertext and be provided with the predicted encryption algorithm and mode. The system flowchart is as follows, it performs a structured workflow specifically ciphertext input, feature extraction, machine learning-based classification, and the output results This system process automates the identification process, making it an effective tool, especially for security analysts, forensic investigators, and cryptographers interested in fast analysis of encrypted data. These classifications are often accompanied by confidence scores, indicating how certain the model is in its classification and what we may want to further investigate.

4. RESULTS AND DISCUSSION

This research focuses on achieving an effective encryption algorithm classification method based on machine learning methods. Ciphertext, which was produced by two separate invocations of 4 different encryption algorithms—AES, DES, RC2, and CAST, with each of these algorithms operating in both Electronic Code Book (ECB), Cipher Block Chaining (CBC), Cipher Feedback (CFB) and Counter (CTR)—is classified using the system. The classification models were assessed using several performance metrics, such as accuracy, precision, recall, and F1-score. Results show that traditional machine learning models such as Random Forest, SVM and Logistic Regression, outperform models such as XG Boost, Light GBM, and Cat Boost. Feature engineering played a pivotal role in the performance of the classification models. There are many frequency based statistical methods introduced to detect encryption algorithms on ciphertext. But differentiation of encryption scheme among various applications, especially for high randomness block ciphers, is out of them. To circumvent this, the new system adopted Hamming Weight Distribution-based feature extraction that captured bitwise changes in the ciphertext to produce more informative features for classification. Recursive Feature Elimination (RFE) using XGBoost was also employed to further optimize the feature set by removing any redundant and less informative features, thus enhancing model efficiency and performance. Various machine learning models were trained and tested to evaluate the model of the proposed system. The performance of the classification model was different, with XGBoost producing the best score of 71.0%, and LightGBM and CatBoost yielding results that were very close (70.87 and 70.73%, respectively). Gradient boosting models were primarily trained on that data, as they could capture the complex feature interactions better than the others. However, classic models like Random Forest and Logistic Regression showed only average results and lacked the ability to learn more complex feature representations that are instrumental in differentiating between various forms of encryption mode.

TABLE 2. Classification Report for Encryption Algorithm Identification

Algorithm	Precision	Recall	F1-Score	Support
AES (0)	0.98	0.76	0.85	1026
CAST (1)	0.99	0.74	0.85	1040
DES (2)	0.54	0.66	0.6	1073
RC2 (3)	0.55	0.68	0.61	1060
Overall	0.71	-	-	4199
Accuracy				
Macro Avg	0.77	0.71	0.73	4199
Weighted Avg	0.76	0.71	0.72	4199

The encryption mode applied plays a vital role in how well a machine learning model works for cryptanalysis. The block ciphers can us tell you how different modes of operation effects on the structure and randomness of the cipher text you receive. For the four encryption modes of ECB, CBC, CFB and CTR, the accuracy of classification was significantly different. The simplest mode to classify was ECB mode, which deterministically encrypts identical plaintext blocks into matching ciphertext blocks. This property allowed a greater distinguishability of ECB-encrypted data, which resulted in a better classification accuracy. We speculate that other modes such as CBC, CFB and CTR add higher levels of randomness, helping to make classification more difficult. Along with the ML models, a web-based encryption classification tool was also developed for real time detection of encryption algorithms. With the trained models, a web application is designed, in which users need only insert ciphertexts and get a prediction for the classification. It was user-friendly, allowing easy interaction and giving specific information about the classification. The web application was evaluated in terms of accuracy, response time, and usability in expecting real-time performance. The classification results on the system were comparable to those of the best performing machine learning models, thus the proposed work can indeed be used in practical cryptanalysis problems. Moreover, the prediction application response time was optimized with fast model deployment to provide predictions with minimum computational inference time.

FIGURE 2: Encryption Algorithm Prediction Web Application Interface

5. CONCLUSION

This study investigates a possible application of machine learning techniques in the classification of encryption algorithms based on ciphertext. We run and tested different model MY XGBoost, MY LightGBM, MY CATBoost with extensive statistical feature extractio methods. As represented in our experimental results, XGBoost has shown the best classification accuracy 71.0% compared to the other models, while then LightGBM (70.87%) and CatBoost (70.73%) are performed closely. Also, high precision and recall scores in AES and CAST algorithms are noticed while DEoS and RC2 are relatively more difficult to differentiate as they are more structurally relevant shares in terms of encryption keys. Our findings are promising, although there are some limitations. Adding more cryptographic features, polishing feature engineering approaches, and using deep learning algorithms for higher-level pattern extraction can help improve the classification performance. Future work will be concentrated on refinement of the model's accuracy, broadening the dataset with different encryption schemes and enhancing the robustness of predictions across different encryption modes.

REFERENCES

[1]. Dileep, A.D., & Sekhar, Chandra, "Identification of Block Ciphers using Support Vector Machines," Proceedings of International Joint Conference on Neural Networks, Vancouver, BC, Canada, July 2006.

[2]. Nagireddy Sreenivasulu, Murthy Hema, & Shri Kant, "Identification of Encryption Method for Block Ciphers using Histogram Method," Journal of Discrete Mathematical Sciences and Cryptography, vol. 13, pp. 319- 328, 2010.

[3]. Manjula, R., & Anitha, R., "Identification of Encryption Algorithm Using Decision Tree," Advanced Computing, pp. 237–246, 2011.

[4]. Cannière, Christophe, Biryukov, Alex, & Preneel, Bart, "An Introduction to Block Cipher Cryptanalysis," Proceedings of the IEEE, vol. 94, pp. 346-356, 2006.

[5]. lbassal, A. M. B., & Wahdan, A. M. A., "Neural Network-Based Cryptanalysis of a Feistel Type Block Cipher," International Conference on Electrical, Electronic and Computer Engineering (ICEEC'04), Cairo, Egypt, pp. 231-237, 2004.

[6]. Albassal, A. M. B., & Wahdan, A. M. A., "Genetic Algorithm Cryptanalysis of a Feistel Type Block Cipher, Proceedings of IEEE International Conference on Electrical, Electronic and Computer Engineering (ICEEC'04), pp. 217–221, September 2004.

[7]. Saxena, G., Classification of Ciphers using Machine Learning, Master's Thesis, Indian Institute of Technology, Kanpur, 2008.

[8]. Duda, R.O., Hart, P.E., & Stork, D.G., Pattern Classification, 2nd edn, 2004.

[9]. Hernandez, J. C., & Isasi, P., "New Results on the Genetic Cryptanalysis of TEA and Reduced-Round Versions of XTEA," Congress on Evolutionary Computation (CEC2004), vol. 2, pp. 2124–2129, June 2004.

[10]. Keliher, L., Meijer, H., & Tavares, S., "Provable Security of Substitution Permutation Encryption Networks Against Linear Cryptanalysis," Canadian Conference on Electrical and Computer Engineering, vol. 1, pp. 37–42, 2000.

[11]. Biham, E., "Cryptanalysis of Multiple Modes of Operation," Lecture Notes in Computer Science, Advances in Cryptology, Proceedings of ASIACRYPT94, 1994.

[12]. Yang, Y., & Pedersen, J., "A Comparative Study on Feature Selection in Text Categorization," ICML 1997, pp. 412–420, 1997.
[13]. Lin, F-T., & Kao, C-Y., "A Genetic Algorithm for Cipher Text-Only Attack in Cryptanalysis," Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, vol. 1, pp. 650–654, 1995
[14]. Ithape, S., & B. P. R., "Identification of Encryption Method for Block Ciphers using Machine Learning Methods," 2023 IEEE 20th India Council International Conference (INDICON), Hyderabad, India, pp. 1228- 1233, 2023. DOI: 10.1109/INDICON59947.2023.10440785
[15]. Zhao, L., Chi, Y., Xu, Z., & Yue, Z., "Block Cipher Identification Scheme Based on Hamming Weight Distribution," IEEE Access, vol. 11, pp. 21364-21373, 2023. DOI: 10.1109/ACCESS.2023.3249753.
[16]. Kaggle Dataset: Encrypted Data. Available at: https://www.kaggle.com/datasets/xuqinyang/encrypted-data.

