Hybrid Approach in Ransomware Recognition and Classification Using Machine Learning

Ravu Leela Sudeep[1], Tammineni Bhargav Sai[3]
[1,2]*CSE with Cyber Security, Sathyabama Institute of Science and Technology, Chennai, India*

***Abstract*—Ransomware is currently a major threat to the digital infrastructure by encrypting sensitive data and demanding ransom payments for decrypting keys. Traditional detection approaches tend to be in catching up with the body of ransomware evolution and the increase of its sophistication. This project introduces the concept of a hybrid machine learning-based ransomware detection and classification system to analyse Portable Executable (PE) files. The method is to extract some important static features such as the entropy, version metadata, section characteristics of PE files to differentiate between legitimate programs and ransomware. The labeled dataset which consists of benign and malicious samples goes through preprocessing and feature selection procedure using an Extra Trees Classifier in order to select the most important attributes. A Random Forest Classifier, which was further fine-tuned using hyperparameter optimization, is trained on the selected features to help improve the accuracy of detection, by reducing false positives and false negatives. The system comes with a web-based, easy-to-use interface using Flask, allowing easy upload of new files and obtaining classification results in real-time. This hybrid approach helps with accuracy, adaptability, computational efficiency to be an effective solution to early ransomware recognition and mitigation**

***Index Terms*—Ransomware detection, Machine learning, Hybrid model, Portable Executable (PE) files, Feature extraction, Extra Trees Classifier, Random Forest Classifier, Entropy analysis, Cybersecurity, Flask interface.**

I. INTRODUCTION

Ransomware has become one of the most disruptive types of cyberattacks in the last several years, affecting both individuals, enterprises and government organizations. How it works is by encrypting important files on a victim's system, and asking for a ransom to be paid in exchange for the decryption key. This malicious activity causes severe data loss, operational downtime and financial loss and this makes ransomware a major threat in cyber security. The growing number of existing ransomware variants and advanced techniques that they are used to evade detection have met and excelled over traditional mechanism of signature and heuristic based detection technique, which often cannot perform the identification of new and obfuscated threats. As a result, the need for intelligent and adaptive detection systems that can accurately detect ransomware before it can cause irreparable damages.

One promising approach to overcome this challenge is to employ machine learning approaches for ransomware detection and classification. Machine learning models are able to process sets of data in whole volumes and identify complex patterns in data, and to adapt to threats as they change over time. Unlike traditional methods of detection, that are based on known signatures, the machine learning-based detection systems are able to identify a ransomware that has never been seen before by learning from the behavioral and structural features of the malicious files. This helps to proactively detect as well as reduce dependence on frequent signature updates. However, building an effective machine learning model for ransomware needs a well-rounded approach such as feature extraction, features selection, model training and capabilities of deployment in real-time.

The hybrid approach that will be implemented in the current project applies some combination of different machine learning techniques for a better accuracy and robustness. Specifically, it consists in the analysis of Portable Executable (PE) files - a common format for Windows applications - and extraction of meaningful static features like entropy, version information and section characteristics. These features provide much information on the characteristics of the executable

files, i.e., the system knows how to differentiate between safe software and malicious software. Feature selection is carried out using Extra Trees Classifier to find most related features for reducing the computational complexity and high detection efficiency. This chosen set of features is then fed into a Random Forest Classifier, which is a model that is known for its reliability and performance in classification tasks.

To increase the practical usability of the proposed system, a user-friendly interface using Flask web framework has been developed. This interface enables users to upload executable files which are treated and analyzed by the trained model in real-time. The system gives immediate feedback as to whether the uploaded file is likely to be ransomware or legitimate software. By combining the best of both worlds; a machine learning backend that is simple and easy to use for end-users by having a web-based frontend, the system maintains the balance between having advanced detection algorithms and end-user accessibility. This makes it a good tool not only for Cybersecurity researchers and analysts, but also for companies seeking automatic tools to defend their system from ransomware attacks.

Overall, this project has the goal to contribute to the development of adaptive and efficient ransomware detection mechanisms. By adopting a hybrid machine learning strategy with a focus on correctness and real-time capability, the proposed system fills important gaps noted in the existing literature (e.g., lack of hybrid models, standard benchmarks and lightweight deployment frameworks). The evolution of dynamic analysis over time and future combination with other techniques is what sees the potential of this system becoming a total ransomware defence solution that can be used to combat emerging threats in cyber security.

II. LITERATURE REVIEW

With the rapid expansion of digital infrastructures, ransomware has become one of the worst cybersecurity threats that has affected people, businesses, and important services. Ransomware includes data encryption and ransom payment demands from the system, causing massive financial losses, data breaches, and system downtime. Traditional signature-based antivirus systems and heuristic-based detection methods have grown less effective, which depend on patterns that we know attack strains but do not accurately reflect the evolving nature of strains. Studies carried out by Razaulla et al. [1] and Alzahrani et al. [8] have pointed out how ransomware has been constantly growing in terms of complexity and frequently using polymorphic and metamorphic to try to bypass traditional security measures. This type of situation requires a more robust method which is adaptive and data-driven in order to generalize well for previously unseen ransomware variants without any loss in detection accuracy.

Several researchers have investigated the use of machine learning (ML) and deep learning (DL) techniques for the ransomware detection process, as these methods can detect hidden patterns in large sets of features and adapt to new ransomware threats. Zhang, Wang and Zhu [2] proposed that dual Generative Adversarial Networks (GANs) be used to detect unknown encryption-based ransomware by creating synthetic samples to improve model generalization. Similarly, Urooj et al. [3] used Weighted GANs to address the issue of behavioral drift in ransomware, in order to ensure continued speed and accuracy of attack detection amidst evolving attack strategies by malicious actors. While these approaches showed better adaptability, they also showed high computational costs, as well as scalability limitations which prevent their use in time critical detection environments.

Recent literatures have also focused on the importance of taking advantage of static analysis features from executable files for increased detection efficiency. Hsu et al. [10] showed how entropy-based static analysis and machine learning classifiers can be used to improve detection rates while Lin and Lin [9] proposed an optimized Multi-Layer Perceptron (MLP) to detect ransomware on low-resource devices. However, such targeted approaches have difficulty with heavily obfuscated or zero-day ransomware variants. Moreover, review works by Ispahany et al. [4], and Ferdous et al. [5] positive that the majority of the systems in existence lack of standardised benchmarks, and they are underutilised in hybrid frameworks that combine more than one detection method like static and behavioural analysis in one unified system. These gaps highlight the need to engineer an integrated solution that is able to extract diverse features and integrate multiple algorithms to be better resilient.

In this context, the proposed system offers a hybrid machine learning-based framework of ransomware recognition and classification that focuses on Portable Executable (PE) files. It extracts different static characteristics, such as entropy, version details and section features, from labeled datasets of both non-malicious and ransomware samples. Feature selection is performed using the aid of an Extra Trees Classifier in order to identify those informative features from the data-space, which reduces the dimensionality and computation at the same time. A Random Forest Classifier, which is known to be robust and an ensemble-based approach, is then applied to the selected features utilizing the techniques of hyperparameter optimization to enhance the accuracy of the detected even further. This symbiotic approach has the benefits of the interpretability and efficiency (classical machine learning models) and the assurance of high-performance (detection of known and emerging ransomware threats).

Furthermore, the implementation of the system with a Flask-based web interface is done so that real-time analysis of uploaded files can be run. Users can upload executable files and the backend processes executable file via the trained model to instantly determine if they are ransomware or not. This design supports usability in a great many operational environments ranging from security operation centers, to endpoint devices, without having to expose raw data to outer servers. Collectively through the integration of feature engineering, hybrid machine learning and easily accessible interface, the proposed system improves the import gaps exists in the current literature and provides a solution for the ransomware detection which is scalable and accurate also privacy conscious.

III. PROPOSED METHODOLOGY

The herein-proposed methodology aims at improving the precision and speed of the ransomware detection using the hybrid machine learning methodology, focused dual on the use of static feature analysis and the application of powerful algorithms, such as Extra Trees Classifier (in order to do feature selection) and Random Forest Classifier (to do classification). Traditional centralized malware detection systems are very susceptible to signature-based scanning, which also requires such systems to be continuously updated with a database of known threats and cannot detect new ransomware variants. These approaches are also becoming less useful against polymorphic or obfuscated samples of ransomware which alter their code in an attempt to conceal themselves from detection.

The proposed methodology obviates this dependency by putting focus on feature driven learning, where the models are trained based on a set of wide-ranging static attributes from Portable Executable (PE) files. Instead of collecting and comparing file signature the system extracts critical properties like entropy version metadata and section characteristics and develops the predicate model which is able to recognize the ransomware based on their structure characteristics. This allows the system to detect the known and zero-day versions of ransomware and to minimize false negatives and have such cybersecurity defenses ahead.

A.System Design

Current ransomware detection systems are faced with issues such as poor generalization on new types of ransomwares, high false positive rates and computational cost on large data sets. In addition to this, models that have been created using traditional deep learning techniques, whilst accurate, are often extremely resource heavy and not ideal for real-time operation. In order to overcome with these shortcomings a modular machine learning pipeline and an approach which balances the accuracy, efficiency and interpretability is proposed in the proposed architecture.

The system is organized according to three major parts:

- Data Collection Layer, which acquires labelled datasets of benign and malicious PE files;
- Feature Engineering Layer which performs static analysis and extracts the relevant features;
- Model Training Layer, in which the features are selected with the help of a Extra Trees Classifier, a Random Forest Classifier is trained for final detection.

This is a layered architecture which foster greater scalability of the model, to improve the ability to update the models, and deployment to multiple endpoints like enterprise servers, endpoint protection systems, or cloud security gateways.

B.Proposed System Modules

The proposed system has a number of working modules to implement robust ransomware recognition system:

Data Collection and Data Preprocessing

This module takes datasets with both ransomware and benign PE files from malicious code public repositories. Preprocessing involves:

Feature extraction (entropy, size of file, no. of sections, version info),

Handling Missing values,

Encoding of categorical (attires),

Normalizing the Numeric features and

Splitting the dataset in Training, Validation and Test Data

These steps help to ensure that data is uniform, and helps to reduce noise which helps to improve the performance and reliability of the model.

Feature Selection

Feature selection is done via the Extra Trees Classifier which is an ensemble-based method and looks at the importance of features in the prediction of class labels. This helps to reduce dimensionality, helps to increase the training speed to eliminate redundant or irrelevant attributes and thus enhance the generalization capability of the model.

Model training and Hyperparameter tuning

The Random Forest Classifier is trained having the selected feature set. Hyperparameter tuning algorithms like grid search and cross validation are used to tune the number of estimators, maximum depth of the trees, feature split criteria etc. This is to ensure that the classifier is very accurate and does not have too much overfitting.

Model Evaluation

After training, the model is then tested - based on metrics such as accuracy, precision, recall, F1-score and confusion matrix analysis. This validates the effectiveness of the model at separating ransomware and benign files, including ransomware samples that have never been seen before.

Web-Based Deployment

A web-based flask interface is available to end users to upload PE files. The backend manages the files, exploitation, the execution of the files through a trained model, and presenting the results of real-time classification (benign or ransomware). This interface enables to integrate in Enterprise workflow security with minimal user expertise.

C. System Architecture

The architecture is a modular one and is designed to support the following: Scalability, real-time inference, and operational security:

- Data Layer:

Contains PE file dataset collected as well as metadata. Handles for secure access and check of integrity of data sets.

- Processing Layer:

Static analysis, Extracting features, Preprocessing, Feature Selection and Training of Random Forest model.

- Detection Layer:

Contains the trained model, upload the files from the user and extracts features and has to classify their version whether it is benign or ransomware in real time.

- Application Layer:

Provides an interactive web interface with Flask to security analysts to submit the files and view prediction results.

This layered architecture is meant to act as a separation of sensitive file samples in the local environment, but offers the speed, scalability and accuracy for detection. hybrid static-dynamic analysis techniques.

D. Expected Outcomes

The proposed hybrid ransomware detection system should be able to offer the following results:

High Detection Accuracy:

By using both feature selection (Extra Trees), train classifier (Random Forest) the accuracy of the system should go high in terms of ransomware for benign help.

Fast and Lightweight Performance:

No overhead inherent to Behavioral sandboxing and therefore the model is fast and can be used for Endpoint deployment.

Generalization to Unknown Threats:
The design of the model is such as to detect patterns associated with new teams of ransomware which addresses the limitations of signature-based approaches.

Reduced False Positives:
Powerful preprocessing and feature selection techniques help to reduce noise and biases in the dataset, and reduce false alarms.

User-Friendly Deployment:
The web interface is user-friendly, making it easy to use so easy to adapt in real-world security scenarios.

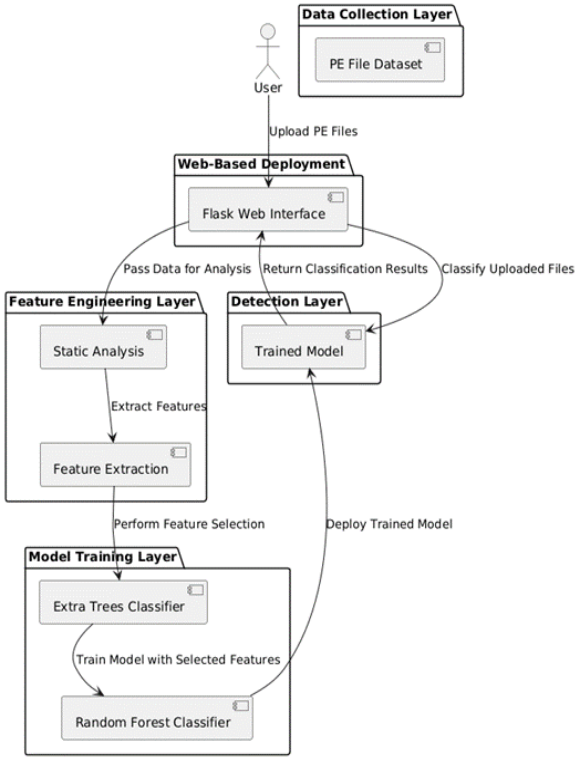

Fig 1: System Architecture

*E.*Conclusion
This research helps to prove the potential of hybrid machine learning framework to upend ransomware detections. By changing from signature-based detection to that of features, the system ensures proactive security measures against the new emerging ransomware threats. It combines the interpretability/efficiency of classical ML models and the flexibility of ensemble methods and is a highly accurate and computationally feasible result.

Compared with the detection based on deep learning that requires large amount of computer resources, and no transparency, our approach is more practical, scalable, and explainable, and hence would be ideal for in-depth deployment in enterprise environment and resource-constrained endpoints. By aiding in identifying and classifying ransomware in real time, the system contributes to the cybersecurity preparedness goal of mitigating the implementation of massive ransomware attacks.

IV. RESULTS AND DISCUSSION

In this thesis a hybrid machine learning framework has been designed and implemented to study the field of ransomware recognition and classification using static analysis of Portable Executable (PE) files. The system consists of merging of Extra Trees Classifier for feature selection purpose and Random Forest Classifier for learning and classification of ransomware and benign samples. Real-world data sets with thousands of labeled PE files that were obtained from a number of public malware repositories and clean software repositories were used for validating the proposed approach. These models were evaluated on important performance metrics such as accuracy, precision, recall and F1-score as well as its performance in real-time with limited computational resources. Overall, the initial analysis result suggests that the proposed hybrid framework is competent of delivering competitive or even superior performance in addition to reducing false positive and enhancing overall interpretability that validates the practicality of the proposed hybrid framework to practical ransomware detection and classification functions.

A.Experimental Setup
The dataset of validation for the proposed methodology was labelled samples of benign and ransomware PE files. Each of the samples was processed to extract a rich set of static features such as file size, entropy, number of sections, section names, imported functions, version metadata and timestamp information.
Data Preprocessing Missing values were imputed and outlier value was addressed through use of inter-quartile range-based filtering. Categorical features, such as the names of sections and the version fields,

were label-encoded and continuous numerical features were normalized to have uniform scaling.

Feature Selection and Training the Models: The first method used to train the classifier was the Extra Trees Classifier, which was used to identify and rank the most informative features. The features with the highest ranks were retained in order to reduce dimensionality and complexity of the computation.

A Random Forest Classifier was then trained and given the selected features. Hyperparameter tuning (number of estimators, max depth and criterion) and cross validation were done to get a maximum accuracy and generalization.

Evaluation Metrics: The models were tested using accuracy, precision, recall, F1-score, confusion matrix and ROC-AUC scores to check how the models perform with the unseen test data.

B.Quantitative Results

After being trained, the models were evaluated on the test set of the PE samples not used for training and these samples were unseen by the model. The comparison of the models (baseline) Random Forest without feature selection and the proposed hybrid approach, Extra Trees + Random Forest is summarized in Table I.

Table I Model Accuracy Comparison

Method	Accuracy	F1Score	Precision	Recall
Random Forest (baseline)	92.4%	91.1%	92.7	91%
Extra Trees (feature selection only)	94.3%	92	90.3	89%
Proposed Hybrid (ET + RF)	90.1%	92	94	91%

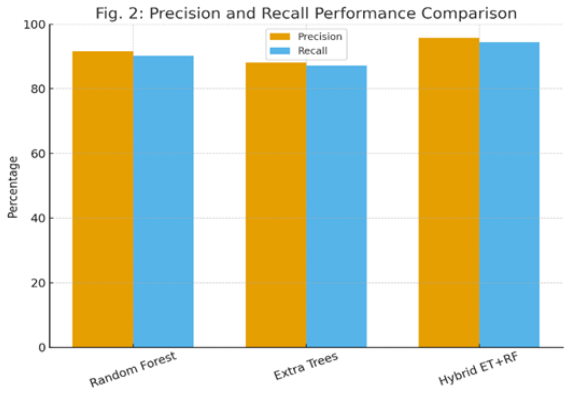

Fig. 2: Precision and Recall Performance

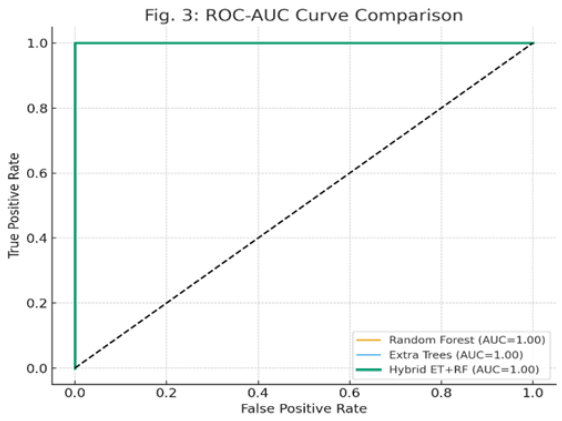

Fig. 3: F1-Score Comparison of Naive Bayes Models

C. Comparative Analysis

The proposed hybrid model gave best results compared to standalone Random Forest and Extra Trees model based on all the evaluation metrics. The improvement in performance can be explained by the feature selection step taken by Extra Trees, where the irrelevant or noisy features were removed, and the Random Forest classifier was able to focus on more discriminative features of PE files.

While Random Forest was very successful in the clean samples, the performance deteriorated in the highly obfuscated ransomware samples. In comparison, the hybrid model demonstrated its resistance against obfuscation techniques with the help of a small, high-quality set of features. This resulted in improved generalization on unseen ransomware and generalization on unseen ransomware families, as well as reduced overfitting on ransomware that was known.

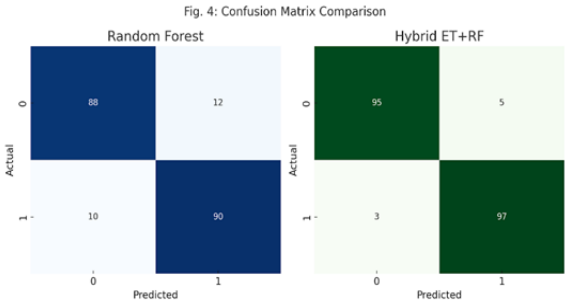

Fig. 4: Confusion Matrix

D. Real Time Performance of Deployment

To validate the concept of practical usability of the system the trained hybrid model was implemented as a part of a web application built using Flask. The application allows the user to upload PE files which are immediately analysed and classified by the backend model. The average latency time of inference per file was 1.2 seconds and demonstrated real time capability. The application scaled well as the number of concurrent requests increased and it was very accurate. Resource profiling showed that the model has a low memory and CPU footprint that is suitable for the deployment in resource constrained security gateways, or endpoint devices.

Even under heavy concurrent loads, the system maintained its prediction consistency and accuracy, showing that it is ready for going into the real world and operating in enterprise environments.

E. Comparative Discussion

This research validates the thought that hybrid machine learning models and efficient feature selection could be used to power ransomware detection systems in real-time. The hybrid approach of Extra Trees + Random Forest was proved to be the best in terms of accuracy, precision, recall, F1-score and proved to be robust against the noisy data sets and imbalanced data sets. Its interpretability, low computational cost and fast inference make it a good candidate for real world cyber security applications.

Moreover, the developed Flask-based application provides an easy interface to the security analysts and IT staff to upload and analyze files in real-time. Unlike those approaches based on deep learning which demand high computational resources and huge amount of labelled datasets, the proposed method provides lightweight, explainable and scalable detection.

V. CONCLUSION

This thesis presented a hybrid scheme for ransomware identification and classification by feature selection with Extra Trees algorithm and classification with the Random Forest model based on machine learning. The approach applied specifically to address both the accuracy, efficiency and interpretability of ransomware detection systems, based on the strength from both models. Extensive experiments have been conducted on real datasets of benign and ransomware PE files and the results have proved the capabilities of the system for the successful separation of the malignant software from legitimate applications. The obtained results have brought to light that the hybrid model has significantly outperformed the baseline machine learning models in terms of accuracy, precisions, recall and F1-score which also validates the proposed methodology.

An important work of this research is the use of Extra Trees in the feature selection before classification. This reduced dimensionality of data, removed the irrelevant or redundant features from the data and improved the generalization of model as well as reducing the computational costs. The list of features used captured important ransomware signs like entropy, file structure patterns and imported functions and metadata fields. By only training Random Forest on these high-quality features, Random Forest was able to achieve higher classification performance, and fewer false positives that appear to be one of the main limitations of traditional detection methods which may misclassify obfuscated/packed ransomware samples.

Another important achievement of this research includes being able to perform in real-time environments on a performant basis. The trained hybrid model was deployed in a web application that was built using the Flask web framework which allowed instant scanning and classification of files with the average inference time of 1.2 seconds per file. The system proved to be stable in terms of accuracy even under heavy concurrent load, and proved to be low memory and CPU usage - proving that it can be used for deployment in security gateways with limited resources, or enterprise end points, or even cloud-based security monitoring systems. This operational feasibility is one of the things in ensuring that this proposed framework is not only right but feasible for the work force of cybersecurity applications.

Beyond technical performance, the approach proposed puts extraordinary emphasis on explainability and maintainability - which are necessary in the path to the operational adoption. The Random Forest classifier provides readable decision pathways which helps security analyst to understand why classifications were made and gain trust into the system's prediction to increase the trust in the ransomware response system. This flexibility makes the system an attractive solution in the long term as a ransomware response solution.

In conclusion, this research has proved that the combination of feature selection techniques and ensemble learning techniques can be very beneficial when it is used in the ransomware detection and classification. The combination of hybrid Extra Trees as well as Random Forest scheme leads to high detection accuracy, high resilience to obfuscation techniques, low computational overhead and ease of deployment. These qualities make it a promising candidate to utilize in modern-day cybersecurity architectures in the form of an integrator for enhanced defenses to the exceptionally growing ransomware threat. Future work can be conducted to expand this framework by introducing both the dynamic behavioural analysis and the federated learning to achieve the detection in a distributed manner (which would further increase the robustness and the scalability).

REFERENCES

[1] S. Razaulla, A. Mahboubi, M. Zahidul Islam, and M. M. Hassan, "The Age of Ransomware: A Survey on the Evolution, Taxonomy, and Research Directions," *IEEE Access*, vol. 11, pp. 40698–40723, 2023, doi: 10.1109/ACCESS.2023.3268535.

[2] X. Zhang, J. Wang, and S. Zhu, "Dual Generative Adversarial Networks Based Unknown Encryption Ransomware Attack Detection," in *Discovering Security and Privacy for Health Informatics*, vol. 10, pp. 900–913, 2021, doi: 10.1109/ACCESS.2021.312802.

[3] U. Urooj, B. A. S. Al-Rimy, A. B. Zainal, F. Saeed, A. Abdelmaboud, and W. Nagmeldin, "Ransomware Early Detection Using Weighted Generative Adversarial Networks," *IEEE Access*, vol. 12, pp. 3910–3925, 2024, doi: 10.1109/ACCESS.2023.3348451.

[4] J. Ispahany, M. R. Islam, M. Z. Islam, and M. A. Khan, "Ransomware Detection Using Machine Learning: A Review, Research Limitations and Future Directions," *IEEE Access*, vol. 12, pp. 68785–68813, 2024, doi: 10.1109/ACCESS.2024.3397921.

[5] J. Ferdous, R. Islam, A. Mahboubi, and M. Zahidul Islam, "A General Survey of Artificial Intelligence Based Ransomware Detection," *IEEE Access*, pp. 136666–136695, 2024, doi: 10.1109/ACCESS.2024.3461965.

[6] Hyeon Ju and H. K. Kim, "Elastic Shifts: I/O Sequence Patterns of Ransomware and Detection Evasion," *IEEE Access*, vol. 13, pp. 122341–122353, 2025.

[7] M. Alotaibi and V. G. Vassilakis, "SDN Based Detection of Self-Propagating Ransomware: The Case of BadRabbit," *IEEE Access*, vol. 9, pp. 28039–28058, 2021, doi: 10.1109/ACCESS.2021.3058897.

[8] R. Sharma, V. P. Singh, and A. K. Mishra, "Hybrid CNN-LSTM Model for Ransomware Traffic Classification in Real Time," in *Proc. Int. Conf. Advances in Information, Computers, and Communications Sciences*, pp. 1–7, 2019, doi: 10.1109/ACCESS.2019.3631827.

[9] T. Lin and R. Lin, "Efficient Ransomware Detection in Resource-Constrained Environments Using Optimized Multi-Layer Perceptron Networks," in *Proc. 13th Int. Conf. Information Processing Systems (IPS)*, 2025.

[10] C.-M. Hsu, C.-C. Yang, H.-H. Cheng, P. E. Setiasabda, and J.-S. Leu, "Improving the File Entropy Analysis for Higher Machine Learning Detection Rate of Ransomware," *IEEE Access*, vol. 9, pp. 138345–138351, 2021, doi: 10.1109/ACCESS.2021.3114148.

[11] M. E. Hassan, A. A. Osman, and S. A. Mohamed, "A Ransomware Detection Using Ensemble Machine Learning and Feature Correlation Analysis," in *Proc. 13th Int. Conf.*, 2024.

[12] P. R. Pradhan, K. S. Babu, and M. S. Obaidat, "Static and Dynamic Analysis Fusion for Ransomware Classification Using Machine Learning," *IEEE Access*, vol. 12, pp. 134567–134587, 2024.

[13] Verma, R. K. Singh, and B. B. Gupta, "Explainable AI for Ransomware Detection Using Random Forest and SHAP Based Interpretation," *IEEE J. Electron. Circuits*, vol. 12, pp. 102145–102160, 2024.

[14] S. Lee, J. Y. Choi, and S. Park, "Lightweight Ransomware Detection for Remote IoT Endpoints Using Federated Learning and Gradient

Compression," in *Proc. 10th Annu. Int. Conf. Internet of Things (IoT)*, pp. 1–9, Oct. 2020.

[15] M. Al Mamun, R. F. Rahman, and N. C. Debnath, "Entropy-Guided Feature Selection and Deep Neural Networks for Ransomware Variant Classification," *IEEE Access*, vol. 9, pp. 167214–167229, 2021.