



# Machine Learning-Based Intrusion Detection for Malicious Email Attachments

Kanak Lata<sup>1\*</sup>, Ritesh Jha<sup>1</sup>, Vijay Kumar Jha<sup>1</sup>

<sup>1</sup> BIT Mesra, Ranchi, Jharkhand, India

\*Corresponding Author: [Phdcs10005.24@bitmesra.ac.in](mailto:Phdcs10005.24@bitmesra.ac.in)

**Abstract:** Email attachments are a common way for cybercriminals to distribute malware. These attachments can include ransomware, spyware, viruses or other malicious software. Traditional antivirus systems are primarily signature-based, and do not recognize new or unknown malware. It has become increasingly difficult to protect against the variety of attacks that are currently being used, especially if they are new variants of malware. This research aims to see if it's possible to automatically detect malicious email attachments using machine learning (ML) algorithms before they are executed. The machine learning (ML) models were trained and tested with a dataset of 100,000 file samples, with half of it is malicious and the other half of it is benign. A well-balanced class distribution minimized bias in model training and gave a fair ground for performance calculation. To ensure data quality, certain non-informative attributes like hash identifiers and timestamp-related attributes were filtered before developing the model. Three supervised learning algorithms – Random Forest (RF), Logistic Regression (LR) and Linear Support Vector (SVC) Classification – were trained and evaluated. Experimental analysis showed that the RF model has the best performance with an average classification precision of 91.28%, while the LR model and the Linear SVC model achieved an average classification accuracy of 74.82% and 76.28%, respectively. The results show that ML can be a valuable tool to recognize potentially dangerous attachments prior to execution. Future studies could involve using real-world email datasets and explainable artificial intelligence (XAI) methods to improve the practical application and interpretability of models.

**Keywords:** Malicious Email Attachments, Malware Detection, Machine Learning, Intrusion Detection, Email Security

## I. Introduction

Email is still one of the most important and basic ways to communicate in schools, offices and between people around the world! Although email is easily accessible and convenient, it is still a key way in which cybercriminals spread their messages because of its ubiquity and the ease of perpetration [1]. Malicious attachments PDF files, Microsoft Office documents, ZIP, executable or macro enabled documents are some of the typical ways that attackers transmit malware, ransomware, trojans, spyware and other payloads.

In the last few years, attacks such as phishing and business electronic mail compromise (BEMC) have increased tremendously, resulting in financial losses and data breaches in multiple industries [2]. Traditional e-mail security methods, such as blacklist filtering and rule-based detection systems and signature-based anti-virus software, are well suited to deal with known threats, but have been shown to be ineffective when encountering zero-day attacks and variants of malware [3]. With the ever-evolving nature of cyber threats, there is a growing need for intelligent and adaptive security measures that can detect malicious attachments before they are executed on a system. Some studies have been conducted on phishing detection, malware analysis and email security with machine learning and deep learning approaches. However, there have been relatively few studies that have focused on the detection of threats at the attachment level using features of the malicious behaviour. Preventing users from interacting with a suspicious attachment can greatly help minimize the chances of malware infection and unauthorized access into a system.

This research aims to identify how effective it is to detect malicious email attachments in a proactive way using machine learning algorithms. To train and test various classification models, a behavioural malware dataset that consists of malicious and benign samples is used. The performances of three classifiers i. Linear SVC, ii. RF iii. LR are compared and the effectiveness of each of them in detecting malicious attachments is investigated.

**The main objectives of this research are:**

- To examine malware behavioural datasets for attachment-related threats.
- To identify malicious and benign attachment samples—Machine learning techniques.
- To compare the performance of multiple machine learning algorithms.
- Smart threat detection in e-mail to enhance proactive security.

**The key contributions of this work are:**

- Development of a machine learning framework for detecting malicious attachments.
- Comparative analysis of the classifiers Random Forest, Logistic Regression and Linear SVC.
- Behavioral analysis of malware features for classification.
- Early detection of attachments that could be malicious to boost email security.

## II. LITERATURE REVIEW

There have been several studies on email security utilizing ML and Natural Language processing (NLP). In [1], summarized the recent NLP techniques used for the detection of phishing e-mail and explored the methods including text preprocessing, feature extraction, and using ML for phishing classification. Their results showed the need for intelligent detection systems to enhance email security in the face of emerging cyber threats.

In [2], the authors conducted a systematic literature review of 77 studies, and they found 11 major challenges that can affect secure software development: exposure to cyber-attack; lack of knowledge; constraints on resources; and risks of unauthorized access.

In [3], presented a deep learning model for phishing electronic mail detection based on CNN, RNN, LSTM and BERT. Through experiments, it was found that the BERT-LSTM hybrid model had the best performance with a precision of 99.61%, proving that deep learning methods could be used to enhance email security systems.

In [4], Opaline Attachment was proposed, an attachment screening system based on DistilBERT for identifying malicious email attachments. The proposed framework uses intelligent threats detection and file sanitization via a sandboxed approach to prevent known and zero-day attachment-based attacks.

To extract the behavioral features from volatile memory, the authors in [5] developed a memory-based malware detection approach with memory forensic techniques. Experimental results showed that the SVM was able to reach an accuracy of 98.5% when using the classifier, and memory analysis proved to be a very effective method in detecting advanced malware threats.

RF algorithm, which is proposed in [6], is a combination of several decision trees to improve the classification accuracy, reduce overfitting and provide robustness to noisy data.

SVM is one of the supervised learning algorithms introduced in [7] which tries to identify an optimum separating hyperplane in a high dimensional feature space, for which complex data can be classified. Because of its high classification and generalization abilities, SVM has been used in many applications such as malware detection, and cyber security.

LR is described as an effective statistical method to classify binary classification problems in [8]. They laid the theoretical groundwork as well as practices of Logistic Regression for predictive modelling and decision making.

Although much effort has been invested in the detection of phishing attacks, classification of the malwares and email security, most existing studies have concentrated on either the content analysis of email messages or the detection of malicious URLs, or on general email or malware behavior. There are relatively few studies that investigate the detection of threats at the attachment level of the behavioral attributes of malware. In addition, there has been little research above performed results of traditional ML classifiers for proactive malicious attachment identification. To address this, the current study presents and tests intrusion detection system (IDS) based on the ML to identify malicious email attachments before they are executed.

**Table 1. Summary of Related Work on ML-Based Email Security and Malware Detection**

| <b>Seq.</b> | <b>Author(s)</b>      | <b>Method</b>         | <b>Application Area</b>  | <b>Key Findings</b>                                     |
|-------------|-----------------------|-----------------------|--------------------------|---------------------------------------------------------|
| [1]         | Salloum et al.        | NLP,ML                | Phishing Email Detection | Reviewed NLP-based phishing detection techniques        |
| [2]         | Khattak et al.        | SLR, CSC - SD Model   | Cybersecurity Challenges | Identified 11 major cybersecurity challenges            |
| [3]         | Atawneh and AL jehani | CNN,RNN,LSTM,BERT     | Phishing Detection       | BERT-LSTM achieved 99.61% accuracy.                     |
| [4]         | Magishashini et al.   | DistilBERT            | Malicious Attachments    | Combined threat detection with attachment sanitization. |
| [5]         | Sihwail et al.        | Memory Forensic + SVM | Malware Detection        | Achieved 98.5% accuracy                                 |

### III. Problem Statement

Electronic mail attachments are a common source of malware attacks and remain a serious cybersecurity challenge. Attackers frequently use shared file formats, such as PDF files, Microsoft Office documents, and compressed files, to distribute malicious software over emails. Most existing email security systems depend on signature-based detection methods, which work well for known malware but are less effective in detecting new or previously unseen attacks.

As a result, malicious attachments can bypass traditional security mechanisms and compromise system resources. Therefore, there is a need for an intelligent intrusion detection approach that can analyze system-level behavioral features and accurately identify malicious email attachments. In this work, supervised ML processes are employed to improve the detection performance by reducing false positives (FPs) and enhancing the reliability of email security.

### IV. Proposed Methodology

#### A. Description of Dataset

The malware dataset used in this paper is publicly available on Kaggle and contains 100,000 malware samples, which are used to train and test the proposed intrusion detection framework. The dataset consists of 50,000 malicious files and 50,000 benign files, which is a good ratio for binary classification problems. The sum of all samples is:

$$\text{Total Samples} = 50,000 + 50,000 = 100,000$$

The data set includes 35 features corresponding to the various behavioral, memory and execution aspects of the files. The attributes include hash identifiers, timestamp information, process-related attributes, memory usage metrics, execution behavior indicators and class labels. These characteristics can be useful when determining whether a specific file is malicious or benign by analyzing the file's behavior.

Both classes have the same number of samples, so that this dataset is suitable to evaluate the model and does not create class imbalance problems. The motives behind the behavior of these types of malicious email attachments are analyzed and the framework suggested is designed to prevent the detections before they can execute and thus enhance email security.

#### B. Data Preprocessing

Data preprocessing was performed to ensure good data quality and computational efficiency. The dataset consists of 35 features in the beginning. The fact that the millisecond timestamp and the hash identifier are not indicative of malware type was reason for their removal from the two attributes used as non-informative. Hash attribute is only used to identify files, whereas the timestamp attribute carries temporal information, irrelevant to identifying malicious and benign files.

The classification attribute was chosen as the target attribute in supervised learning. Eliminating two irrelevant attributes and omitting target label from the input feature set, 32 input features were selected for training and testing of the model.

$$\text{Number of input features} = 35 - 2 - 1 = 32$$

The resulting characteristics include behavioral, memory and execution-based characteristics which are used by the machine learning models to categorize files as malicious or benign.

### C. Train-Test Split

The dataset was split into training and testing sets to evaluate the performance of the machine learning models on new data. The data was divided using an 80:20 split ratio where 80% of the data was used for training the models and 20% was used for testing and evaluation.

Training Data = 80%

Data was split by the `train_test_split` function in Scikit-learn library. The training set was utilized to learn patterns of malicious and benign files while the testing set was utilized to check the generalization abilities of the trained models. This methodology provides an impartial measure of the performance of classification on a data set never seen before.

### D. Machine Learning Models

To classify malicious and benign email attachments, this study implements and compares three supervised machine learning algorithms.

#### 1. Random Forest:

RF is an ensemble learning method that aggregates several decision trees and enhances the classification performance and lessens overfitting. It can not only capture the goodness of individual trees but also has a strong ability to capture the complex pattern of the data and has strong predictive ability. Random Forest was chosen as the main classification model in this study, because of its high generalization capability and its tolerance to noise [6].

#### 2. Logistic Regression:

A popular supervised learning algorithms for binary classification problems is Logistic Regression. It uses a logistic function to approximate the chance of an instance fit in to a specific class and offers a simple but effective baseline solution to classification problem. Logistic Regression was used as a referent model in this research for comparison [8].

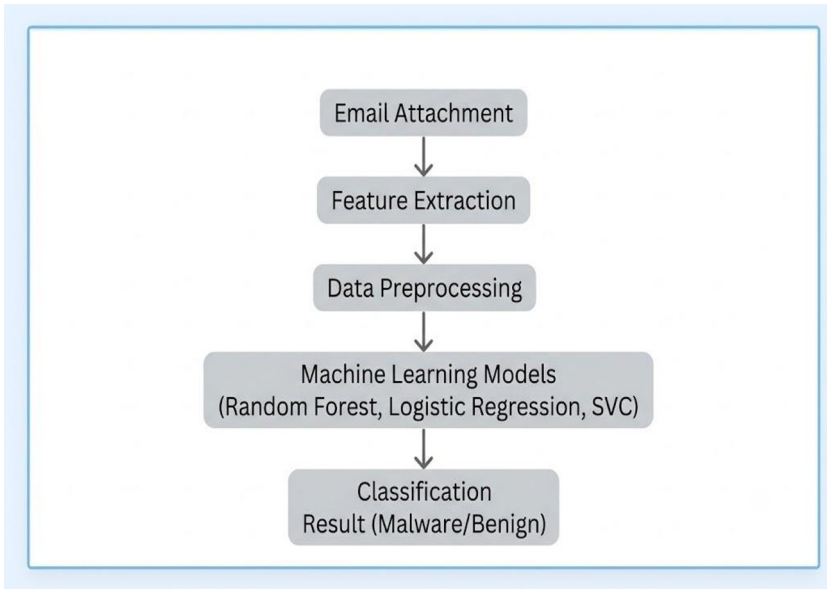
#### 3. Linear Support Vector Classification (Linear SVC):

Linear SVC is a supervised learning algorithm that is used for classification purposes and is based on the concept of a hyperplane, which is a plane that is part of the linear model. It is especially appropriate for high dimensional data sets and offers efficient and reliable classification results. Linear SVC was also selected to be added to the study to see if it could be effective in detecting malicious email attachments [7].

The accuracy, confusion matrix, Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) scores were used to assess the performance of the models implemented. The metrics were used to compare the classification accuracy of the models and to find the model which is most appropriate to detect malicious email attachments.

## V. System Architecture

The architecture proposed shows how machine learning is used to detect malicious email attachments. The process starts with receiving an email attachment, and then the features of the behavior, memory, and execution in the email are extracted. These attributes then are passed through preprocessing, and some irrelevant information (such as the hash identifiers and the timestamp information) is discarded. Preprocessed data is fed to the ML classifiers such as Linear SVC, LR, and RF. The trained model regulates if an attachment is malicious or not during the classification process by looking at the extracted features. When a file is malicious, the system gives an alert message to block the execution of the file, otherwise, it marks it as safe for further use. The overall workflow of the proposed system is summarized in Fig. 1



**Fig. 1. Architecture of the Proposed ML -Based Malicious Email Attachment Detection Framework**

## VI. Experimental Results

The accuracy, confusion matrix, Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC) scores were used to assess proposed machine learning models. The experiments were conducted using python and machine learning libraries such as Scikit-learn, Pandas, Matplotlib.

### A. Accuracy Comparison

The algorithms used were Linear SVC, Logistic Regression, and Random Forest, all of which are machine learning algorithms, and were implemented and evaluated to identify malicious email attachments. The results of the models' performance are shown in Table II.

**Table II: Accuracy Comparison of Machine Learning Models**

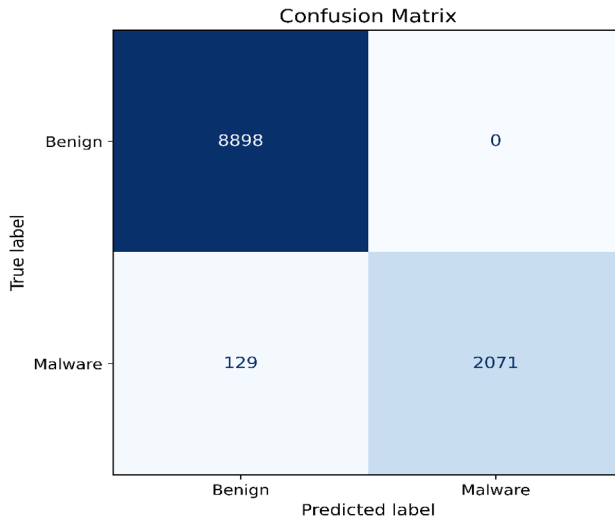
| Model               | Accuracy |
|---------------------|----------|
| Linear SVC          | 76.28%   |
| Logistic Regression | 74.82%   |
| Random Forest       | 91.28%   |

The experimental judgement show that the **RF** classifier reached the highest classification accuracy of **91.28%**, outperforming **Linear SVC (76.28%)** and **LR (74.82%)**. The higher performance of the RF model can be credited to its ensemble learning mechanism, which combines multiple decision trees to improve prediction accuracy and reduce the influence of noisy observations. In contrast, Logistic Regression and Linear SVC are linear classifiers and therefore have limited capability to model the complex non-linear relationships present in system-level behavioral features associated with malicious email attachments. These findings indicate that ensemble-based learning provides a more effective solution for malicious email attachment detection than conventional linear classification techniques.

## B. Confusion Matrix

RF was used as the best model and its performance was tested using a confusion matrix. Gives detailed predictions in terms of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). TP is a file that is flagged correctly as malicious while TN is a file that is flagged correctly as not malicious. An FP is benign files that are identified as malicious, while an FN is malicious files identified as benign. When dealing with malware detection, getting rid of FN is an important step to take because malicious files that go undetected can pose serious security problems. The confusion matrix results show that the RF classifiers have a good classification accuracy, and the confusion rate is low, it is effective to distinguish malicious and benign attachments. This confirms the usability and reliability of the suggested model in the field of malicious email attachment detection.

Fig. 2 presents the confusion matrix of the RF classifier, illustrating the distribution of correctly and incorrectly classified malicious and benign samples.



**Fig. 2. Confusion Matrix for the Classification Performance of the Random Forest Model**

The confusion matrix provides a detailed evaluation of the classification performance by illustrating correctly and incorrectly classified samples. The proposed RF classifier achieved a high number of correct predictions while maintaining a low misclassification rate. A lower number of false positives reduces the possibility of legitimate email attachments being incorrectly blocked, whereas a lower false negative rate minimizes the risk of malicious attachments bypassing the detection system. These observations confirm the effectiveness of the proposed framework in distinguishing malicious and benign email attachments.

## C. ROC Curve

The Receiver Operating Characteristic (ROC) curve was used to evaluate the classification ability of the ML models based on the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR). The nearer the ROC curve is to the upper left corner of the graph, the better the classification the curve will be able to differentiate between malicious and benign samples more effectively [9, 10].

The experimental result showed that the RF classifier outperformed the LR and LSVC in terms of ROC. This means that it has a greater ability to identify malicious email attachments and fewer false positives. The RF model shows the best performance on the ROC curve, suggesting it is a good model for balancing detection sensitivity and classification reliability [11, 12].

Moreover, the ROC analysis confirms the suitability of the proposed intrusion detection system based on the RF algorithm in terms of high performance and reliability in detecting malware. The larger the area under the ROC curve, the better the model's ability to distinguish between malicious and benign samples. [9]. Fig. 3 Presents the ROC curves of the evaluated machine learning models [9, 11].

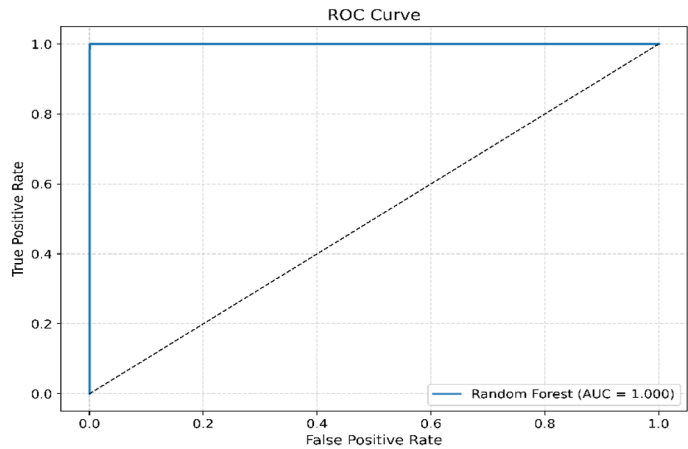
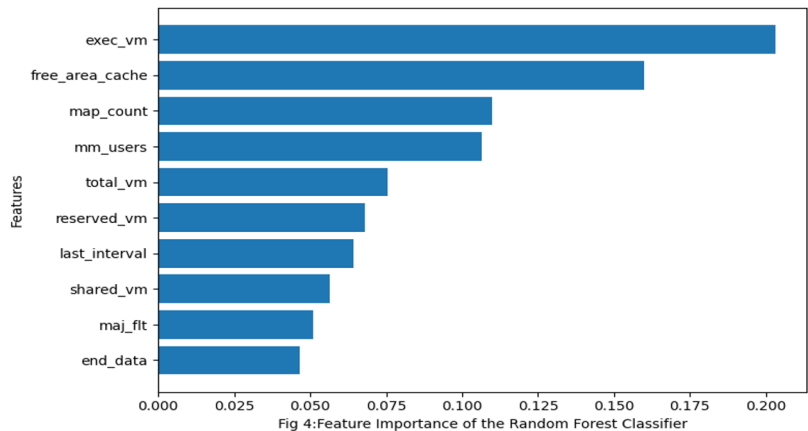


Fig. 3. Comparative Classification Performance of Machine Learning Models Using ROC Analysis

The ROC analysis further validates the effectiveness of the proposed intrusion detection framework. A ROC curve closer to the upper-left corner represents better discrimination between malicious and benign samples. Among the evaluated classifiers, the RF model exhibited the best overall ROC characteristics, indicating a strong balance between detection sensitivity and false alarm rate. This demonstrates that the proposed model can reliably detect malware while maintaining stable classification performance.

D. Feature Importance Analysis

The feature importance analysis was performed using the RF classifier to identify the most influential features in classifying malicious and benign email attachments. The importance score assigned to each feature reflects its relative contribution to the decision-making process of the model.



**Figure 4** illustrates the top ten most important features identified by the RF classifier. Among these, `exec_vm`, `free_area_cache`, and `map_count` contributed most significantly to the classification process, indicating that execution- and memory-related behavioral characteristics play a major role in distinguishing malicious and benign email attachments. Other important features, including `mm_users`, `total_vm`, `reserved_vm`, `last_interval`, `shared_vm`, `majflt`, and `end_data`, also influenced the prediction process to varying degrees. These findings suggest that system-level behavioral features provide valuable information for identifying malicious activities and enhance the overall effectiveness of the proposed intrusion detection framework.

## E. Discussion

The experimental results demonstrate the feasibility of ML techniques for detecting malicious email attachments. Random Forest classification model outperformed in terms of the best classification accuracy as compared to other evaluated classification models which were Linear SVC (76.28%), LR (74.82%), and RF (91.28%). This improvement is attributed to the ensemble property of RF, which is able to learn the more complex pattern in the feature space through the combination of several decision trees, and also restrains the overfitting.

LR and Linear SVC, on the other hand, had lower classification accuracies, indicating that they were not as effective at capturing the complex interactions in the data. To further add to the reliability of the evaluation process, 5-fold cross validation approach was taken, which involved evaluation of the model performance for multiple splits of the data.

To conclude, it can be noted that the RF approach is one of the promising approaches for early detection of malicious email attachments. However, the study was done with a malware collection that is available to the public, not with the actual email attachment data. Future research will include testing the proposed framework with real-world data sets and incorporating explainable artificial intelligence (XAI) tools to build transparency and interpretability within the models and improve their threat analysis capabilities.

## VII. Conclusion

This study introduced a ML approach for malicious email attachment detection based on behavioral features of malware. A publicly available set of 100,000 samples was used to train and test the classification models. Data preprocessing was used in an attempt to enhance the quality of the data and eliminate non-informative attributes before developing the models. [12].

Three supervised ML algorithms (Random Forest, Logistic Regression, Linear Support Vector Classification (Linear SVC)) were implemented and evaluated. The model with the highest classification accuracy was the Random Forest model, with an accuracy of 91.28%, indicating that it performed best for differentiating between malicious and benign files. Results show the usefulness of ML-based methods to identify malicious attachments before they are executed, and improve protection methods for emails. [9, 12]. The proposed model is a potential solution to ensure the security of email and detect malicious attachments early. It can help proactively detect threats, thereby reducing the chance of malware infections and improving the security of the organization's computing environment.

## VIII. Future Work

Future work can be performed according to the proposed framework to evaluate the framework with the real-world email attachment data sets in enterprise environment. This assessment would give a more realistic idea of the performance of the model when deployed in real-world situations and give it a greater applicability to different attack scenarios. [12]

In the future, more advanced ML techniques, i.e. Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and transformer networks, could also be explored, potentially enhancing the detection of new and sophisticated malware threats. Additionally, if the incorporation of Explainable Artificial Intelligence (XAI) techniques can provide increased interpretability to the model predictions, it would allow security analysts to understand more about what affects the classification outcomes [12].

A fascinating direction to look is combining and assessing the framework in real-time email security solutions. These deployment studies would help determine the scalability, detection delay, and the efficiency of the system for enterprise deployments.

## References

- [1] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques," *IEEE Access*, vol. 10, pp. 65703–65727, 2022.
- [2] K.-N. Khattak, Sabahat, N. Iqbal, Zeeshan, and J. Choi, "Cyber Security Challenges: A Strategic Model for the Software Development Organizations," 2026.
- [3] S. Atawneh and H. Aljehani, "Phishing Email Detection Model Using Deep Learning," *Electronics*, vol. 12, no. 20, Art. no. 4261, 2023, doi: 10.3390/electronics12204261.
- [4] C. Magishashini, S. Prabu, P. Ravivarma, and A. Akash, "Development of Opaline Attachment Defence System for Proactive Detection and Sanitization of Malicious Email Files," 2026.
- [5] R. Sihwail, K. Omar, and K. A. Zainol Ariffin, "A Malware Detection and Classification Approach Based on Memory Forensics," *IEEE Access*, vol. 8, pp. 43738–43752, 2020.
- [6] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [7] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [8] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.
- [9] V. R. Hananto, U. Serdült, and V. Kryssanov, "A Text Segmentation Approach for Automated Annotation of Online Customer Reviews, Based on Topic Modeling," *Applied Sciences*, vol. 12, no. 7, Art. no. 3412, 2022.
- [10] A. Ayodeji, A. Di Buono, I. Pierce, and H. Ahmed, "Wavy-attention network for real-time cyber-attack detection in a small modular pressurized water reactor digital control system," *Nuclear Engineering and Design*, vol. 424, Art. no. 113277, 2024, doi: 10.1016/j.nucengdes.2024.113277.
- [11] S. P. Jani and M. A. Khan, Eds., *Applications of AI in Smart Technologies and Manufacturing: Volume 2*, 1st ed. Boca Raton, FL, USA: CRC Press, 2025.
- [12] A. Punitha, S. Syedakbar, and S. Jeyasudha, Eds., *Intelligent and Sustainable Systems: AI, Green IoT, and Adaptive Automation in Electrical and Communication Technologies*, 1st ed. Boca Raton, FL, USA: CRC Press, 2026.

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

