International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0205
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1256 – 1260

Static and Dynamic Malware Analysis Using Machine Learning

K. Anusha[1], A. Charesshmaa[2], Y. Jayalakshmi[3], A. Gangadhar[4] K. Hrushikeswarreddy[5]

[1]Assistant professor, Dept. of CSE, Annamacharya Institute of Technology & Science, Boyanapalli, Andhrapradhesh, India.

[2,3,4,5]UG Scholar, Dept. of CSE Annamacharya Institute of Technology & Science, India.

Email ID: *anushaajay11153@gmail.com[1], charesshmaaaluru@gmail.com[2], jayalakshmi0131@gmail.com[3], gangadharroy9999@gmail.com[4], kommahrushikeshwarreddy@gmail.com[5]*

Abstract

The necessity for efficient, automated detection systems that can accurately identify malicious software has been brought to light by the growing sophistication of malware assaults. Despite their value, conventional static as well as dynamic evaluation tools sometimes lack the flexibility to adjust to changing infection strategies. In order to improve detection capabilities, this study employs a machine learning-based strategy that combines static and dynamic analysis of malware. To guarantee data quantity and importance for analysis, the system uses a large malware dataset and then applies data preparation techniques including feature scaling and normalization. The Extra-Trees-Classifier streamlines the classification process by identifying the most informative features through feature selection optimization. A Random Forest model, the main classifier, is used to evaluate the generated data and categorize files as either malware-free or infected. With a high precision of 99.42%, this model effectively and with little mistake distinguishes between harmful and benign files. This system offers a dependable, high-performance solution for proactive malware identification, which is crucial for contemporary cybersecurity applications, by fusing strong feature engineering with sophisticated classification approaches.

Keywords: *Cybersecurity, Data Preprocessing, Feature Scaling, Machine Learning, Malware Analysis, Malware Detection, Random Forest and Dynamic Analysis.*

1. Introduction

In order to safeguard sensitive data and maintain system integrity, effective malware detection techniques have become more necessary in recent years due to the exponential increase in digital threats [1]. Malware variations are become increasingly sophisticated as technology develops, frequently eluding conventional antivirus software and avoiding detection systems. Malware assaults may cause serious operational and financial consequences for both persons and businesses, ranging from the theft of private information to the disruption of vital infrastructure. Because of this, cybersecurity is facing an urgent challenge that calls for more flexible, automated methods to precisely and successfully identify and categorize dangerous data. Millions of distinct malware samples are discovered annually, making malware attacks one of the most common cyberthreats globally, according to global data. One of the most costly forms of malware, ransomware, is expected to cost billions of dollars worldwide, and projections showed a 50% rise in ransomware attacks in 2023 alone. Additionally, more than 30% of all recorded security-related problems involved phishing attempts, which frequently spread malware and impact businesses of all sizes [2]. The COVID-19 pandemic dramatically increased the issue as firms transitioned to remote work arrangements, leaving their networks at greater risk. Attackers took

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0205
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1256 – 1260

advantage of lax security measures [3] brought about by an increase in online transactions and telecommuting, which resulted in a significant rise in malware assaults, especially those involving phishing, ransomware, and spyware. Attempts to strengthen cybersecurity defenses have developed as a result of this growing susceptibility, although conventional methods frequently lack the flexibility needed to handle malware's ever-changing nature [4]. The problem is obvious: in order to identify and react to malware quickly and accurately, new, smart technologies must be put in place.

2. Literature Survey

Liu et al. proposed Information security is seriously threatened by the explosive growth of malware. Conventional anti-virus software has trouble identifying new applications and categorizing unknown infections [5]. Data processing, decision-making, and detection are all components of a machine learning based analysis of malware system. After testing on more than 20,000 malware cases, it was able to detect new malware and categorize unknown malware with a 98.9% accuracy rate. Rathore et al. over the past ten years, malware has increased dramatically, costing businesses money. Deep learning and machine learning methods [6] are now being used by antivirus software businesses to analyze and detect viruses. While Big Auto-Encoders are excessive for feature reduction, Random Forest performs better with opcode frequency than Deep Neural Networks. Variance Threshold and other basic functions work better. Future directions, constraints, open research topics, and other obstacles are covered throughout the lesson. Gavrilut et al. in order to reduce false positives, the study offers a flexible framework for applying machine learning methods to differentiate between infection and clean data. Both cascade kernelized and cascade one-sided perceptron [7] are used in the framework. Vinaya Kumar et al. in the digital era, malware assaults represent a serious security risk, yet the available detection methods are inefficient and time-consuming. Machine learning algorithms (MLAs) are helpful for malware research since recent malwares create new versions using evasive strategies [8]. More precise malware vs benign file categorization is now possible because to improved extraction and selection of features approaches that capture key characteristics of dynamic and static information. Mahindru et al. proposed a web-based structure for detecting malware on Android devices, is introduced in the research paper. The framework builds a model using several machine learning techniques [9] by utilizing dynamic analysis and characteristics from real-world applications. Xu et al. presents an original structure for hardware-assisted detection of malware that uses machine learning to monitor and categorize access to memory patterns. This method boosts automation and decreases user input on certain malware signatures [10].

3. Data Collection and Preprocessing

High-quality information gathering and careful preparation are essential for accurate analysis in malware detection. The first step in the data collection procedure is to compile a complete dataset that contains both dangerous and benign files. This stage is critical because the detection algorithm cannot successfully generalize across multiple threat situations without a well-balanced dataset that represents distinct malware kinds, families, and benign samples. The dataset generally contains of static file features, such metadata and code groups, as well as changing data, such as calls to system and network activities acquired during program execution in an appropriately controlled setting. The capacity of the system to manage novel or unknown malware variants is improved by ensuring variety in data sources. After being gathered, the raw data is cleaned to get rid of duplicate entries, damaged files, and irregularities that can interfere with the detection process. Data cleaning [11] ensures that only pertinent information is kept for additional analysis by reducing noise in the dataset. This stage may involve addressing missing values that may otherwise provide biased findings and eliminating superfluous or unnecessary characteristics that do not aid in malware identification. Normalization methods [12] like min-max scaling or z-score normalization are frequently employed to put characteristics within a constant range because of the considerable dimensionality and unpredictability in malware data. Since malware detection algorithms rely on certain

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0205
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1256 – 1260

traits that differentiate harmful from benign behavior, feature engineering is crucial to gleaning valuable insights from unprocessed data. Static properties that aid in identifying structural trends inside files, such byte patterns, import functions, and opcode frequency ranges, are frequently the source of key features. An organized partitioning technique guarantees efficient training of the detection system while maintaining enough data for objective assessment. The test set is put aside for evaluating the system's performance in the actual world, the validation set is used to adjust hyperparameters [13], and the training data set is used to identify patterns.

4. Principles and Methods

The first step in the process for this kind of malware detection tool is the thorough acquisition of a varied dataset that includes both benign files and different malicious samples. The machine learning algorithms are trained and assessed using this dataset as the basis. In order to guarantee that the framework can identify and distinguish between different traits displayed by different malicious software, it incorporates a broad variety of malware kinds (as shown in Fig.1). To make sure the model runs on a constant scale which is essential for algorithms that depend on distance metrics the pertinent characteristics are then standardized. Techniques for feature extraction are used to extract valuable information from the data, collecting crucial features including file metadata, execution behavior, and other pertinent traits that aid in distinguishing (Figure 1)

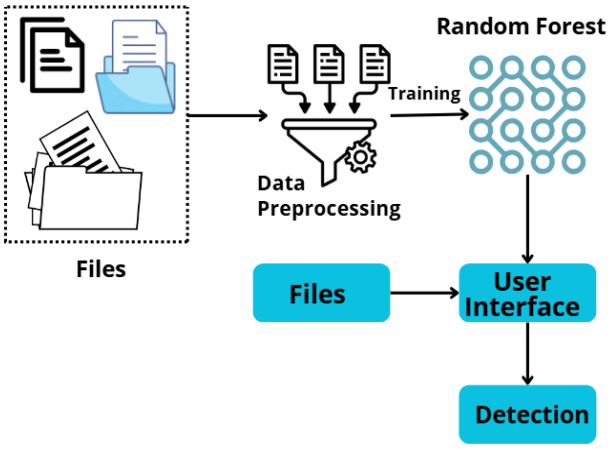

Figure 1 Working Methodology

4.1. Random Forest Classifier

In addition to its excellent classification accuracy and capacity to handle big datasets, the Random Forest approach [14] is a potent and adaptable machine learning technique that is frequently employed in malware detection. As a collaborative learning method, Random Forest builds a group of decision trees that have all been trained on various dataset subsets. Random Forest solves the drawbacks of single models by combining the predictions of several decision trees, increasing stability and lowering the risk of overfitting. Because of its resilience and versatility, this model is a good option for determining if a file is dangerous or safe, particularly when handling the complicated, multifaceted data that is frequently encountered in malware research. This is especially helpful in situations when malware files have a range of behavioral and structural traits [15]. Random Forest has several built-in benefits when it comes to managing unbalanced datasets, which are frequently encountered in malware detection. Random Forest can manage databases where the amount of benign files far outweighs the quantity of malicious ones since every tree in the forest has been developed on a distinct sample of data. This feature reduces the possibility that the model may become skewed in favor of the majority of class, which could otherwise result in a large number of false negatives. Random Forest increases the model's overall efficacy by distributing the predictions throughout several trees, which guarantees that even occurrences of minority classes like uncommon malware types have some likelihood of being discovered.

5. Results

An easy-to-use streamlit interface makes it possible to view the malware detection system's results, which facilitates effective file classification and straightforward result interpretation. Through the program, users may upload files directly. The system swiftly analyzes these files and indicates if they are deemed safe or possibly dangerous. This simplified method greatly speeds up reaction times when detecting threats by ensuring that users even those with little technical knowledge can take use of the capabilities of sophisticated machine learning

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0205
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1256 – 1260

frameworks for finding malware in real-time. User accessibility is given top priority in the interactive interface design, which makes it easy to navigate and quickly retrieve detection findings. (Figure 2)

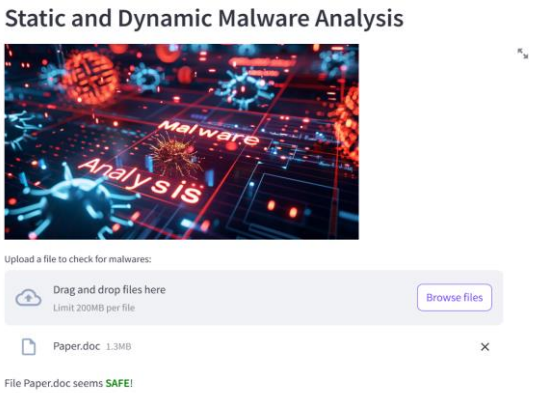

Figure 2 User Interface for Detection

The accuracy of the system is a reflection of the efficacy of thorough data preparation methods, which guarantee that only pertinent properties are examined. This preprocessing pipeline, which includes operations like extraction of features and scaling, improves detection efficiency and accuracy while preserving a high degree of precision in separating malicious files from safe ones. Both common and uncommon malware may be detected because to the model's high accuracy rate and sensitivity to a variety of malware kinds, which helps the system adjust to new and changing threats. In the current cybersecurity environment, where new malware varieties constantly pose new difficulties, this flexibility is essential. (Figure 3)

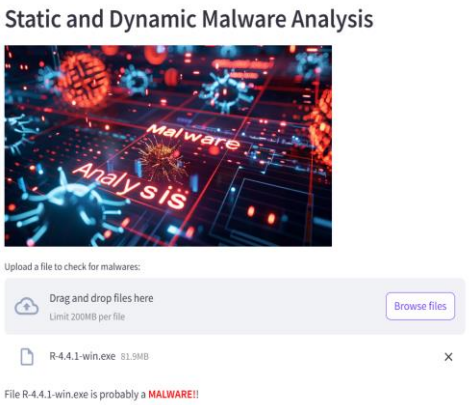

Figure 3 User Interface for Detection

Additionally, the system is made to provide in-depth analysis and reporting, which is advantageous for cybersecurity experts or IT managers. The tool offers a description of several traits and behaviors found during analysis, providing more insights into possible dangers than just identifying infected files (as shown in Fig.2&3). Users may learn more about the characteristics and potential source of each malware incident thanks to this thorough reporting, which can help them develop more focused mitigation techniques. The comprehensive data also facilitates trend analysis and historical monitoring, enabling firms to trace the kinds and frequency of attacks they have faced over time. The outcomes demonstrate a comprehensive, user-friendly, and effective malware detection solution offered by the streamlit platform. The system is a useful tool for both people and enterprises, enabling proactive and knowledgeable malware control due to its high accuracy, quick detection capabilities, and comprehensive reporting features.

Conclusion

This malware detection solution uses sophisticated machine learning algorithms and a well-organized streamlit display for real-time analysis to identify harmful files in a very efficient and user-friendly manner. Strong data preprocessing and the model's flexibility in handling different kinds of malware produce a high degree of precision that guarantees accurate identification of both well-known and new threats. Accessibility and usability are improved by the system's simplified and interactive interface, which enables users of various technical skill levels to categorize files, evaluate detection results, and comprehend the confidence of each categorization with ease. Furthermore, the program's ability to analyze big files with little latency meets real-world, high-throughput cybersecurity requirements, making it appropriate for both individual and corporate use. Users are empowered to make well-informed decisions about threat mitigation and management by this tool's clear, well-organized findings and insights into the traits of discovered malware. This system's overall efficacy, usability, and scalability demonstrate its potential to be a useful tool in the ever-changing field of identifying

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2025.0205
e ISSN: 2584-2854
Volume: 03
Issue:04 April 2025
Page No: 1256 – 1260

and strengthening safety measures against a constantly changing threat landscape.

References

[1]. Aslan, Ömer Aslan, and Refik Samet. "A comprehensive review on malware detection approaches." IEEE access 8 (2020): 6249-6271.
[2]. Goenka, Richa, Meenu Chawla, and Namita Tiwari. "A comprehensive survey of phishing: Mediums, intended targets, attack and defence techniques and a novel taxonomy." International Journal of Information Security 23.2 (2024): 819-848.
[3]. Azam, Hamza, et al. "Wireless Technology Security and Privacy: A Comprehensive Study." (2023).
[4]. Raghib, Raghib, and Dr Syed Mohammad Raghib. "Cyber Security and Data Protection in India: A National Concern." Cyber Security and Data Protection in India: A National Concern (September 03, 2024) (2024).
[5]. Liu, Liu, et al. "Automatic malware classification and new malware detection using machine learning." Frontiers of Information Technology & Electronic Engineering 18.9 (2017): 1336-1347.
[6]. Rathore, Hemant, et al. "Malware detection using machine learning and deep learning." Big Data Analytics: 6th International Conference, BDA 2018, Warangal, India, December 18–21, 2018, Proceedings 6. Springer International Publishing, 2018.
[7]. Gavriluţ, Dragoş, et al. "Malware detection using machine learning." 2009 International multiconference on computer science and information technology. IEEE, 2009.
[8]. Vinayakumar, R., et al. "Robust intelligent malware detection using deep learning." IEEE access 7 (2019): 46717-46738.
[9]. Mahindru, Arvind, and Amrit Lal Sangal. "MLDroid—framework for Android malware detection using machine learning techniques." Neural Computing and Applications 33.10 (2021): 5183-5240.
[10]. Xu, Zhixing, et al. "Malware detection using machine learning based analysis of virtual memory access patterns." Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017. IEEE, 2017.
[11]. Chu, Xu, et al. "Data cleaning: Overview and emerging challenges." Proceedings of the 2016 international conference on management of data. 2016.
[12]. Saranya, C., and G. Manikandan. "A study on normalization techniques for privacy preserving data mining." International Journal of Engineering and Technology (IJET) 5.3 (2013): 2701-2704.
[13]. Boddapati, Mohan Sai Dinesh, et al. "Creating a Protected Virtual Learning Space: A Comprehensive Strategy for Security and User Experience in Online Education." International Conference on Cognitive Computing and Cyber Physical Systems. Cham: Springer Nature Switzerland, 2023.
[14]. Rigatti, Steven J. "Random forest." Journal of Insurance Medicine 47.1 (2017): 31-39.
[15]. Duckworth, Renée A., Ahva L. Potticary, and Alexander V. Badyaev. "On the origins of adaptive behavioral complexity: developmental channeling of structural trade-offs." Advances in the Study of Behavior. Vol. 50. Academic Press, 2018. 1-36.