IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Detection And Identification Of Malware Using Random Forest Algorithm

[1]Mr. ARAVINDHAN K, [2] Ms. NIHAL BABA,

[1]Student, [2]Assistant Professor,

[1] Mr. ARAVINDHAN K, M.sc CFIS, Department of Computer Science Engineering, Dr.MGR UNIVERSITY, Chennai, India

[2] Ms. NIHAL BABA, Assistant Professor, Cyber forensics and Information Security, University of Madras, Chepauk, Chennai, India

Abstract: Malware is an increasing threat to cybersecurity, causing significant harm to systems, data, and networks worldwide. As cyberattacks become more sophisticated, it's crucial to develop reliable and effective methods for detecting and preventing malware. This project addresses this issue by using the Random Forest algorithm, a well-established machine learning technique known for its ability to handle complex datasets and deliver high accuracy. The model is trained to analyse features extracted from executable files, such as file size, code structure, and patterns in behaviour, to distinguish between benign and malicious software. One of the key advantages of the Random Forest approach is its ability to adapt to new and evolving types of malware, ensuring it stays effective as threats change. Additionally, its resistance to overfitting and scalability make it a robust choice for detecting even the most sophisticated malware. Ultimately, this project aims to enhance cybersecurity by providing a flexible, precise, and scalable solution for malware detection.

Index Terms - Malware Detection, Random Forest Algorithm, Machine Learning, Predictive Accuracy, Malicious Software, Malware Identification.

I. INTRODUCTION

The paper “Detection and identification of malware using random forest algorithm” focusses on detecting the malicious file that is present in our device. The algorithm's[1] ensemble nature enhances its ability to handle high-dimensional data, mitigate overfitting, and deliver accurate results in identifying malicious software. By implementing this approach, the project aims to create a reliable and scalable solution for early detection of malware, contributing to more robust cybersecurity measures

The paper focuses on detecting and identifying malware using a machine learning-based approach with the Random Forest algorithm. Random Forest is a powerful machine learning method that creates multiple decision trees using random samples of data and features. Each tree makes a prediction, and the final result is based on the majority vote or average of all trees from that the result will be highly accurate and most reliable. [2]

The "Detection and Identification of Malwares Using Random Forest Algorithm" offers significant potential in improving cybersecurity by enabling accurate detection of various malware types, including viruses, ransomware, and Trojans. It is adaptable to evolving threat landscapes and scalable for analyzing large datasets, making it suitable for organizations of all sizes. The solution can be integrated into antivirus software and intrusion detection systems, providing real-time protection and enhancing security frameworks. [3]

The research questions would be framed around understanding the malware complexity and how to detect the new complex malwares[4] without confusing with legit files and it is really a hard part in this project because nowadays the malicious content patterns and structures are improving day by day like the security standards increases.

The solution can be integrated into antivirus software and intrusion detection systems, [5] providing real-time protection and enhancing security frameworks. Additionally, the project serves as a foundation for further research in advanced machine learning techniques for malware detection and contributes to raising awareness about intelligent, automated cybersecurity measures.

II. LITERATURE SURVEY

Ajay Kumar, Kumar Abhishek, Shishir Kumar Shandilya, Muhammad Rukunuddin Ghalib et al.,[1] Malware Analysis Through Random Forest Approach: This paper gives unique and comprehensive detail along with a proposed system for malware detection using Machine learning and Deep Learning techniques by integrating both behavior-based detection methods and signature-based methods.

Yi-chen Wu, You-lun Chang et al.,[6] Ransomware Detection on Linux Using Machine Learning with Random Forest Algorithm: Ransomware continues to give a significant threat to cybersecurity, particularly affecting critical data that is running on Linux. This paper focusses on the random forest algorithm for detecting ransomware on Linux systems offers a significant advance method, leveraging machine learning to enhance detection accuracy.

Seong Il Bae, Gyu Bin Lee, Eul Gyu Im et al.,[7] Ransomware detection using machine learning algorithms: The number of ransomware variants has increased rapidly every year, and ransomware needs to be mentioned separately from the other types of malwares to protect users' machines from ransomware-based attacks. Ransomware is like other types of malwares in some aspects, but other characteristics are clearly different.

Mohammed S. Alam, Son T. Vuong et al.,[8] Random Forest Classification for Detecting Android Malware: Android is the most popular OS (operating system) for smartphones. It is also an easy target for the malware authors.

Ban Mohammed Khammas.,[9] Ransomware Detection using Random Forest Technique: The current study involved using random forest classifier with a comprehensive analysis to the effect of both tree and seed numbers on the ransomware detection. The results showed that 100 number of trees with seed number of 1 achieved best results in terms of time-consuming and with more accuracy.

Carti Irawan, Teddy Mantoro, Media Anugerah Ayu et al.,[10] Malware Detection and Classification Model Using Machine Learning Random Forest Approach: Computers are easily infiltrated by various malware programs that can interfere with and even damage user file.

Alan Mills, Theodoros Spyridopoulos, Phil Legg et al.,[11] Efficient and Interpretable Real-Time Malware Detection Using Random-Forest: Malicious software, often described as malware, is one of the greatest threats to modern computer systems, and malware authors continue to develop more advance methods to access and compromise critical data.

III. PROPOSED METHODOLOGY

This project aims to detect and identify the malware that present in the files by using the powerful Random Forest algorithm which average the all the trees and choose a decision which sounds more acurate, with the past data of the malware structures like coding, origin of the file etc. and python is the programming language used to do this project work because of its readability, ease of use and extensive library characteristics. Goals and constraints:

- To secure the data and privacy of the user.
- To protect the computational resources of the organization.
- To avoid the risks.

3.1 MALWARE

Malware, short for malicious software, encompasses any type of software intentionally designed to disrupt, damage, or gain unauthorized access to computer systems, networks, or data. Cybercriminals, commonly referred to as hackers, create these programs to achieve a variety of harmful objectives, such as stealing sensitive information, encrypting data for ransom, spying on user activities, or damaging files and infrastructure. Some common types of malware include viruses, worms, Trojans, ransomware, spyware, and adware—each operating in unique ways to exploit system vulnerabilities.

In recent times, [12] malware has evolved to become increasingly sophisticated, often employing advanced techniques to bypass traditional detection systems. These modern threats are frequently polymorphic, meaning they can change their code to evade signature-based detection methods that rely on identifying known patterns. As a result, conventional antivirus tools are often ineffective against newly developed or modified malware strains.

3.2 RANDOM FOREST ALGORITHM

Random Forest is a versatile supervised machine learning algorithm that is capable of handling both classification and regression problems. It is based on the principle of ensemble learning, which involves combining the predictions of several models to produce a more accurate and stable result than any individual model could achieve on its own.

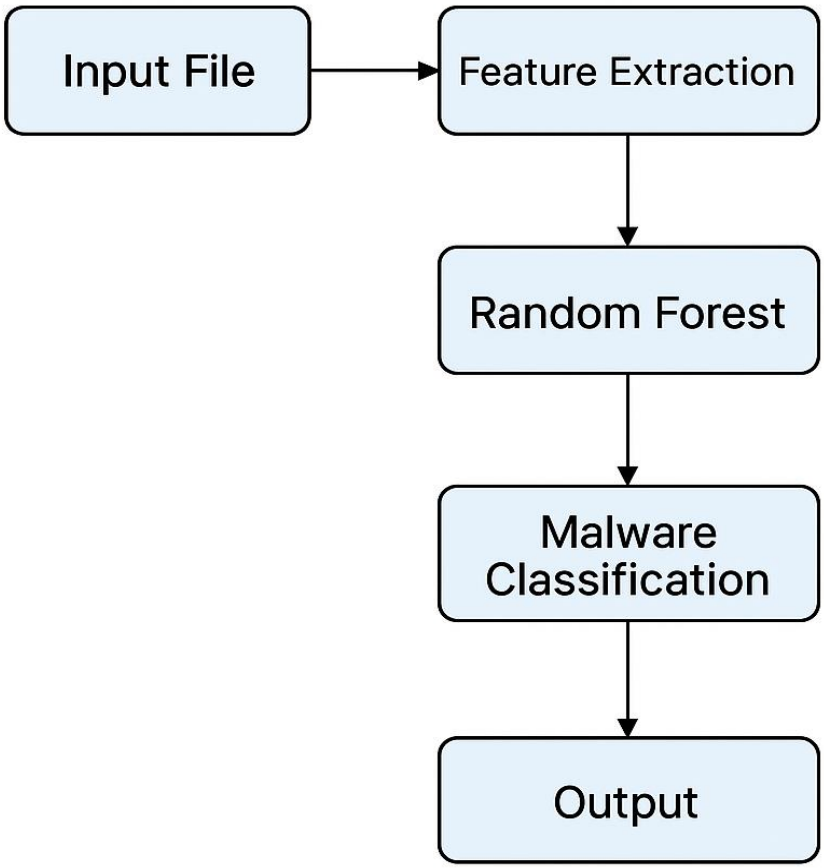
The core idea behind [13] Random Forest is to create a "forest" composed of numerous decision trees. Each tree in the forest is trained on a different random subset of the training data, which is selected through a technique called bootstrapping (random sampling with replacement). Additionally, when splitting nodes in each tree, the algorithm selects from a random subset of features rather than considering all available features. This randomness helps to reduce the correlation between individual trees and enhances the diversity of the model, which in turn improves its generalization capability.

3.3 INPUT FILE

The system is designed to support a wide range of file types, allowing users to upload virtually any format for malware analysis. This includes commonly used file formats such as images (e.g., .jpeg, .jpg, .heic), videos (e.g., .mp4, .mov, .avi, .mkv), and audio files (e.g., .mp3, .aac, .flac, .wav). In addition to multimedia formats, the system also accepts documents and compressed files such as PDFs, Word documents (.doc, .docx), and archive files like .zip and .rar.

This broad compatibility is essential in today's cybersecurity landscape, as cybercriminals are increasingly embedding malicious code within files that appear harmless at first glance. Attackers often disguise malware within files that mimic legitimate content—such as photos, videos, or documents—making it more difficult for users to suspect any foul play. For example, a seemingly normal PDF or a song file might contain hidden scripts or payloads that, when opened, trigger malicious activities in the background.

3.4 SYSTEM ARCHITECTURE



IV. FINDINGS

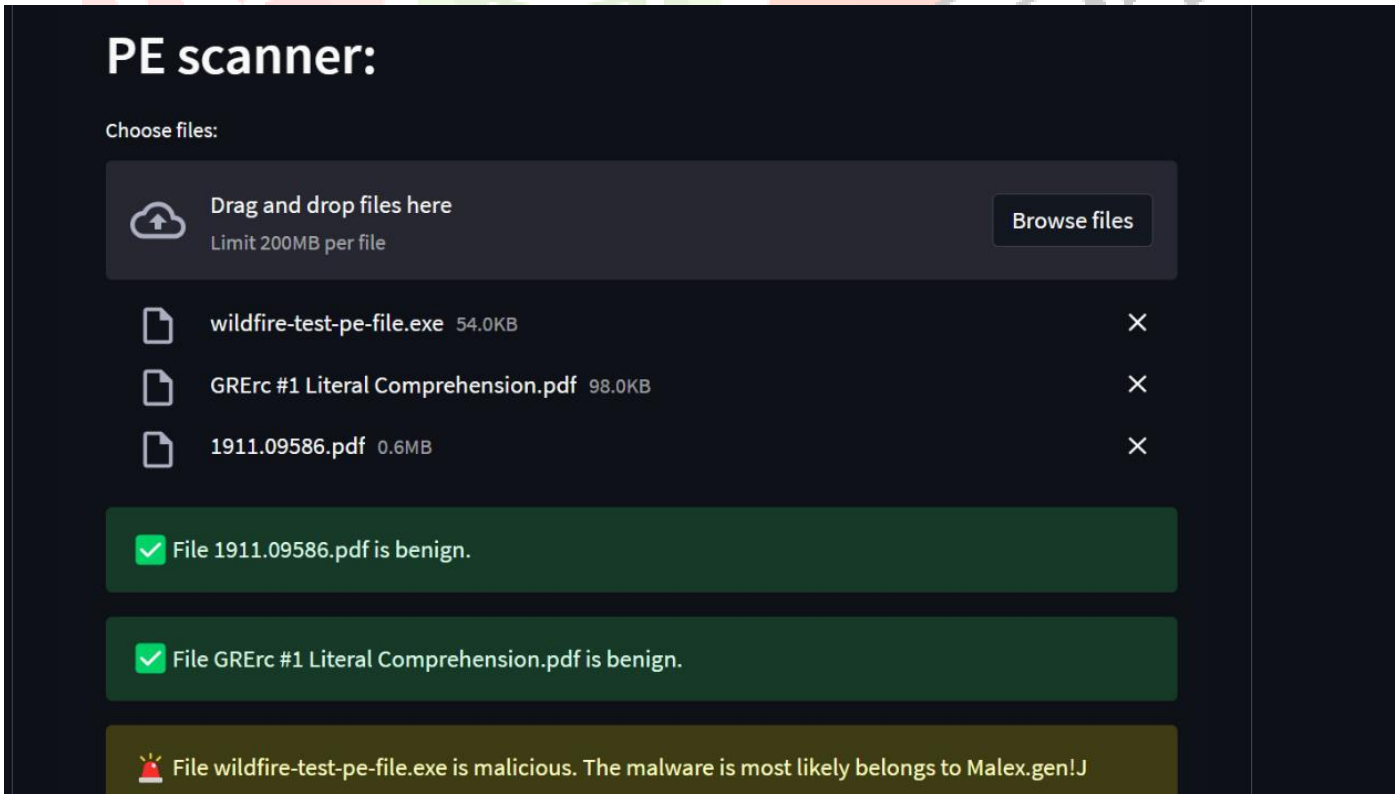


Figure 4.1 files that are detected and classified as legit or malware

As illustrated in Figure 4.1, the system provides a user-friendly interface that allows users to upload files from their device to determine whether the files are safe or malicious. This can be done easily either by dragging and dropping the desired files into the designated area or by manually browsing and selecting the files from the device's storage.

Once the files are uploaded, the system begins scanning them using a detection algorithm designed to identify malware. After the scan is complete, the interface displays the results in a clear and intuitive manner. Each file is marked with a specific visual indicator based on the outcome of the scan.

A green checkmark icon is used to signify that a file is legitimate—meaning it is free of any malicious content and considered safe for use. On the other hand, if a file is found to contain harmful elements or behaves suspiciously, it is flagged with a warning icon (often represented by an alert symbol), indicating that the file poses a potential threat and may contain malware.
This visual feedback makes it easy for users, even those without technical expertise, to quickly understand which files are secure and which ones should be handled with caution or removed from their system. The overall design enhances user experience by combining simplicity with effective malware detection capabilities.

V. ACKNOWLEDGEMENT

I would like to express our sincere gratitude to all those who contributed to the successful completion of this research work.

First and foremost, we extend our heartfelt thanks to Dr. M.G.R. Educational and Research Institute, Chennai, for providing us with the necessary infrastructure and academic environment to carry out this project.

I deeply thankful to Ms. Nihal Baba, Assistant Professor, Cyber forensics and Information Security, University of Madras, Chepauk, Chennai, India, for her invaluable guidance, continuous support, and insightful feedback throughout the research. Her expertise and mentorship were instrumental in shaping the direction and quality of this work.

I also extend our appreciation to our colleagues and peers who provided constructive suggestions and moral support throughout this journey. Special thanks to the faculty of the Department of Computer Science Engineering for their encouragement and academic assistance.

VI. CONCLUSION

In conclusion this paper focuses on detecting and identifying the Malwares in the file to maintain the privacy of the user and to maintain the integrity of the data. By implementing this we can protect the digital asset and we can avoid the risk of getting attacked by the hackers and to maintain the uninterrupted work flow.

The future work for the paper Detection and Identification of malware using random forest algorithm can be integrated with Antivirus software and other security measures to avoid the malwares in the first place and the data sets can be updated by facing the new set of malwares in the future, and it will be helpful to train the AI data models.

REFERENCES

[1] Kumar, A., Abhishek, K., Shandilya, S. K., & Ghalib, M. R. (2020). Malware analysis through random forest approach. Journal of Web Engineering, 19(5–6), 795–818. https://doi.org/10.13052/jwe1540-9589.195610

[2] Akinwole, A. K., Asafe, Y. N., & Oludayo, O. I. (2023). Malware detection system using mathematics of Random Forest classifier. International Journal of Advances in Scientific Research and Engineering, 9(3), 45–53. https://doi.org/10.31695/IJASRE.2023.9.3.6

[3] Kamalakanta Sethi, Shankar Chaudhary, Bata Krishna Tripathy, and Padmalochan Bera. (2017). A novel malware analysis for malware detection and classification using machine learning algorithms DOI:10.1145/3136825.3136883

[4] Imamverdiyev, Y., & Baghirov, E. (2024). Evasion techniques in malware detection: Challenges and countermeasures. Problems of Information Technology. https://jpit.az/en/journals/334s

[5] Durrani, O. K. (2024). Machine learning for cybersecurity: Intrusion detection, malware analysis, vulnerability assessment.

[6] Ransomware Detection on Linux Using Machine Learning with Random Forest Algorithm: Yi-chen Wu and You-lun Chang

[7] Ransomware detection using machine learning algorithms: Seong Il Bae, Gyu Bin Lee, Eul Gyu Im

[8] Random Forest Classification for Detecting Android Malware: Mohammed S. Alam; Son T. Vuong

[9] Ransomware Detection using Random Forest Technique: Ban Mohammed Khammas

[10] Malware Detection and Classification Model Using Machine Learning Random Forest Approach: Carti Irawan, Teddy Mantoro, Media Anugerah Ayu

[11] Efficient and Interpretable Real-Time Malware Detection Using Random-Forest: Alan Mills; Theodoros Spyridopoulos; Phil Legg

[12] A Survey on Malware Detection Using Data Mining Techniques: Yanfang Ye, Tao Li, Donald Adjeroh, S. Sitharama Iyengar

[13] Biau, G. (2012). Analysis of a Random Forests Model. Journal of Machine Learning Research, http://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf