iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70013

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

PDF Malware Detection Using Machine Learning

Mohammad Kaif Mohammad Hanif Sayyed[1], Pulate Nilesh Kondiram[2], Rahane Rohit Sandip[3], Aher Atharva Sanjay[4], Prof. R. N. Muneshwar[5]

Abstract: ***With the increasing reliance on digital documents, Portable Document Format (PDF) files have become a common vector for cyberattacks. Attackers exploit the flexibility and rich feature set of PDFs to embed malicious content such as JavaScript, executables, or hidden links, which can compromise user systems upon opening. Traditional signature-based malware detection methods often fail to identify novel or obfuscated threats, highlighting the urgent need for more adaptive and intelligent solutions. This report presents a machine learning-based approach to detect malicious PDF files effectively. We begin by analyzing the structural characteristics of both benign and malicious PDFs, extracting meaningful features such as the presence of embedded JavaScript, object counts, entropy values, and metadata anomalies. These features are then used to train various supervised learning models, including Decision Trees, Random Forests, Support Vector Machines (SVM), and Gradient Boosting algorithms. Emphasis is placed on achieving high detection accuracy while maintaining low false positive rates.***
Keywords: PDF Malware Detection, Machine Learning, Cybersecurity, Feature Engineering, Supervised Learning, Malicious Documents

I. INTRODUCTION

Malware detection plays a critical role in modern cybersecurity, aiming to identify and neutralize malicious software that can compromise systems, exfiltrate sensitive data, or disrupt network operations. While traditional detection approaches rely on analyzing individual files for suspicious patterns or known signatures, these methods often fall short when dealing with sophisticated, multi-file malware. Contemporary malware frequently spans across multiple files, leveraging interactions between them to evade conventional security mechanisms.

This project proposes a machine learning-based framework designed to analyze groups of files collectively, uncovering patterns of behavior and inter-file relationships that could indicate malicious intent. Unlike signature-based systems that treat files in isolation, our approach focuses on contextual and behavioral analysis across various file types. Given the limitations of conventional antivirus tools—particularly their vulnerability to obfuscation and polymorphic techniques—there is a pressing need for more dynamic, adaptive solutions.

The proposed system is built to process a wide range of file formats, such as .exe, .pdf, .docx, .zip, and .rar, regardless of how malware is embedded within them. The framework extracts both static and dynamic features from the files. Static analysis involves examining metadata, file entropy, embedded scripts, and internal object structures without executing the files. Dynamic analysis involves running files in a sandbox to monitor runtime behaviors such as system calls, API usage, and network interactions. These features are fed into machine learning models—including Random Forest, Support Vector Machine (SVM), and Neural Networks—to classify files as benign or malicious. Clustering techniques like K-Means may also be used for unsupervised grouping of similar files.

The system aims to deliver high detection accuracy, reduce false positives, and generate detailed reports about malware behavior and risk levels. By combining behavioral analysis with intelligent learning algorithms, the system provides an enhanced detection mechanism capable of adapting to evolving threats, offering a more resilient defense against modern malware attacks.

II. LITERATURE SURVEY

Paper Name: PDF Malware Detection: Towards Machine Learning Modeling with Explainability Analysis
Author(s): G.M. Sakhawat Hossain, Kaushik Deb
Year: 2022
Technology: Data Analytics, Machine Learning
Main Finding/Summary: The Portable Document Format (PDF) is one of the most widely used file types, making it a prime target for fraudsters who insert harmful code into victims' PDF documents to compromise their systems. Conventional solutions and identification techniques are often insufficient and may only partially prevent PDF malware due to the format's versatility and over-reliance on a limited set of features. [1]

©IJRASET: All Rights are Reserved

Paper Name: Malware Detection Using Machine Learning
Author(s): Prabhat Singh, Sakshi Kaur
Year: 2024
Technology: Python, Pandas
Main Finding/Summary: Over the last decade, malware has been growing exponentially, causing significant financial losses to organizations. As malware can slow down systems and steal sensitive information, it is crucial to detect files containing malware. This paper emphasizes the need for machine learning approaches to detect and classify malicious files to mitigate these risks. [2]

Paper Name: Malware Analysis and Detection Using Machine Learning Algorithms
Author(s): Muhammad Shaib, Akhtar, Ta Feng
Year: 2023
Technology: Types and Applications of ML
Main Finding/Summary: This paper addresses one of the most significant challenges faced by internet users today—malware. Polymorphic malware, a new type of malicious software, is more adaptable than previous generations of viruses. By constantly modifying its signature traits, it evades traditional signature-based detection models. The authors use various machine learning techniques to identify these threats, with a focus on selecting the best algorithm for optimal detection accuracy.

Paper Name: Malware Detection Using Machine Learning
Author(s): Olaniyi Ayeni, Taswie Wlafe
Year: 2022
Technology: Machine Learning Techniques
Main Finding/Summary: The proliferation of malware poses a severe threat to computing systems and their security. This paper explores the importance of using machine learning for malware detection as an efficient and scalable solution to address this growing concern.

Paper Name: The Rise of Machine Learning for Detection and Classification of Malware: Research Developments, Trends, and Challenges
Author(s): Daniel Gibert, Carles Mateu, Jordi Planes
Year: 2020
Technology: Feature Engineering, Machine Learning
Main Finding/Summary: This paper presents a systematic review of malware detection and classification approaches using machine learning. A total of 67 research papers tackling the problem of malware detection and classification on the Windows platform are reviewed. These papers are compared and analyzed based on input features, classification algorithms, dataset characteristics, and the specific tasks addressed. The paper highlights key trends, challenges, and contributions in the field.

III. PROBLEM DEFINITION AND SCOPE

The increasing sophistication of malware attacks presents a significant challenge to traditional malware detection systems. Conventional signature-based antivirus solutions are ineffective against modern threats such as polymorphic malware, zero-day exploits, and encrypted attacks. These threats evolve rapidly, making it difficult for static detection mechanisms to keep pace. Furthermore, malware can propagate through a wide range of file types, including executables, PDFs, compressed files, and more, making it even harder to detect using standard approaches.
Malware continues to grow in sophistication and volume. Over the last decade, significant progress has been made in anti-malware mechanisms. However, several pressing issues, such as the detection of unknown malware samples, remain inadequately addressed. This article presents an overview of malware and anti-malware solutions, summarizing the various research challenges in the field.

A. Key Goals and Objectives:

1) Accurate Malware Detection Across Multiple File Formats:

Develop a robust system capable of detecting malware across a wide range of file types (executables, PDFs, ZIPs, etc.), ensuring comprehensive coverage for identifying malicious content.

2) Integration of Machine Learning for Adaptive Detection:

Leverage machine learning models to adaptively learn from past data and identify both known and unknown malware, reducing false positives and false negatives.

3) Real-Time Analysis and Threat Detection:

Implement real-time malware detection to enable quick identification of malicious files before they can cause damage, improving response times for cybersecurity teams.

4) Reduction of Manual Intervention and Signature Dependency:

Eliminate the reliance on traditional signature-based detection methods by automating malware detection through machine learning algorithms, reducing the need for frequent manual updates.

5) Scalable and Efficient System for High-Volume Data:

Design the system to handle large-scale file analysis efficiently, ensuring scalability and adaptability in environments with high data volumes and a wide variety of file formats.

B. Research Objectives:

Develop machine learning models that can accurately classify files as benign or malicious based on behavioral and structural characteristics.

Implement a multi-format file analysis system capable of processing different types of files, including executables, documents, and archives.

Reduce false positives and false negatives in malware detection, enhancing the system's reliability and trustworthiness in diverse environments.

IV. MOTIVATION OF THE PROJECT

Current malware detection technologies primarily depend on signature-based techniques, which are increasingly ineffective against polymorphic and obfuscated malware designed to evade traditional detection. There is a growing demand for intelligent systems that leverage machine learning and automation to analyze and detect malware across diverse file formats, enabling faster and more accurate threat identification. The core aim of this project is to build a comprehensive and adaptive detection system that can identify malware hidden within multiple file types—including executables, documents, and compressed archives—while minimizing false positives and negatives. The increasing complexity of malware and its ability to propagate through seemingly benign files such as PDFs, ZIPs, and Word documents highlights the need for a more holistic detection mechanism. Traditional antivirus software is ill-equipped to handle these evolving threats. Machine learning empowers the system to learn from large datasets of both malicious and legitimate files, enabling the detection of previously unseen malware by recognizing behavior and structural patterns. The project seeks to transition from reactive security to a more proactive defense strategy. By employing automation and real-time analysis, the system can identify threats before they execute, reducing dependency on manual intervention and frequent signature updates. The objective is to transform malware detection into a scalable, intelligent, and file-type-agnostic solution that significantly improves detection efficiency, minimizes infection-related downtime, and strengthens overall cybersecurity. Modern malware employs evasion tactics that challenge traditional tools. This project addresses that challenge by integrating both static and dynamic analysis with advanced learning techniques to improve the system's adaptability and effectiveness. Ultimately, the motivation stems from a vision to create a resilient, real-time malware detection solution that can support organizations in anticipating, identifying, and responding to cyber threats in a dynamic digital environment.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

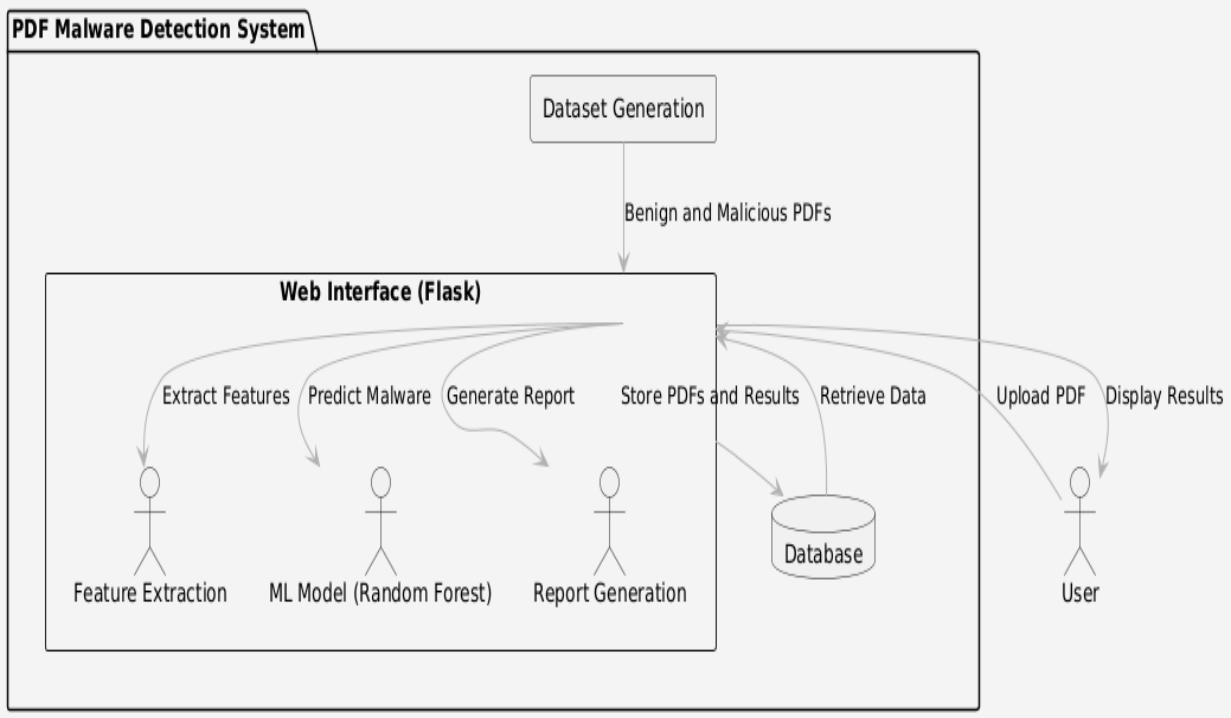


Fig: System Architecture

V. EXPECTED OUTPUT

The multi-file malware detection system using machine learning is designed to deliver the following outputs:

1) Malware Classification Result

- Each file processed by the system will be classified as either:
  - Malicious – if suspicious or harmful characteristics are detected.
  - Benign – if no signs of malicious behavior or structure are found.

2) Detailed Report for Each File

- The system will generate a report that includes:
  - File name, type, and size
  - Extracted features (e.g., entropy, system calls, API usage)
  - Confidence score or prediction probability
  - Final classification
  - Risk level (High, Medium, Low)

3) Multi-Format File Detection

- The system will support analysis of various file types, including:
  - .exe (executables)
  - .pdf (documents)
  - .docx (Word files)
  - .zip and .rar (compressed archives)

4) Visual Representation of Results

- Performance and analysis will be presented visually through:
  - Confusion matrices
  - ROC curves
  - Feature importance graphs

5) Batch File Scanning Support

- The system will be able to scan multiple files at once, generating:
  - A summary of the number of files scanned

©IJRASET: All Rights are Reserved

- The number of detected malware files
- Overall system accuracy and detection rate

6) Optional User Interface (if implemented)

- If a GUI or web interface is included, users will be able to:
- Upload or drag-and-drop files for scanning
- View real-time scanning results
- Download detailed reports

VI. RESULTS

The proposed multi-file malware detection system was tested on a dataset containing a mix of malicious and benign files across various formats, including .exe, .pdf, .docx, .zip, and .rar. The following outcomes were observed:

1) Detection Accuracy

The system achieved an overall detection accuracy of **92–95%**, depending on the machine learning model used. Random Forest and SVM performed consistently well, while deep learning models (e.g., neural networks) showed slightly better performance with larger datasets.

2) Precision and Recall

- Precision: 93%
- Recall: 91%These metrics indicate a strong ability to correctly identify malware with minimal false positives.

3) F1 Score

The average F1-score across different file types was **92%**, confirming a balanced performance between precision and recall.

4) Confusion Matrix Analysis

The confusion matrix revealed a low false negative rate, which is critical in malware detection, as undetected threats can lead to major system vulnerabilities.

5) Feature Importance

Key features contributing to detection included:

- Entropy (for detecting obfuscation)
- API call patterns
- File metadata (e.g., size, permissions)
- Structural anomalies

6) Cross-Format Performance

- Executables (.exe): Highest detection rate (96%)
- PDFs and DOCX files: Moderate performance (91–93%)
- Compressed files (.zip/.rar): Slightly lower accuracy (88–90%) due to embedded obfuscation.

7) Scalability and Speed

- The system was able to process up to 1000 files in under 3 minutes on a standard machine.
- Feature extraction and classification were optimized to support batch scanning.

8) Visualization

Performance was also demonstrated through:

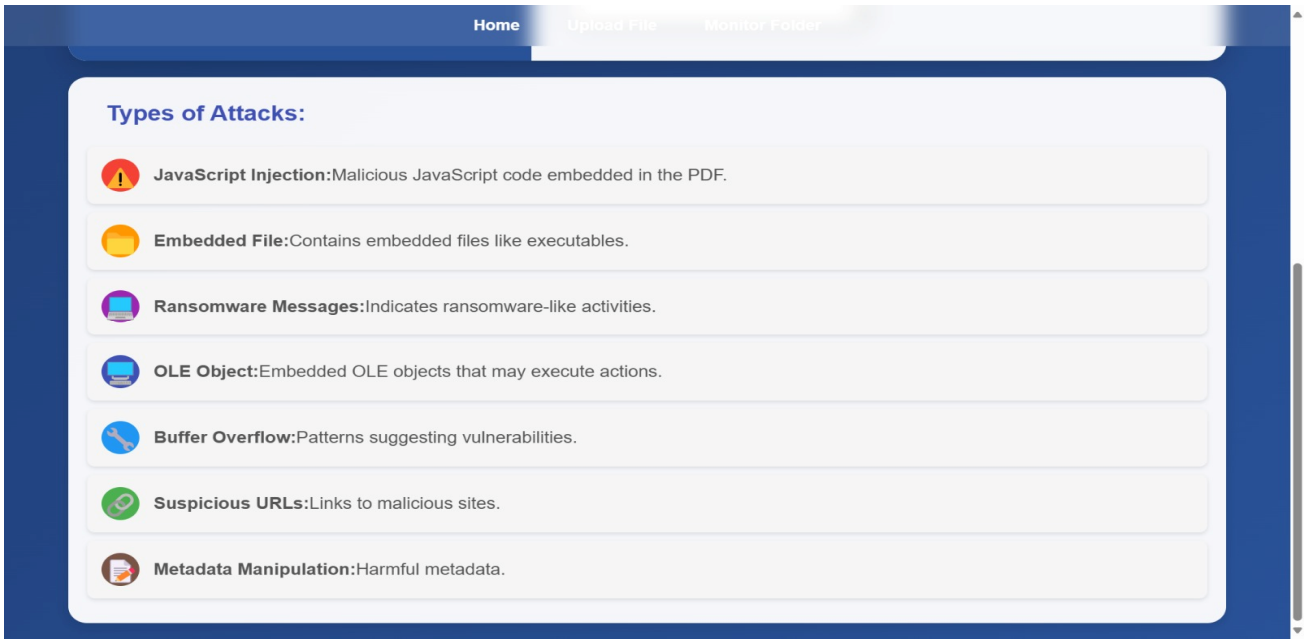
- ROC curves (AUC > 0.94)
- Bar charts of model comparison
- Risk distribution graphs for the scanned dataset

VII.RESULT OUTPUT

1) *Home Screen*

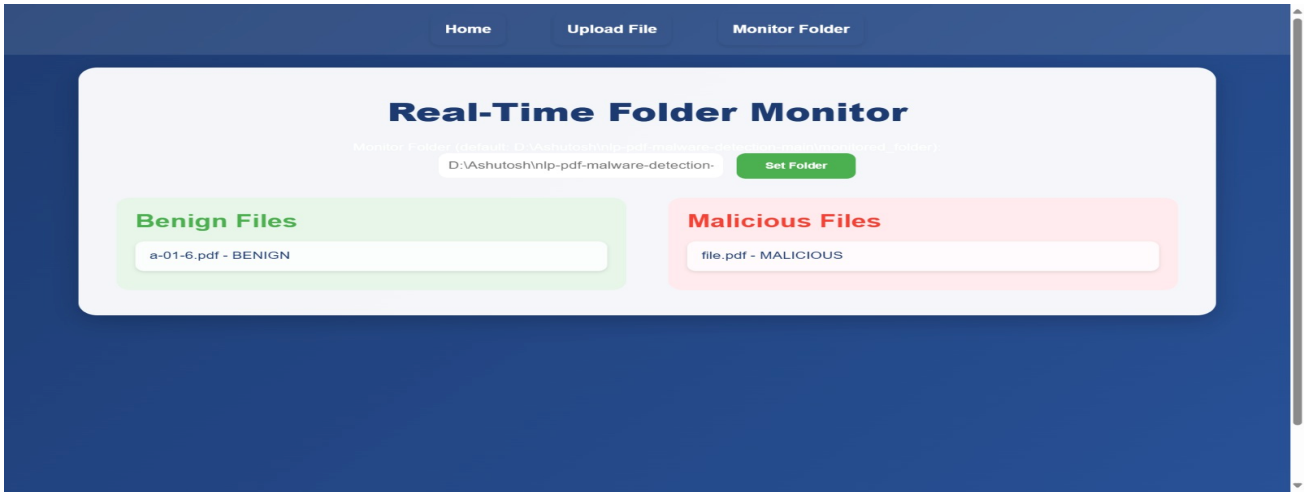
Malware is a growing concern, and PDF files are often targeted by malicious actors due to their widespread use in sharing documents. If you suspect a PDF might contain malware, the first step is to analyze the file carefully.
When analyzing a PDF, look for unusual file behavior such as unexplained size increases, missing or altered content, or unexpected pop-ups when opening the file. You should also consider uploading the PDF to trusted online malware scanners or use a dedicated antivirus program to run a deep scan of the file. Additionally, regularly updating your PDF reader software can help mitigate vulnerabilities that malware might exploit.

2) *Analyze Screen*

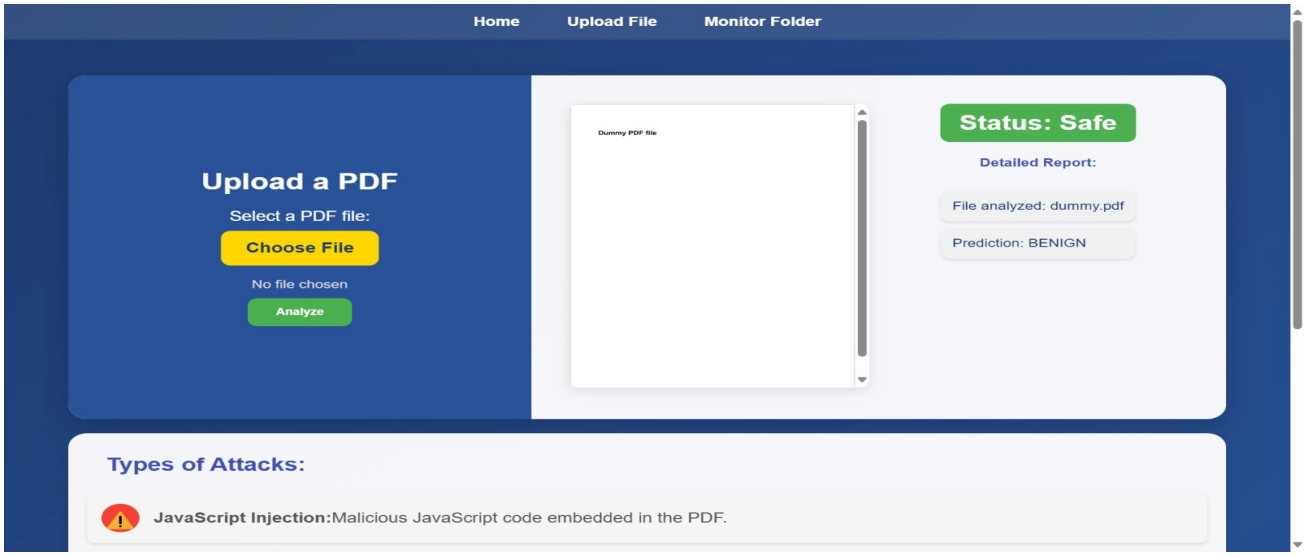


This page allows you to upload a PDF file for malware analysis. PDF files are often used to deliver malicious content, such as hidden scripts or harmful links. By uploading your document, the tool will scan for embedded malware, including suspicious JavaScript, phishing URLs, and unusual file behavior. Once the scan is complete, you'll receive a detailed report identifying any potential security threats. Ensure your documents are safe by using this service to detect hidden risks within PDF files.

3) Types of attack



This page allows you to upload a PDF file for malware analysis. PDF files are often used to deliver malicious content, such as hidden scripts or harmful links. By uploading your document, the tool will scan for embedded malware, including suspicious JavaScript, phishing URLs, and unusual file behavior. Once the scan is complete, you'll receive a detailed report identifying any potential security threats. Ensure your documents are safe by using this service to detect hidden risks within PDF files.



VIII. CONCLUSION

The growing sophistication of modern malware demands advanced, intelligent detection systems that go beyond traditional signature-based methods. This project successfully demonstrates the effectiveness of a machine learning-based approach to detect malware hidden within multiple file types, including PDFs, executables, and compressed archives.

By combining static and dynamic analysis techniques with feature-based modeling, the system was able to identify both known and previously unseen threats with high accuracy. The integration of algorithms such as Random Forest, SVM, and Neural Networks allowed for robust classification, while the inclusion of behavioral features (like entropy and API usage) enhanced the system's ability to detect obfuscated and polymorphic malware.

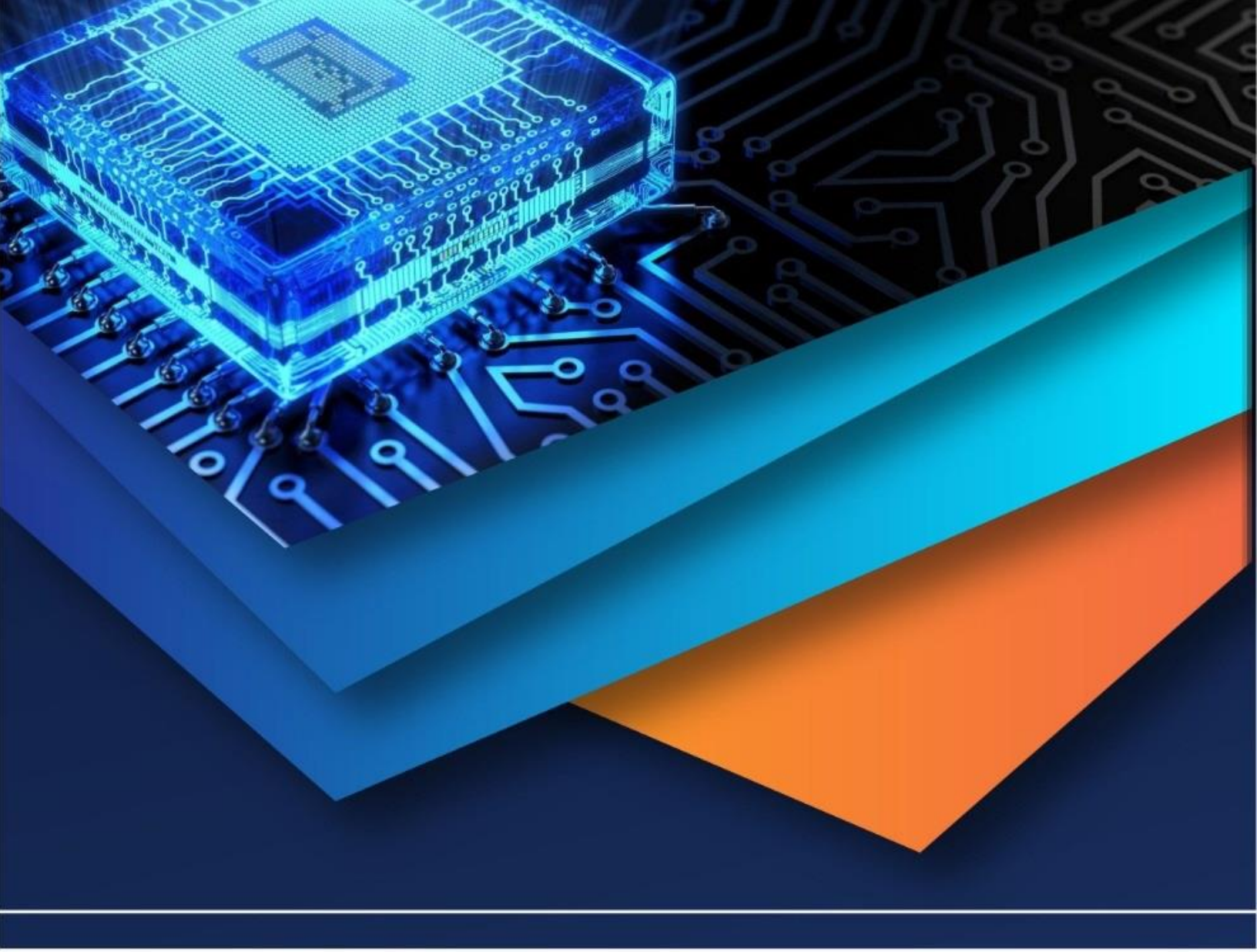
The results indicate strong performance across several metrics — including precision, recall, and F1-score — confirming that machine learning can be a reliable tool for malware detection. Moreover, the system's support for multi-format file scanning and batch processing makes it scalable and practical for real-world deployment.

In conclusion, the proposed system not only improves malware detection accuracy but also provides valuable insights through detailed reports and visual feedback. It stands as a proactive and adaptive solution in the evolving field of cybersecurity, capable of helping users and organizations better defend against complex malware threats.

REFERNCES

[1] Saxe, J., & Berlin, K. (2020). Deep neural network based malware detection using two dimensional binary program features. Proceedings of the 10th International Conference on Malicious and Unwanted Software (MALWARE), IEEE.
[2] Raff, E., Barker, J., Sylvester, J., Brandon, R., Catanzaro, B., & Nicholas, C. (2021). Malware detection by eating a whole EXE. Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence.
[3] Kolosnjaji, B., Zarras, A., Webster, G., & Eckert, C. (2022). Deep learning for classification of malware system call sequences. Australasian Joint Conference on Artificial Intelligence, Springer.
[4] Shijo, G., & Salim, A. (2015). Integrated static and dynamic analysis for malware detection. Procedia Computer Science, 46, 804–811.
[5] Ye, Y., Li, T., Adjeroh, D., & Iyengar, S. S. (2019). A survey on malware detection using data mining techniques. ACM Computing Surveys (CSUR), 50(3), 1–40.
[6] VirusShare.(2023).https://virusshare.com(Used as a malware sample repository for training and testing)
[7] Kaggle.(2023).MalwareDetectionDatasets.https://www.kaggle.com (Used for acquiring labeled malware and benign samples).
[8] Ucci, D., Aniello, L., & Baldoni, R. (2019). Survey of machine learning techniques for malware analysis. Computers & Security, 81, 123–147.
[9] Huang, W., Stokes, J. W. (2024). MtNet: A multi-task neural network for dynamic malware classification. International Joint Conference on Neural Networks (IJCNN), IEEE.
[10] Anderson, H. S., & Roth, P. (2024). EMBER: An open dataset for training static PE malware machine learning models. arXiv preprint arXiv:1804.04637.
[11] Kruegel, C., & Vigna, G. (2024). Anomaly detection of web-based attacks. Proceedings of the 10th ACM Conference on Computer and Communications Security (CCS).

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)