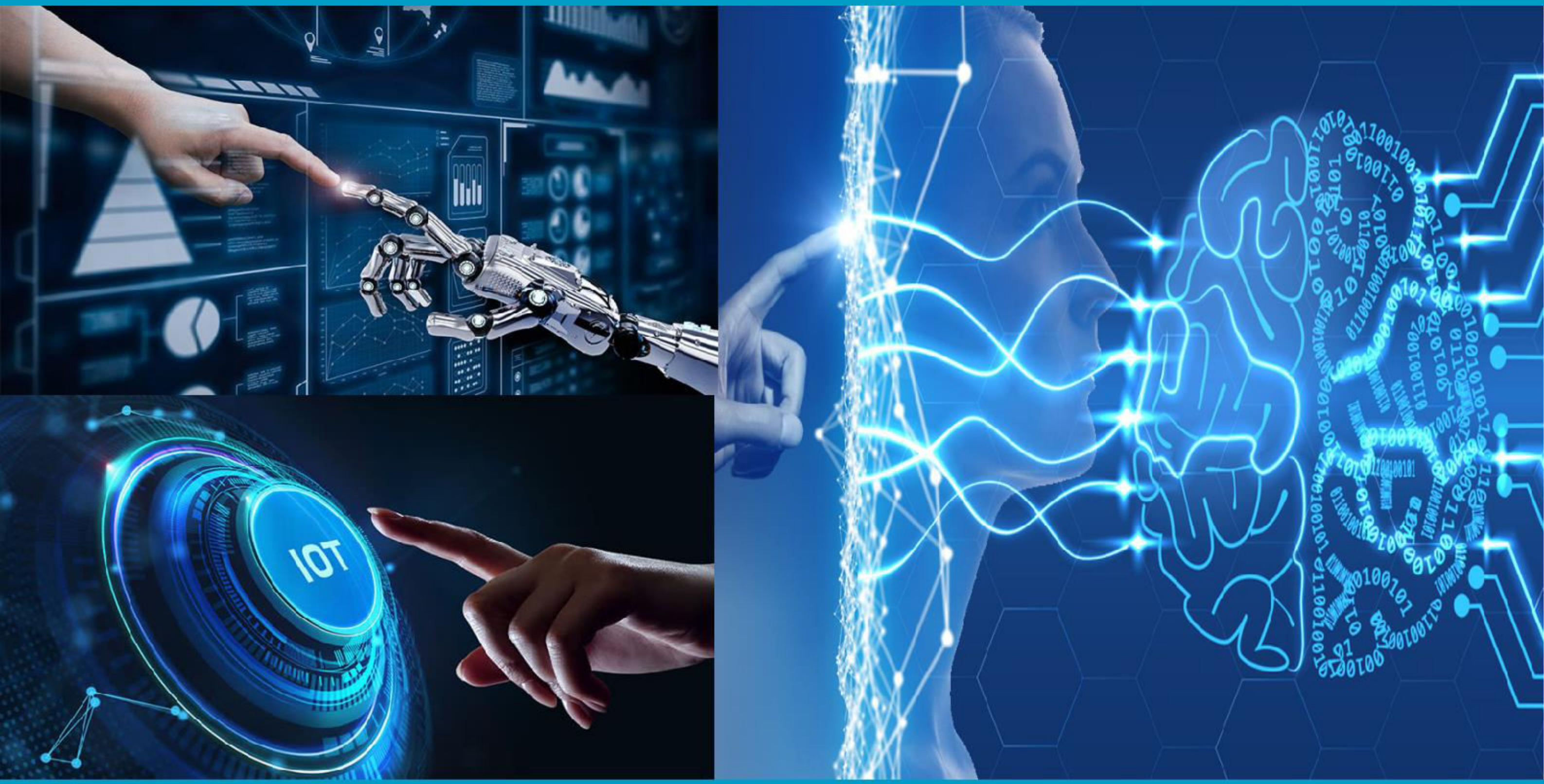
IJMRSETM

e-ISSN: 2395 - 7639



INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

Volume 12, Issue 5, May 2025

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.214

☎ +91 99405 72462 +9163819 07438 ✉ ijmrsetm@gmail.com @ www.ijmrsetm.com

International Journal of Multidisciplinary Research in Science, Engineering, Technology & Management (IJMRSETM)

| ISSN: 2395-7639 | www.ijmrsetm.com | Impact Factor: 8.214 | A Monthly Double-Blind Peer Reviewed Journal |

| Volume 12, Issue 5, May 2025 |

Malware Detection and Classification using Machine Learning

Prof. Suhail Shaha[1], Mr. Deekshith Gowda C R[2], Mr. Nithin G B[3], Mr.Tony Alex M[4], Mr. U Jayasimha[5]

Assistant Professor, Dept. of ISE, YIT Moodbidri, Mangalore, Karnataka, India[1]

B.E Student, Dept. of ISE, YIT Moodbidri, Mangalore, Karnataka, India [2]

B.E Student, Dept. of ISE, YIT Moodbidri, Mangalore, Karnataka, India[3]

B.E Student, Dept. of ISE, YIT Moodbidri, Mangalore, Karnataka, India[4]

B.E Student, Dept. of ISE, YIT Moodbidri, Mangalore, Karnataka, India [5]

ABSTRACT: With the rapid evolution of malware, traditional signature-based detection methods struggle to keep pace. This study presents a robust machine learning-based framework that leverages dynamic analysis via the Cuckoo Sandbox to monitor malware behavior in a controlled environment. A custom feature extraction and selection module identifies key behavioral indicators to enhance detection accuracy while reducing computational overhead. Multiple machine learning models, including Decision Trees, Random Forests, SVMs, and deep learning, are employed and integrated using ensemble soft voting to boost reliability and generalization. Experimental results on benchmark datasets demonstrate superior performance over traditional approaches, highlighting the framework's scalability and efficacy in modern cybersecurity applications.

KEYWORDS: Malware, Classification, Machine Learning, Deep Learning, Polymorphism, Ensemble Models

I. INTRODUCTION

Malware remains a persistent and evolving threat in today's interconnected digital ecosystem. Designed to infiltrate, damage, or exploit systems, malware comes in diverse forms such as viruses, trojans, adware, and spyware, each leveraging unique evasion techniques. Traditional detection approaches, while once effective, now struggle to keep pace with the increasing sophistication of modern attacks. To address this, advanced methods integrating dynamic behavioral analysis and machine learning have emerged as promising alternatives. This survey explores the application of Cuckoo Sandbox for dynamic feature extraction and examines the role of machine learning algorithms in enhancing detection and classification accuracy. The aim is to provide a comprehensive overview of current trends, challenges, and innovations shaping the future of malware defense strategies.

II. METHODOLOGY

This survey investigates a hybrid malware detection methodology combining static and dynamic analysis, network traffic inspection, and machine learning. Suspicious files are executed in Cuckoo Sandbox to capture real-time behavioral data, including API calls, file modifications, and network activities. Concurrently, static features such as file headers and control flow are extracted without execution. Network traffic analysis helps identify communication patterns indicative of command-and-control behavior. Relevant features are selected using Recursive Feature Elimination (RFE), and multiple classifiers—including Random Forests and Neural Networks—are trained. Ensemble learning with soft voting consolidates predictions, improving robustness and classification accuracy across diverse malware types.

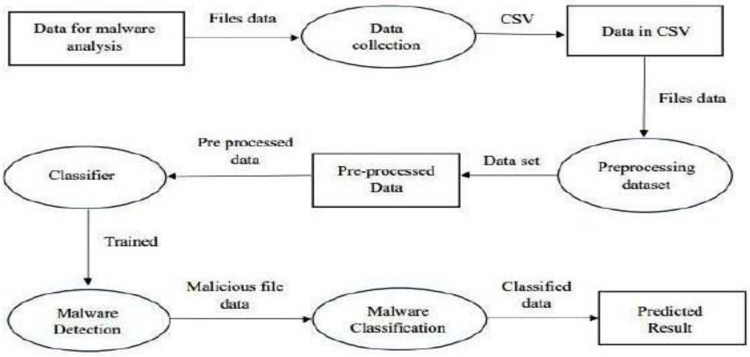


Fig.1 Architecture Model for Malware Analysis

III. LITERATURE REVIEW

Recent studies have focused on integrating static and dynamic analysis with machine learning to enhance malware detection accuracy. Techniques such as API call monitoring, feature extraction from sandbox reports, and hybrid analysis models have demonstrated effectiveness. Ensemble methods and deep learning architectures, including CNNs and stacked models, improve detection of obfuscated and fileless malware. Research also highlights the use of image-based representations and memory forensics to capture subtle malware behavior. Tools like Cuckoo Sandbox and frameworks like Streamlit have enabled scalable and interactive solutions. Overall, literature confirms that multi-feature and multi-layered approaches significantly outperform traditional signature-based detection systems.

IV. RESULT AND DISCUSSION

The proposed ensemble model was evaluated using metrics like accuracy, precision, recall, F1-score, and AUC- ROC, confirming its strong predictive performance. Among tested classifiers, Random Forest and Neural Networks achieved the highest accuracy and generalization on the 102-feature dataset. The integration of Recursive Feature Elimination improved model efficiency by eliminating irrelevant features. Ensemble learning via soft voting further enhanced robustness by minimizing the impact of individual classifier weaknesses. ROC analysis showed a strong trade-off between sensitivity and specificity, validating the model's applicability in real- world malware detection. Continuous monitoring and retraining are essential to maintain performance against evolving malware threats.

V. CONCLUSION

The comparative analysis confirms that machine learning techniques, particularly Random Forest and Gaussian Naive Bayes, are highly effective for binary malware detection. Ensemble and tree-based classifiers consistently outperformed others across both datasets, achieving perfect or near-perfect evaluation scores. Neural networks and KNN also demonstrated strong performance, though slightly hindered by model complexity. Support Vector Machine and Logistic Regression lagged behind, indicating limitations in handling intricate malware patterns. Multiclass classification results further highlight the promise of XGBoost and Neural Networks in identifying malware families. Overall, the study validates that well-tuned ensemble methods offer the most robust and scalable solutions for real-world malware detection.

REFERENCES

[1] V Esraa Odat,” Detecting Malware families and sub families using machine learning algorithm: Empirical Study, in International Journal Of Advance Computer Applications”,2022.

[2] [2] B A S. Dilhara,” Classification of Malware Using Machine Learning and Deep Learning Techniques”, in International Journal of Computer Apllications,2021.

[3] Gibert, D: Mateu: Planes, J. The rise of machine learning for detection and classification of malware: Research developments, trends and challenges. J.Netw. Comput. Appl. 2020,153,102526.

[4] Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Berlin, Heidelberg: Springer Berlin Heidelberg.

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING, TECHNOLOGY AND MANAGEMENT

+91 99405 72462 +91 63819 07438 ijmrsetm@gmail.com

www.ijmrsetm.com