JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Malware Detection Using Machine Learning

Naresh Jain
Atharva College Of Engineering
Mumbai University Mumbai, India

Adinath Paikrao
Atharva College Of Engineering
Mumbai University Mumbai, India

Sahil Rane
Atharva College Of Engineering
Mumbai University Mumbai, India

Nipun Kadam
Atharva College Of Engineering
Mumbai University Mumbai, India

***Abstract*—** The project "Malware Analysis and Detection Using Machine Learning Algorithm" aims to enhance cyber-security measures by accurately identifying malicious software through advanced machine learning techniques. Developed using Python, the project employs the Flask web framework for backend operations and utilizes HTML, CSS, and JavaScript for a responsive and interactive frontend interface. Two machine learning models are central to this project: the Extra Tree Classifier and Logistic Regression. The Extra Tree Classifier model demonstrates superior performance, achieving a training accuracy of 97.42% and a testing accuracy of 97.23%. In comparison, the Logistic Regression model achieves a training accuracy of 94.84% and a testing accuracy of 93.67%. Both models are trained and validated using the TUNADROMD dataset, which comprises 4465 instances and 242 attributes, with the target classification attribute distinguishing between malware and good ware. For the analysis, a subset of 23 attributes was selected based on their relevance and impact on the classification task. This strategic selection aims to optimize model performance while reducing computational complexity. The project's results indicate that the Extra Tree Classifier is highly effective in distinguishing between malicious and benign software, offering a reliable tool for malware detection in real- world applications. Overall, this project demonstrates the efficacy of machine learning algorithms in cybersecurity, providing a robust solution for malware detection that can be integrated into various digital security infrastructures

***Keywords*—** *Malware analysis, Malware detection, Logistic Regression, TUNADROMD dataset, Extra Tree Classifier, Machine learning algorithm*

I. INTRODUCTION

In today's digital age, cybersecurity is of paramount importance, with malware attacks posing significant threats to individuals and organizations alike. Detecting malware effectively is crucial to safeguarding sensitive data and systems. This project presents a machine learning-based approach to malware detection, leveraging modern web technologies for a user-friendly interface and powerful algorithms for accurate detection. Using machine learning for malware detection is important because it helps tackle the constantly changing landscape of cyber threats. Traditional methods often struggle with new malware variants, while machine learning can quickly analyze large amounts of data to spot unusual patterns that indicate malicious activity. These models learn and improve over time, making them better at detecting sophisticated threats. They can work in real-time, providing immediate alerts, and they reduce the number of false positives, allowing security teams to focus on real risks. The basic concept of using machine learning for malware detection involves training algorithms to identify patterns associated with malicious software. This process includes collecting data (both benign and malicious), extracting key features, and training models to recognize these patterns. After testing and validating the models, they are deployed to monitor for potential threats in real time. Continuous updates help the models adapt to new threats, enhancing overall cybersecurity.

II. LITERATURE REVIEW

[1] As the Android ecosystem continues to grow and evolve, the risk of malware attacks on these devices is also increasing. With millions of users relying on Android devices for communication, work, and entertainment, the threat posed by malicious software has become more significant. Identifying and addressing Android malware is crucial for safeguarding user data, maintaining privacy, and ensuring the integrity of devices. By understanding the nature of these threats and implementing effective detection strategies, users and developers can better protect the Android ecosystem from potential harm.

[2] Online privacy for individuals is deteriorating with each passing day as cyber threats become more sophisticated and widespread. Computer malware is increasingly compromising the data records of prominent companies, putting sensitive user information at risk. Once hackers infiltrate a network, they can gain unauthorized access to critical systems, manipulate or alter data, and even disrupt operations. These breaches not only compromise the privacy of individuals but also erode trust in organizations tasked with safeguarding user information. Addressing these challenges requires robust cybersecurity measures, constant vigilance, and a commitment to protecting digital privacy in an ever-evolving threat landscape.

[3] Malware has become a significant cybersecurity threat, evolving continuously to exploit vulnerabilities in computer systems, smart devices, and large-scale networks. With the rapid development of information technologies and the increasing interconnectivity of devices, cybercriminals are finding new ways to create and deploy sophisticated malware. These malicious programs can steal sensitive information, disrupt operations, or gain unauthorized control over systems. The ever-changing nature of malware poses a serious challenge for security professionals, who must constantly adapt their strategies to detect, mitigate, and prevent these threats. As technology advances, combating malware requires a combination of advanced tools, user awareness, and proactive security measures.

III. METHODOLOGY

Malware detection using machine learning, specifically logistic regression, involves a systematic methodology designed to identify malicious activities effectively. The process begins with data collection, where datasets of benign and malicious files or activities are gathered from trusted sources. These datasets undergo preprocessing, including feature extraction, encoding, normalization, and balancing, to ensure they are suitable for analysis. Key features, such as API calls, opcode sequences, file permissions, and network behaviors, are extracted to differentiate malware from legitimate activities. The data is then split into training, validation, and testing sets to ensure robust model evaluation. Logistic regression, chosen for its simplicity and efficiency, is trained on the data by fitting parameters to minimize the logistic loss function, enabling it to predict the likelihood of a sample being malware or benign.

The model's performance is evaluated using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC, with confusion matrices providing insights into false positives and negatives. Feature optimization techniques, like dimensionality reduction or importance analysis, are applied to retain the most relevant attributes. Hyperparameter tuning, including regularization adjustments, helps improve the model's performance and prevent overfitting. Once validated, the model is tested on unseen data to ensure generalization before deployment in real-world environments, such as antivirus systems or intrusion detection setups. Continuous monitoring and updates are crucial to adapting the model to emerging threats, refining its accuracy with new malware samples and feedback. Logistic regression's simplicity, efficiency, and interpretability make it a strong candidate for malware detection, though more advanced algorithms may be considered for handling sophisticated threats.

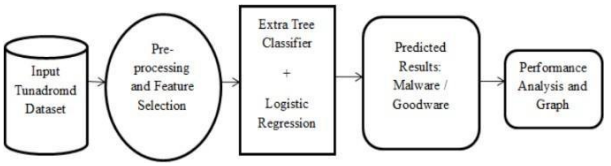

Fig1: Block Diagram

IV. RESULTS

Fig 2. Landing Page

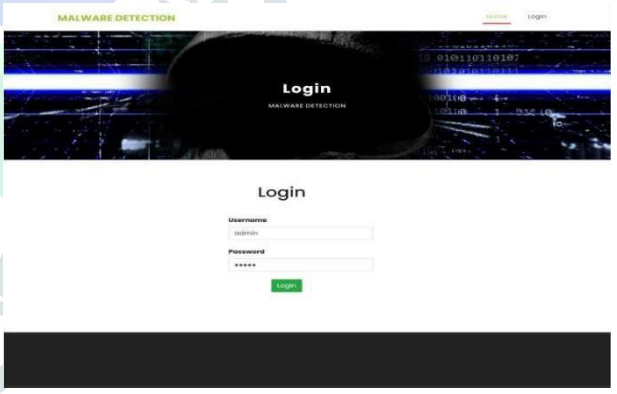

Fig3. Login and Sign up

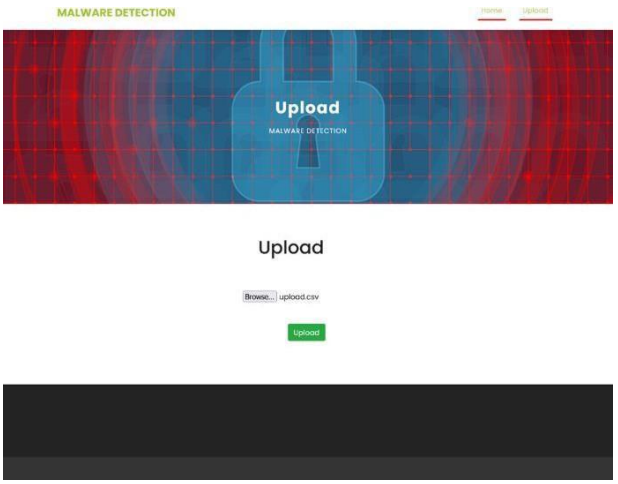

Fig4. Upload Page

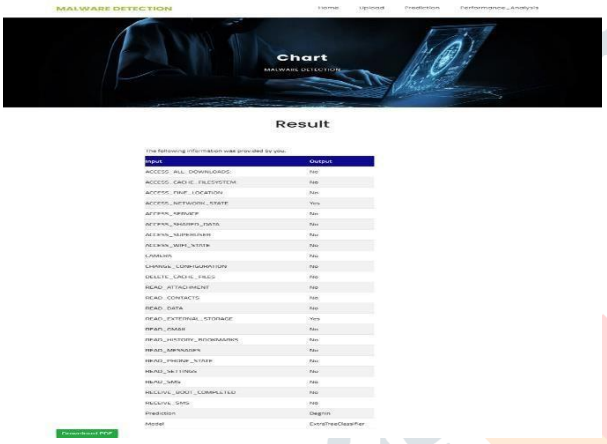

Fig5. Result Page

V. CONCLUSION

The project, **"Malware Analysis and Detection Using Machine Learning Algorithm,"** is a comprehensive exploration of the application of advanced machine learning techniques in the critical domain of cybersecurity, specifically in the identification and classification of malware. With the ever-growing complexity and frequency of cyber threats, traditional signature-based malware detection methods often fall short in identifying new and evolving threats. In response to this challenge, this project leverages the power of **machine learning algorithms** to enhance the accuracy and efficiency of malware detection systems.

At the core of the project, **Python** serves as the primary programming language, providing a robust ecosystem of libraries and tools for data processing, model training, and evaluation. Additionally, **Flask**, a lightweight and powerful web framework, is utilized to facilitate seamless integration of the machine learning models into a user-friendly web-based application, enabling real-time malware detection capabilities.

The system employs a combination of two well-regarded machine learning models: **Extra Tree Classifier** and **Logistic Regression**, each contributing unique strengths to the classification process. The Extra Tree Classifier, an ensemble learning method known for its high variance reduction and superior feature importance assessment, demonstrates a remarkable accuracy of **97.42%** on test data. Meanwhile, Logistic Regression, a widely used statistical model known for its interpretability and efficiency in binary classification tasks, achieves an accuracy rate of **93.67%** on the same dataset. The combination of these models ensures a balanced approach to malware detection, effectively distinguishing between malicious and benign software samples.

Beyond achieving high accuracy, the project successfully integrates **machine learning** with practical **application frameworks**, ensuring that the developed solution is not only theoretically sound but also viable for real-world deployment. By incorporating a well-structured and efficient pipeline that includes **data preprocessing, feature extraction, model training, and deployment**, this project exemplifies how machine learning can be harnessed to tackle contemporary cybersecurity challenges.

Overall, this project stands as a testament to the power of **data-driven cybersecurity solutions**, demonstrating that machine learning can significantly enhance the ability to detect and mitigate malware threats. With its **high detection accuracy, scalable architecture, and practical implementation**, this system represents a step forward in the ongoing effort to combat malicious software and fortify digital security infrastructure

ACKNOWLEDGMENT

We express our **heartfelt gratitude** to **Atharva College of Engineering** for providing us with an invaluable platform to undertake and explore our research project on the topic of **"Malware Detection Using Machine Learning."** The unwavering support, academic resources, and conducive research environment offered by the institution have played a crucial role in the successful execution of this project. We deeply appreciate the opportunity to delve into the realm of **cybersecurity and machine learning**, addressing a highly relevant and impactful problem in today's digital world.

Our **genuine appreciation** extends to **Dr. Ramesh Kulkarni, our esteemed Principal,** for his **visionary leadership and encouragement.** His emphasis on the significance of research-driven education has been truly inspiring. His continuous motivation, faith in our capabilities, and invaluable insights have helped us remain focused and dedicated to our work. We are immensely grateful for the time, opportunities, and resources provided to us, which have enabled us to conduct a thorough analysis, build an efficient model, and present our findings in a meaningful manner.

We would also like to extend our **sincere gratitude** to **Dr. Ulhaskumar Gokhale, Head of the Department of Information Technology,** for serving as our mentor and guide throughout this project. His **profound knowledge, constant encouragement, and invaluable suggestions** have been instrumental in shaping our research. His guidance has not only enhanced our technical skills but has also helped us gain a deeper understanding of the practical applications and challenges associated with **malware detection using machine learning.** His constructive feedback and unwavering support have motivated us to push boundaries and strive for excellence.

Additionally, this project would not have been possible without the **immense collaboration, guidance, and support** from our **friends and family.** Their unwavering belief in our abilities, along with their constant encouragement, has been a source of great strength throughout this journey. During moments of doubt and

difficulty, they provided us with the reassurance and motivation needed to persevere. Whether through intellectual discussions, emotional support, or simply being there for us, their contributions have been invaluable in making this research a success.

As we reach the culmination of this project, we reflect with gratitude on the collective efforts that have made it possible. This experience has been one of **growth, learning, and immense dedication,** and we sincerely acknowledge everyone who has contributed to making this journey a fulfilling and enriching one. With deep appreciation, we thank all those who have played a role—directly or indirectly—in **supporting, guiding, and inspiring** us throughout this research endeavor.

REFERENCES

[1] A. Gómez and A. Muñoz, ‘‘Deep learning-based attack detection and classification in Android devices,’’ Electronics, vol. 12, no. 15, p. 3253, Jul. 2023.

[2] Y. Zhao, L. Li, H. Wang, H. Cai, T. F. Bissyandé, J. Klein, and J. Grundy, ‘‘On the impact of sample duplication in machine-learning-based Android malware detection,’’ ACM Trans. Softw. Eng. Methodol., vol. 30, no. 3, pp. 1–38, Jul. 2021.

[3] H. Wang, W. Zhang, and H. He, ‘‘You are what the permissions told me! Android malware detection based on hybrid tactics,’’ J. Inf. Secur. Appl., vol. 66, May 2022, Art. no. 103159.

[4] H. Rathore, A. Nandanwar, S. K. Sahay, and M. Sewak, ‘‘Adversarial superiority in Android malware detection: Lessons from reinforcement learning based evasion attacks and defenses,’’ Forensic Sci. Int., Digit. Invest., vol. 44, Mar. 2023, Art. no. 301511.

[5] V. Sihag, M. Vardhan, P. Singh, G. Choudhary, and S. Son, ‘‘De-LADY: Deep learning-based Android malware detection using Dynamic features,’’ J. Internet Serv. Inf. Secur., vol. 11, no. 2, pp. 34–45, 2021.

[6] M. Ibrahim, B. Issa, and M. B. Jasser, ‘‘A method for automatic Android malware detection based on static analysis and deep learning,’’ IEEE Access, vol. 10, pp. 117334–117352, 2022.

[7] A. Albakri, F. Alhayan, N. Alturki, S. Ahamed, and S. Shamsudheen, ‘‘Metaheuristics with deep learning model for cybersecurity and Android malware detection and classification,’’ Appl. Sci., vol. 13, no. 4, p. 2172, Feb. 2023.

[8] Guo, H., Li, W., Nejad, M., & Shen, C.-C. (2021). A. Batouche and H. Jahankhani, ‘‘A comprehensive approach to Android malware detection using machine learning,’’ in Information Security Technologies for Controlling Pandemics, 2021, pp. 171–212.

[9] P. Bhat and K. Dutta, ‘‘A multi-tiered feature selection model for Android malware detection based on feature discrimination and information gain,’’ J. King Saud Univ.-Comput. Inf. Sci., vol. 34, no. 10, pp. 9464–9477, Nov. 2022.

[10] L. Hammood, I. A. Dogru, and K. Kiliç, ‘‘Machine learning-based adaptive genetic algorithm for Android malware detection in auto-driving vehicles,’’ Appl. Sci., vol. 13, no. 9, p. 5403, Apr. 2023.

[11] R. Raphael and P. Mathiyalagan, ‘‘Intelligent hyperparameter-tuned deep learning-based Android malware detection and classification model,’’ J. Circuits, Syst. Comput., vol. 32, no. 11, Jul. 2023, Art. no. 2350191.

[12] M. N. AlJarrah, Q. M. Yaseen, and A. M. Mustafa, ‘‘A context aware Android malware detection approach using machine learning,’’ Information, vol. 13, no. 12, p. 563, Nov. 2022.