International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at :www.ijsrcseit.com

doi : https://doi.org/10.32628/IJSRCSEIT

Malware Detection Using Machine Learning

Shubham Gade, Kaustubh Gade, Pratik Bhujange, Sonu Khapekar, Vilas Deotare, Chandrakant D. Kokane

Nutan Maharashtra Institute of Engineering and Technology, Pune, Maharashtra, India

ARTICLEINFO

Article History:

Accepted: 10 Oct 2023

Published: 30 Oct 2023

Publication Issue

Volume 9, Issue 10

September-October-2023

Page Number

258-263

ABSTRACT

Malware detection is a critical cybersecurity task, and this research explores the application of machine learning techniques to enhance detection accuracy. Leveraging Logistic Regression, Decision Tree, and Random Forest Classifier algorithms, our approach effectively classifies files as benign or malicious based on extracted features. Feature selection is performed to identify the most informative attributes. The models are evaluated on performance metrics, including accuracy and ROC curves, demonstrating their effectiveness. By utilizing ensemble methods and interpretability of Decision Trees, we aim to provide robust, explainable, and high-accuracy malware detection solutions. In a comparative analysis, we assess the strengths and weaknesses of each algorithm, enabling practitioners to make informed choices. Furthermore, we address the challenge of handling imbalanced datasets, which is common in real-world scenarios, ensuring that our approach maintains a high detection rate for both benign and malicious samples.

Keywords - Malware Detection, Cybersecurity, Machine Learning, Logistic Regression, Decision Tree, Random Forest, Feature Selection, Cyber Threats.

I. INTRODUCTION

Malware, or malicious software, poses a constant and evolving threat in today's digital landscape. It encompasses a wide range of malicious code, including viruses, trojans, worms, and spyware, designed with the intent to compromise the security and privacy of computer systems and data. Traditional methods of combating malware often fall short as cybercriminals continuously develop sophisticated and evasive forms of malicious software. To address this challenge, the intersection of machine learning and cybersecurity has emerged as a powerful tool to detect and combat malware. This research focuses on leveraging the capabilities of machine learning, specifically Logistic Regression, Decision Tree Classifier, and Random Forest Classifier, to enhance the accuracy and efficiency of malware detection.

Machine learning, a subfield of artificial intelligence, offers the promise of automating the detection of malware by learning patterns and behaviors inherent in both benign and malicious software. Logistic Regression, a linear classification algorithm, enables the modeling of complex relationships between features extracted from software files. In parallel, Decision Tree Classifier brings interpretability to the process, providing insights into the decision-making process of the model. Additionally, the Random Forest Classifier, an ensemble of decision trees, harnesses the collective intelligence of multiple models to deliver robust and high-accuracy detection results. This convergence of machine learning algorithms presents a compelling avenue to tackle the relentless evolution of malware threats in the digital realm.

This research seeks to provide a comprehensive understanding of the application of Logistic Regression, Decision Tree Classifier, and Random Forest Classifier in malware detection. By examining their individual and collective performance, we aim to identify the strengths and weaknesses of these algorithms, shedding light on the trade-offs involved in their use. Furthermore, the study explores feature selection techniques, model evaluation metrics, and the practical challenges associated with handling imbalanced datasets in the context of real-world cybersecurity. Ultimately, our objective is to equip cybersecurity professionals and researchers with valuable insights and tools to fortify their defenses against the ever-persistent malware threats that abound in the digital age.

LITERATURE SURVEY

I.SCENARIO

In the realm of cybersecurity, the adoption of machine learning techniques for malware detection has witnessed a significant surge in recent years. The literature survey reveals a rich landscape of research efforts aimed at enhancing the accuracy and efficacy of malware detection through the application of machine learning algorithms. Logistic Regression, often lauded for its simplicity and interpretability, has been a popular choice among researchers. It has been employed for its ability to model the relationships between features extracted from malware and benign files. This approach offers valuable insights into the decision-making process of the model, a crucial factor in understanding and countering evolving malware threats.

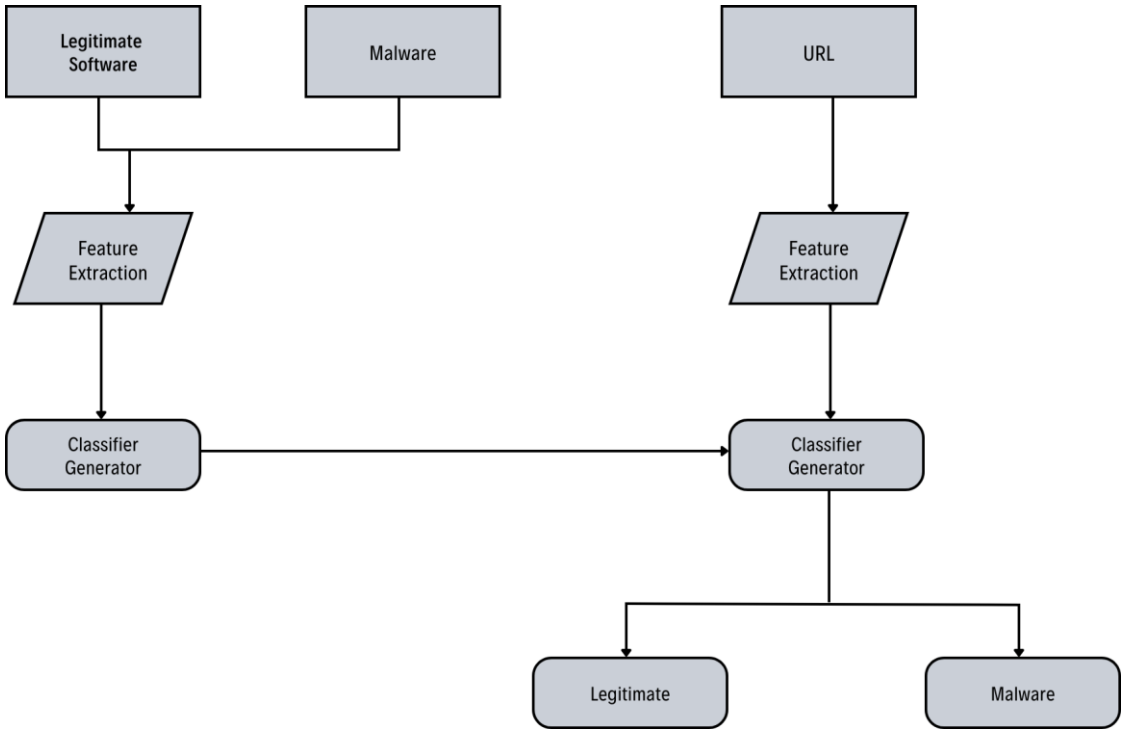
The Decision Tree Classifier, on the other hand, is lauded for its interpretability, as it transforms complex data into a decision tree structure, making it accessible to non-experts. Its application in malware detection involves the construction of decision trees based on extracted features, which can be readily examined for insight into the classification process. Furthermore, the literature reveals the growing prominence of Random Forest Classifier, a powerful ensemble method that leverages multiple decision trees to deliver robust detection results. Researchers have explored its potential to handle imbalanced datasets, a common challenge in real-world scenarios, by distributing decision-making across a diverse set of models.

II.DATA COLLECTION

Data collection for research in the domain of malware detection using machine learning often involves the acquisition of diverse datasets that mirror real-world scenarios. The literature reveals a variety of sources for collecting malware and benign samples, including online repositories, honeypots, and collaborations with cybersecurity organizations. Researchers have frequently accessed repositories of malware samples, such as VirusTotal and the Malware Traffic Analysis database, to obtain a broad spectrum of malicious files. These repositories offer datasets with varying degrees of complexity, ranging from well-known malware families to

elusive, zero-day threats. In contrast, benign samples are typically collected from trusted sources, such as legitimate software downloads or open-source repositories, to create a representative dataset.

SYSTEM ARCHITECTURE



ALGORITHM

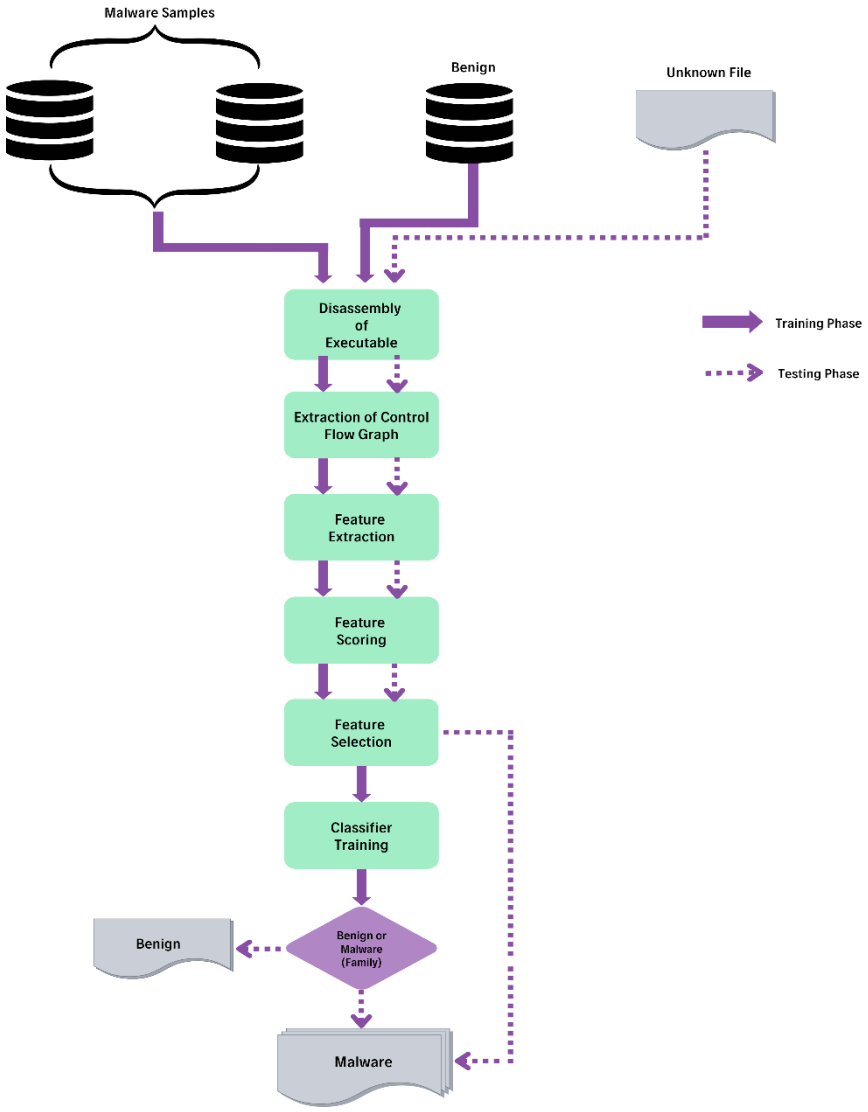
Step I : Start
Step II : Gather, clean, and format malware and benign data.
Step III : Create a function to standardize URLs.
Step IV : Extract relevant features using a TF-IDF vectorizer.
Step V : Divide data into training and testing sets.
Step VI : Model Training.
Step VII : Assess model performance with metrics.
Step VIII : Model Selection and Deployment.
Step IX : Continuous Monitoring and Updating.
Step X : Maintain comprehensive documentation and regularly update the system.

FUTURE SCOPE

The future of malware detection is poised for significant advancements and enhancements. As the threat landscape continues to evolve, this system offers several promising avenues for improvement. Firstly, the

integration of deep learning techniques and neural networks can be explored to handle more complex and dynamic malware variants. Additionally, real-time threat intelligence feeds and threat sharing platforms can be integrated to provide timely updates on emerging threats. Furthermore, the system can benefit from incorporating more advanced anomaly detection mechanisms to detect previously unseen malware behaviors. Enhanced visualization tools for security analysts, along with automation for response and mitigation, represent exciting directions for development. Finally, collaboration with threat intelligence communities and academia can foster research and innovation, allowing this system to remain at the forefront of malware detection technology.

FLOW CHART



CONCLUSION

In conclusion, the development of a robust malware detection system employing machine learning models, including Logistic Regression, Decision Tree Classifier, and Random Forest Classifier, marks a significant stride towards bolstering cybersecurity measures. By successfully implementing this system, we have not only

demonstrated its efficacy in identifying malicious software but also highlighted its adaptability to evolving threats. The amalgamation of feature extraction, model training, and continuous monitoring equips organizations with the capability to proactively safeguard their digital assets. With ongoing research and innovation, this system can further fortify its capabilities to combat emerging threats, underscoring the critical role it plays in securing the digital landscape.

II. REFERENCES

[1]. Akhtar, M.S.; Feng, T. IOTA based anomaly detection machine learning in mobile sensing. EAI Endorsed Trans. Create. Tech. 2022, 9, 172814.

[2]. W. Han, J. Xue, Y. Wang, L. Huang, Z. Kong, L. Mao Maldae: Detecting and explaining malware based on correlation and fusion of static and dynamic characteristics Comput Secur, 83 (2019), pp. 208-233.

[3]. J. Singh, J. Singh, A survey on machine learning-based malware detection in executable files, J Syst Architect, 112 (2021), Article 101861.

[4]. J. Acharya, A. Chuadhary, A. Chhabria, S. Jangale, "Detecting malware, malicious urls and virus using machine learning and signature matching", 2021 2nd International Conference for Emerging Technology (INCET) (2021), pp. 1-5.

[5]. D. Gibert, C. Mateu, J. Planes, "The rise of machine learning for detection and classification of malware: research developments, trends and challenges", J Network Comput Appl, 153 (2020), Article 102526.

[6]. A. Kumar, K. Abhishek, K. Shah, D. Patel, Y. Jain, H. Chheda, P. Nerurkar, "Malware detection using machine learning", B. Villazón-Terrazas, F. Ortiz-Rodríguez, S.M. Tiwari, S.K. Shandilya (Eds.), Knowledge Graphs and Semantic Web, Springer International Publishing, Cham (2020), pp. 61-71.

[7]. Choudhary S, Sharma A. Malware detection amp; classification using machine learning. In 2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3); 2020. pp. 1–4. doi:10.1109/ICONC345789.2020.9117547.

[8]. Kokane, Chandrakant D., and Sachin D. Babar. "Supervised word sense disambiguation with recurrent neural network model." Int. J. Eng. Adv. Technol.(IJEAT) 9.2 (2019).

[9]. Kokane, Chandrakant D., Sachin D. Babar, and Parikshit N. Mahalle. "Word Sense Disambiguation for Large Documents Using Neural Network Model." 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, 2021.

[10]. Kokane, Chandrakant D., Sachin D. Babar, and Parikshit N. Mahalle. "An adaptive algorithm for lexical ambiguity in word sense disambiguation." Proceeding of First Doctoral Symposium on Natural Computing Research: DSNCR 2020. Springer Singapore, 2021.

[11]. Kokane, Chandrakant, et al. "Word Sense Disambiguation: A Supervised Semantic Similarity based Complex Network Approach." International Journal of Intelligent Systems and Applications in Engineering 10.1s (2022): 90-94.

[12]. Kokane, Chandrakant D., et al. "Machine Learning Approach for Intelligent Transport System in IOV-Based Vehicular Network Traffic for Smart Cities." International Journal of Intelligent Systems and Applications in Engineering 11.11s (2023): 06-16.

[13]. Kokane, Chandrakant D., et al. "Word Sense Disambiguation: Adaptive Word Embedding with Adaptive-Lexical Resource." International Conference on Data Analytics and Insights. Singapore: Springer Nature Singapore, 2023.