IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MALICIOUS URL AND PE-HEADER-BASED MALWARE DETECTION

[1]Sudhanva S Bhatt, [2]Sanchitha S Hebbar, [3]Samrudh J, [4] Gayathree T V

[1]Student, [2]Student, [3]Student, [4]Student
Department of Information Science and Engineering,
RNS Institute of Technology , Bangalore ,India

Abstract: This survey study examines malware detection methods and focuses on a particular project that employs machine learning to find potentially hazardous URLs and identify infected files. Malware is any software that carries out harmful operations, such data theft or espionage. In Kaspersky Labs' definition from 2017, malware is "a type of computer programme designed to infect a legitimate user's computer and cause harm in various ways."

The several polymorphic layers present in contemporary malware apps make it increasingly challenging to identify malware using conventional signature-based techniques. Since these layers either conceal from view or update themselves automatically on a regular basis, antivirus software finds it difficult to identify them.

Machine learning provides the answer by teaching models to recognise both positive and negative file attributes. This feature allows malware to be detected independent of previous experience by enabling the recognition of hazardous patterns.

The study that is being presented makes use of a Random Forest classifier for PE-header-based malware detection and Logistic Regression to discover potentially harmful URLs.

contains**.**

***Index Terms* - Component, formatting, style, styling, insert.**

I. Introduction

1.1 Background:

Cybersecurity faces a significant threat from malware, making the development of robust detection mechanisms imperative for protecting computer systems. Among various approaches, machine learning stands out as a promising method to enhance the accuracy and adaptability of malware detection systems.

1.2 Objectives:

Signature-based malware detection techniques are excellent at locating previously identified malware that antivirus vendors have cataloged. Their inability to identify polymorphic malware—malware that may change its signatures—remains a drawback. These systems also have trouble with newly discovered malware instances that don't have established signatures. Heuristics-based detectors thus frequently have insufficient accuracy, which leads to a significant amount of inaccurate information (Baskaran and Ralescu, 2016). The explosive growth of polymorphic viruses creates a critical need for novel detection techniques.

2. Related Work:

Despite their limited application, machine learning algorithms for malware detection are not new. Numerous research projects were undertaken in this area with the goal of determining the accuracy of various techniques.

In order to create a detection system, Dragos Gavrilut changed many perceptron algorithms as described in his study "Malware Detection Using Machine Learning." His range of accuracy for different methods was from 69.90% to 96.18%. It should be mentioned that the algorithms with the highest accuracy also produced the highest number of false positives; the algorithm with the highest accuracy produced 48 false positives. The best balanced strategy obtained 93.01% accuracy with a low false-positive rate and sufficient accuracy. Gavrilut et al. (2009).

The publication "A Static Malware Detection System Using Data Mining Methods" provided extraction techniques based on J48 Decision Trees, Naive Bayes, Support Vector Machines, and PE headers, DLLs, and API calls. With hybrid PE header & API function feature type and PE header feature type of 99% and 99.1%, respectively, the J48 algorithm yielded the best overall accuracy. (Jambaljav, Baldangombo, and Horng 2013)

3.1 PE-Header-Based Malware Detection:

3.1.1 Feature Extraction:

Detailed Feature Analysis: To evaluate the importance of the PE header features in detecting malicious patterns, a thorough analysis will be conducted. Every characteristic will be examined closely to determine how it affects the overall detection accuracy, shedding light on the value of specific PE header attributes.

3.1.2 Preprocessing Steps:

Data Cleaning: Strict preparation procedures will be carried out to guarantee the accuracy and dependability of the dataset before analysis. By eliminating outliers, inconsistent data, and missing values in the PE-header information, this will enhance the dataset's overall robustness.

3.1.3 Model Tuning:

Hyperparameter Tuning: To maximize its hyperparameters, the Random Forest classifier will go through a rigorous tuning procedure. To do this, the parameters associated with the model will need to be adjusted, with special emphasis paid to those that significantly affect the algorithm's ability to detect malware.

3.1.4 Evaluation Metrics:

Performance metrics: In addition to accuracy, a number of metrics, such as precision, recall, F1-score, and area under the ROC curve, will be used to assess the model's overall effectiveness. These metrics will be chosen to provide a more comprehensive understanding of the model's capabilities by taking into account how well they apply to malware detection.

3.1.5 Cross-Validation:

Cross-validation method: To confirm the durability of the model, a k-fold cross-validation method will be employed. This approach will ensure that the performance metrics remain consistent throughout multiple folds, providing a reliable assessment of the model's capacity for generalization.

3.1.6 Comparison with Decision Tree:

Comparison Metrics: The Random Forest and Decision Tree classifiers will be thoroughly compared in terms of accuracy, precision, recall, and F1-score, among other metrics.

3.2 Malicious URL Detection:

3.2.1 Algorithm Choice:

Using its prowess in binary classification tasks, logistic regression will be used to train the model. This decision's justification is that Logistic Regression is a feasible alternative for identifying harmful URLs due to its interpretability and efficiency in processing large datasets.

3.2.2 Data Cleaning:

Pandas will be utilized to manipulate data effectively the raw URL data will be preprocessed using a proprietary vectorizer before being cleaned up. Managing any discrepancies, removing superfluous data, and getting the URL strings ready for model training are the goals of this step.

3.2.3 Sanitization:

To efficiently filter URLs, a specific sanitization function will be put into place. This function will ensure a standardized and sanitized dataset for feature extraction by addressing concerns like special characters, encoding variances, and other potential abnormalities in the URL strings.

3.2.4 Feature Extraction:

The Term Frequency-Inverse Document Frequency (Tf-idf) approach intended for the extraction of text features. This method assesses the significance of every term in a URL throughout the dataset, enabling the model to concentrate on pertinent characteristics for the identification of malicious URLs. By capturing both local and global term importance, the Tf-idf strategy improves the discriminative capability of the prototype.

3.2.5 Whitelist Filter:

A whitelist filter will be applied to URL filtering in order to enhance the model's functionality. To allow the model to distinguish between known safe URLs and possibly harmful ones, this filter will contain a predetermined list of benign URLs. The whitelist filter adds another level of protection to the detection process, increasing accuracy and lowering false positives.

Fig 1 : Random Forest Classifier Performance Graph.

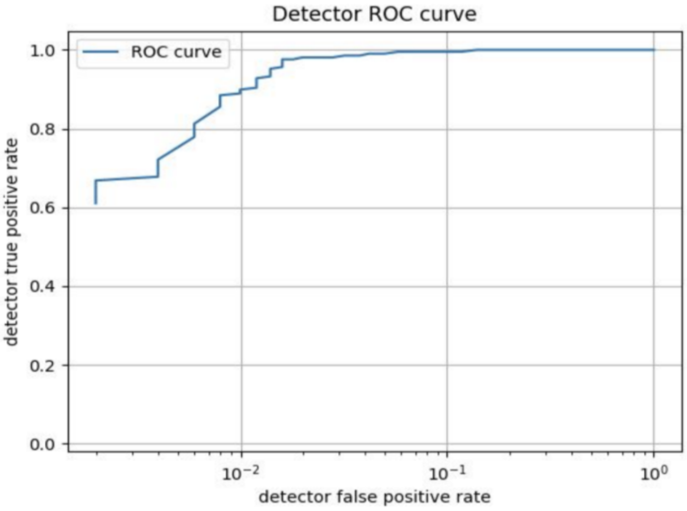
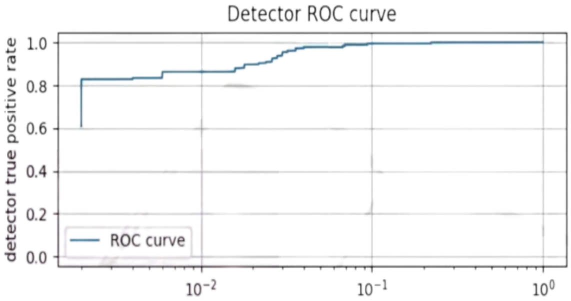


Fig 2 : Logistic Regression Classifier Performance Graph.



4. Advantages and Disadvantages

4.1 Advantages:

Enhanced Accuracy:

- Machine learning algorithms may be trained on large datasets, which helps them become more accurate over time as they encounter more examples of both malware and normal behavior.
- Real-time detection: This capability allows many models for machine learning to respond quickly to emerging dangers.
- Adaptability: models for machine learning can adapt to new and evolving malware types without the need for human updates. Being adaptable is crucial in the dynamic realm of cyberattacks..

Behavioral Analysis:

- Because ML models are able to examine the behavior of files and programs, it is now possible to identify zero-day or unknown threats by observing their behavior, as opposed to solely depending on known signs.

Reduced False Positives:

- By considering a range of qualities and behaviors, models for machine learning can reduce false positives in comparison to traditional signature-based detection strategies.

4.2 Disadvantages:

Data Quality:

- The quality and representativeness of the training data have a significant impact on how well models for machine learning perform. A biased or inadequate set of data could make the model perform poorly.

Overfitting:

- Machine learning systems perform well on the training set when they are overfit, but they cannot generalise to newly observed data. This may lead to false positives and negatives.

Evasion Techniques:

- In order to evade machine learning detection, malware creators may employ various evasion techniques or design their creations to resemble trustworthy behavior.

Resource Intensive:

- Due to their high computational cost, advanced models for machine learning require strong hardware to be deployed, especially in real-time applications.

Lack of Explainability:

- Many models for machine learning—especially complex ones like deep neural networks—are thought of as "black boxes," which means that it could be challenging to figure out how a specific finding was made or why a particular detection was made.

5. Conclusion

One significant development in cybersecurity that will assist fortify digital defenses is the use of machine learning for malware identification. The work demonstrates how the integration of flexibility, immediate response, and advanced behavioral analysis with machine learning may revolutionized hazard detection.

The power of this method lies in its capacity to adjust and gain knowledge from the ever-changing landscape of cyber threats. Machine learning's ability to self-adapt to new threats and its effectiveness in identifying and removing malware make it a valuable tool in the fight against cyber adversaries.

The project does, however, also highlight the difficulties and complications that are there. The necessity for constant monitoring and improvement is highlighted by problems including the requirement for high-quality training data, the possibility of overfitting, and the never-ending game of cat and mouse with changing escape strategies.

In order for machine learning to be effective in malware detection, it must be able to strike a delicate balance between innovation and fixing innate limitations as the digital ecosystem changes. It's a dynamic field that needs ongoing research, collaboration, and adaptation to stay one step ahead of dishonest actors. A successful cybersecurity journey necessitates not just maximizing the benefits of machine learning but also navigating and minimizing its obstacles in order to establish a comprehensive defense against a continually evolving range of cyber attacks.

6.REFERENCES

[1] IEEE Xplore - Performance evaluation of machine learning classifiers in malware detection

[2] IEEE Xplore - IOTA based anomaly detection machine learning in mobile sensing

[3] Nikam, U.V.; Deshmuh, V.M. Performance evaluation of machine learning classifiers in malware detection. In Proceedings of the 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballari, India, 23–24 April 2022; pp. 1–5. [CrossRef]

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

[5] Sethi, K.; Kumar, R.; Sethi, L.; Bera, P.; Patra, P.K. A novel machine learning based malware detection and classification framework. In Proceedings of the 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), Oxford, UK, 3–4 June 2019; pp. 1–13.

[6] Abdulbasit, A.; Darem, F.A.G.; Al-Hashmi, A.A.; Abawajy, J.H.; Alanazi, S.M.; Al-Rezami, A.Y. An adaptive behavioral-based incremental batch learning malware variants detection model using concept drift detection and sequential deep learning. IEEE Access 2021, 9, 97180–97196. [CrossRef]

[7] Feng, T.; Akhtar, M.S.; Zhang, J. The future of artificial intelligence in cybersecurity: A comprehensive survey. EAI Endorsed Trans. Create. Tech. 2021, 8, 170285. [CrossRef]

[8] Sharma, S.; Krishna, C.R.; Sahay, S.K. Detection of advanced malware by machine learning techniques. In Proceedings of the SoCTA 2017, Jhansi, India, 22–24 December 2017.

[9] Chandrakala, D.; Sait, A.; Kiruthika, J.; Nivetha, R. Detection and classification of malware. In Proceedings of the 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 8–9 October 2021; pp. 1–3. [CrossRef]

[10] Zhao, K.; Zhang, D.; Su, X.; Li, W. Fest: A feature extraction and selection tool for android malware detection. In Proceedings of the 2015 IEEE Symposium on Computers and Communication (ISCC), Larnaca, Cyprus, 6–9 July 2015; pp. 714–720.

[11] Akhtar, M.S.; Feng, T. Detection of sleep paralysis by using IoT based device and its relationship between sleep paralysis and sleep quality. EAI Endorsed Trans. Internet Things 2022, 8, e4. [CrossRef]

[12] Gibert, D.; Mateu, C.; Planes, J.; Vicens, R. Using convolutional neural networks for classification of malware represented as images. J. Comput. Virol. Hacking Tech. 2019, 15, 15–28. [CrossRef]