Malicious code detection using Machine Learning

Kakoli Banerjee
Department of Computer Science & Engineering
JSS Academy of Technical Education
Noida, India

Dr. Pradeep Kumar
Department of Computer Science & Engineering
JSS Academy of Technical Education
Noida, India

Ritika Garg
Department of Computer Science & Engineering
JSS Academy of Technical Education
Noida, India

Reeti Agarwal
Department of Computer Science & Engineering
JSS Academy of Technical Education
Noida, India

Rajvi Nanda
Department of Computer Science & Engineering
JSS Academy of Technical Education
Noida, India

Abstract—Malware, derived from "Malicious software," is a comprehensive term encompassing any software intentionally crafted to disrupt, damage, or illicitly access computer systems. Modern antivirus and anti-malware tools offer effective protection against various malware attacks. Nevertheless, due to the constantly changing landscape of malicious activities, it is imperative to curate an up-to-date database of previous malware instances. This repository serves as a valuable resource for anticipating the characteristics of future attacks and facilitating swift responses.

INTRODUCTION

Despite the significant strides made in enhancing cybersecurity mechanisms and their ongoing evolution, malware continues to pose a formidable threat in the realm of cyberspace. Malware analysis employs techniques from various domains, including program and network analysis, to scrutinize malicious samples, aiming to gain a comprehensive understanding of their behavior and evolutionary patterns. In the perpetual cat-and-mouse game between malware developers and analysts, each progress in security technology is swiftly met with a corresponding evasion strategy. The effectiveness of innovative defensive measures is contingent on the properties they leverage. For instance, a detection rule relying on the MD5 hash of a known malware can be easily evaded through standard techniques like obfuscation or more advanced methods such as polymorphism or metamorphism. These techniques alter the binary code of the malware, consequently changing its hash, while keeping its behavior intact. In contrast, establishing detection rules that encompass the semantics of a malicious sample is significantly more challenging to evade. This is because malware developers would need to implement more intricate modifications. A primary objective of malware analysis is to identify additional properties that can enhance security measures and heighten the difficulty of evasion. Machine learning emerges as a logical choice to facilitate this knowledge extraction process. Numerous studies in the literature have embraced this direction, employing various approaches with diverse objectives and outcomes.

The objective of this research paper is to review and systematize the current body of literature that employs machine learning to aid in the field of malware analysis.

This survey is crafted for security analysts, including security-focused reverse engineers and software developers, seeking to leverage machine learning to automate aspects of malware analysis and streamline their workload. Despite the escalating threat of mobile malware, Windows continues to be the primary focus across all existing platforms. Conventional rule-based systems rely on predefined rules and patterns to detect suspected malware. However, these systems face challenges in adapting to novel and evolving fraud strategies, leading to numerous false negatives and potential financial losses.

The first concern pertains to overcoming contemporary anti-analysis techniques, including encryption. The second issue revolves around the precision of modeling malware behavior, which is influenced by the selection of operations considered for analysis. The third challenge relates to the obsolescence and unavailability of datasets utilized in evaluations, impacting the significance and reproducibility of results. To address these issues, we propose several guidelines for preparing appropriate benchmarks for malware analysis through machine learning. Additionally, we highlight key emerging trends worthy of more in-depth investigation, such as malware attribution and triage. Moreover, we introduce the innovative concept of malware analysis economics, acknowledging the existing trade-offs between analysis accuracy, time, and cost. These considerations become crucial when designing a malware analysis environment.

RELATED WORK

The mentioned table includes the various research work in the field of malware detection and there main objective is included with the approaches involved.

TABLE 1. RELATED WORK ON VARIOUS APPROACHES OF MALWARE DETECTION

S. No	Author/ Yr	Purpose	Findings	Limitations
1.	Rushiil Deshmukh et al(2021)	Utilizing Machine Learning and Deep Learning for the Classification of Malware	Examines two methodologies for classifying malware: Utilizes a machine-learning approach to predict the specific class of malware to which each data point belongs among the nine available classes.	The availability of increased computing power enables the possibility of training models for a larger number of epochs. The model's accuracy can potentially be enhanced by incorporating assembly (asm) files.
2	Oladimeji Kazeem et al(2023)	Enhancing Fraud Detection through Machine Learning	Explore the application of machine learning algorithms in the realm of fraud detection and prevention. Create and implement a real-time monitoring system designed for the purpose of detecting fraud.	Utilizing ensemble models, such as Random Forest or Gradient Boosting Machines, has the capacity to elevate the accuracy of fraud detection. Enhance the quality of the training dataset by implementing more sophisticated data preprocessing procedures to effectively handle missing or noisy data.
3.	IJRASET,et al (2022)	Comprehensive Survey on Malware Detection Using Machine Learning.	The objective of this research is to discover a machine learning-based solution to address the challenges posed by malware.	Signature-based techniques encounter two significant challenges: they are ineffective in detecting new or unknown malware, and they can be easily evaded by malware variants. Implementing dynamic techniques offers flexibility but can be a time-consuming process.
4.	Muhammad Shoaib Akhtar (2022)	Advancements in Malware Analysis and Detection through Machine Learning Algorithms	This research paper demonstrates that DT, CNN, and SVM exhibited strong performance in terms of detection accuracy.	The accuracy can be improved further.
5.	S. Soja Rani et al (2020)	A comprehensive survey on various methodologies for malware detection employing	This research paper presents a comprehensive examination of the evolution of concealment techniques.	Classification-specific methods depend on the combination of various feature selection techniques.

		machine learning techniques.		
6.	Mohammed ALTAIY (2023)	Utilizing Deep Learning Algorithms for Malware Detection	This research paper aims to identify malware by utilizing the dataset generated by CTU University in 2011, which combines certified botnet traffic with normal network activity.	Difficulty in training Deep Learning Models.
7	Malak Aljabri et al (2022)	Utilizing Machine Learning Techniques for the Detection of Malicious URLs	This paper concentrates on reviewing research studies concerning the utilization of machine learning algorithms for the detection of malicious URLs. The article introduces various taxonomies and provides comparative results, contributing valuable insights to the field of malicious URL detection.	The SVM algorithm employed faces limitations in handling large or noisy datasets.
8	Ferhat Ozgur et al(2020)	Utilizing Machine Learning for the Detection of Malicious URLs	This survey suggests enhancing classifier performance in the detection of malicious websites by incorporating host-based and lexical features from the associated URLs. Random Forest models and Gradient Boosting classifiers are employed to develop a URL classifier, utilizing URL string attributes as features.	By employing J48, SVM, KNN, and NB machine learning algorithms, a comprehensive comparison can be conducted to evaluate their processing time and accuracy performance in the detection of malicious URLs. This comparative analysis aims to assess the strengths and weaknesses of each algorithm, providing insights into their efficiency and effectiveness for the specific task of identifying malicious URLs.
9	Gwanghyun Ahn et al(2022	Machine Learning-Based Method for Detecting Malicious Files	In this study, the enhancement of detection accuracy was achieved by leveraging the benefits of dynamic analysis for the detection of malicious files.	The drawbacks of both static and dynamic analysis methods are addressed, enhancing the ability to accurately detect malicious code while improving the speed at which this detection is conducted.
10	Mohamed Baset(2016)	Applying Machine Learning for the Detection of Malware	The survey's findings indicated that the accuracy of the results claimed by the authors may be influenced by certain factors, such as sample size, the number of features, or the technique employed for feature extraction.	The evaluation results lack coverage of string and byte n-grams features.

COMPARISON OF MALICIOUS CODE DETECTION MODELS

The table described below shows the advantages and disadvantages of different machine learning models.

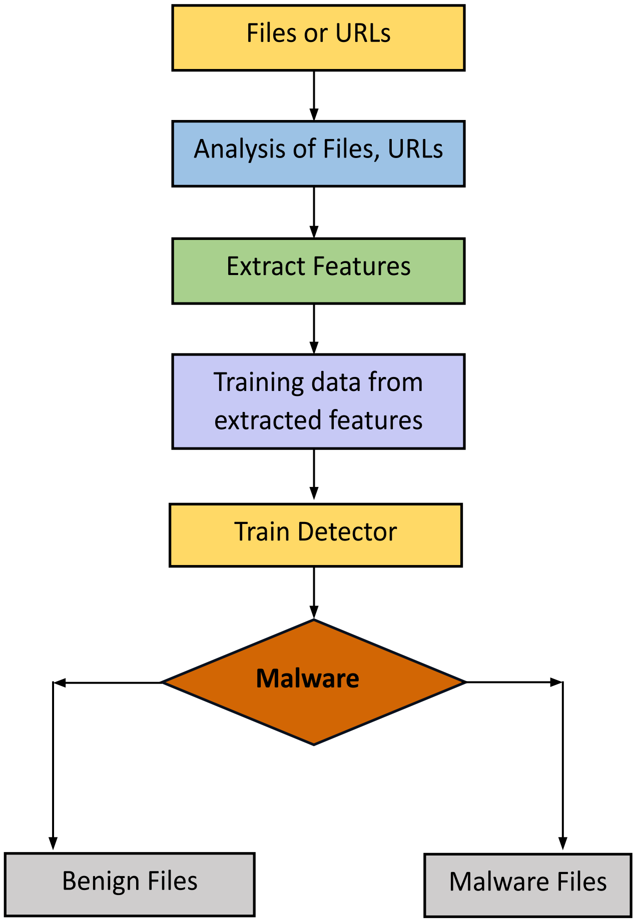
Model Name	Advantages	Disadvantages
Naïve Bayes	It enables rapid and highly scalable model building and scoring, with a linear scalability observed concerning the number of predictors and rows.	In cases where the test data contains a categorical variable representing a category not included in the training dataset, the Naïve Bayes model assigns a zero probability to that category. Consequently, the model is unable to generate predictions for this particular category.
Logistic Regression	Predicts the correlation between input features and the probability of a transaction being fraudulent. It is favored for fraud detection due to its clear readability and simplicity.	Logistic regression assumes a linear relationship between the independent variables and the log-odds of the dependent variable. This assumption may limit its applicability in scenarios with complex, non-linear relationships.
Decision Tree	Capable of handling non-linear correlations between features and the target variable, decision trees offer the advantage of being well-suited for discerning intricate patterns in fraud detection.	Decision trees are prone to overfitting, especially when the tree is deep and captures noise or outliers in the training data. Pruning or limiting the tree depth can help mitigate this issue.
Support Vector Machines	Capable of managing high-dimensional data and non-linear relationships.	The model might exhibit bias toward the majority class (legal transactions) when dealing with unbalanced datasets, resulting in decreased performance in accurately identifying the minority class (fraudulent transactions).
Random Forest Classifier	Integrates multiple decision trees to enhance accuracy and address intricate fraud patterns.	Might not yield favorable results for small datasets (those with few features) as the impact of randomness is significantly diminished.

METHODOLOGIES

The proposed methodology employs various machine learning algorithms to discern the authenticity of files and URLs, particularly focusing on malware classification. Data points are categorized into nine malware classes using three distinct ML models: logistic regression, Random Forest, and Multilayered Perceptron classifier. The models operate on key parameters such as file size in bytes, hex-code uni-gram, hex-code bi-gram, and the final feature matrix. Each model has its unique classification factor, with regularization for logistic regression, number of nodes for Random Forest, and the number of hidden layers for the Multilayered Perceptron classifier.

In the alternative Deep Learning approach, the methodology leverages convolutional operations. This involves the application of a filter of a predetermined size across groups of pixels within an image. The resulting output is generated by the interactions among these pixels, producing an image that faithfully represents the spatial and temporal dependencies embedded in the data.

In the end after applying different ML models on the given datasets we can get that whether the given file or URL is benign or malicious.



Datasets
First task is to collect data points in order to train, test and validate a ML model. These data points consist of input features (attributes or variable) and corresponding output labels. The primarily goal is to use the datasets to train a model that can learn patterns, relationships or trends.

Datasets in ML are typically divided into three subsets: Training set, Validation Set and Testing set.

Pre-Processing
The second step involves the meticulous process of cleaning, transforming, and organizing raw data into a format suitable for training a Machine Learning (ML) model. Effective pre-processing is crucial as the quality of the input data directly influences the accuracy and reliability of the trained model.

Feature Extraction
In this step, the objective is to convert raw data into a condensed representation or a set of features that encapsulates the crucial information for the learning task. This process entails selecting, combining, and transforming input variables to enhance the performance of the Machine Learning (ML) model.

Feature Selection
In this step, the focus is on selecting a subset of features that are relevant and significant from the original set. The primary goal is to enhance model performance, mitigate overfitting, and improve interpretability.

CONCLUSION

The first concern pertains to overcoming contemporary anti-analysis techniques, including encryption. The second issue revolves around the precision of modeling malware behavior, which is influenced by the selection of operations considered for analysis. The third challenge relates to the obsolescence and unavailability of datasets utilized in evaluations, impacting the significance and reproducibility of results. To address these issues, we propose several guidelines for preparing appropriate benchmarks for malware analysis through machine learning. Additionally, we highlight key emerging trends worthy of more in-depth investigation, such as malware attribution and triage. Moreover, we introduce the innovative concept of malware analysis economics, acknowledging the existing trade-offs between analysis accuracy, time, and cost. These considerations become crucial when designing a malware analysis environment.

REFERENCES

[1]Rushiil Deshmukh, Angelo Vergara, Debtanu Bandyopadhyay, Kevin Huang, et al. “Malware Classification using Machine Learning and Deep Learning.”

[2]Oladimeji Kazeem. “Fraud Detection using Machine Learning.” DOI:10.13140/RG.2.2.12616.29441

[3]Pritam Ahire, Mohanki Shreya, Shreya Shinde, Preeti Pisal, Manasi Manikumar “A Survey on Malware detection using ML”. ISSN: 2321-9653

[4]Muhammad Shoaib Akhtar, et al.Malware Analysis and Detection using Machine Learning Algorithms.

[5]S. Soja Rani, S. R. Reeja “A survey on different approaches for malware detection using machine learning techniques”

[6]Mohammed ALTAIY “Malware Detection Using Deep Learning Algorithms”

[7]Malak Aljabri, Hanan S. Altamimi, Shahd A. Albelali, Maimunah AL-Harbi “Detecting Malicious URL’s using Machine Learning Techniques”

[8]Ferhat Ozgur Catak, Kevser Sahinbas, Volkan Dortkardes “Malicious URL Detection Using Machine Learning”

[9] Gwanghyun Ahn, Kookjin Kim, Wonhyung Park Dongkyoo Shin ”Malicious File Detection Method Using Machine Learning”

[10]Mohamad Baset “Machine Learning for Malware Detection”