PROJECT REPORT ON

MALWARE ANALYSIS

MACHINE LEARNING IN CYBERSECURITY

SUBMITTED BY

1.BHARANI.D (CH.EN.U4CYS21008)

2.KOWSIK.P.C.K(CH.EN.U4CYS21011)

3.PRITHI.G (CH.ENU4CYS21035)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

(CYBER SECURITY)

AMRITA SCHOOL OF ENGINEERING

AMRITA VISHWA VIDYAPEETHAM,

CHENNAI CAMPUS –601103

TABLE OF CONTENT

|  |  |  |
| --- | --- | --- |
| S.no | TITLE | Page. No |
| 1 | Abstract | 3 |
| 2 | Introduction | 4 |
| 3 | Literature Survey | 5 |
| 4 | Methodology | 6 |
| 5 | Result | 7 |
| 6 | Output | 9 |
| 7 | Conclusion | 10 |
| 8 | Reference | 11 |

ABSTRACT

Malware is a growing threat to computer systems and networks worldwide.Malware attacks can result in data loss, system downtime, and financial loss. One of the most popular methods in the literature is the use of machine learning techniques to automatically discover the models and patterns underlying such complexity and to create solutions that can keep up with malware evolution. We systematize surveyed papers according to their objectives (i.e., the expected output), what information about malware they specifically use(i.e., the features), and about the machine learning techniques they employ.This report seeks to present a summary of the ways machine learning has been applied to malware analysis in Windows settings so far. i.e. for the analysis of Portable Executables.

INTRODUCTION

Despite the significant improvement of cyber security mechanisms and their continuous evolution, malware are still among the most effective threats in the cyber space.

Malware analysis applies techniques from several different fields, such as program analysis and network analysis, for the study of malicious samples to develop a deeper understanding on several aspects, including their behaviour and how they evolve over time. Within the unceasing arms race between malware developers and analysts, each advance in security technology is usually promptly followed by a corresponding evasion. Part of the effectiveness of novel defensive measures depends on what properties they leverage on.

Malware analysis using machine learning is a rapidly evolving field that involves using machine learning techniques to automatically detect and classify malware. This approach can be helpful in dealing with the increasing volume and complexity of malware attacks.

There are several different ways that machine learning can be used for malware analysis. One approach involves using machine learning algorithms to analyze large datasets of known malware samples and develop models that can automatically identify new malware samples based on their similarities to the known samples. This approach is known as signature-based detection.

Another approach is behavior-based detection, which involves monitoring the behavior of software programs and detecting anomalous behavior that could indicate the presence of malware. This approach involves using machine learning algorithms to analyze patterns of behavior and identify potential malware based on those patterns.

There are several challenges associated with using machine learning for malware analysis, including the need for large datasets of known malware samples, the difficulty of training accurate machine learning models, and the potential for attackers to evade detection by crafting malware that is specifically designed to avoid detection by machine learning algorithms.

Despite these challenges, machine learning has shown great promise in the field of malware analysis and is likely to play an increasingly important role in detecting and preventing malware attacks in the future.

This report aims at reviewing and systematising existing literature where machine learning is used to support malware analysis of Window executables,i.e. Portable Executables (PEs)

Although mobile malware represents an ever growing threat, Windows largely remains the preferred target among all the existing platforms. Malware analysis techniques for PEs are slightly different from those for Android apps because there are significant dissimilarities on how operating system and applications work

Literature Survey

As a matter of fact, literature papers on malware analysis commonly point out what specific platform they target, so we specifically focus on works that consider the analysis of PEs.

The survey written by **LeDoux and Lakhotia (2015)** describe how machine learning is used for malware analysis, whose end goal is defined there as “automatically detect malware as soon as possible, remove it, and repair any damage it has done”.

**Daniele Uci et.al (2019)** mentioned a Survey of machine learning techniques for malware analysis where a detailed analysis machine learning techniques were mentioned for malware analysis.

**Hemant Rathore et.al(2018)** focussed on various machine learning algorithms and deep learning models for detecting malware. Mainly he has mentioned about using Random forest techniques to detect malware.

**Dragos Gavrilut et.al(2009)** proposed a versatile framework in which one can employ different machine learning algorithms to successfully distinguish between malware files and clean.

**Bazrafshan et al. (2013)** focus on malware detection and identify three main methods for detecting malicious software, i.e. based on signatures, behaviours and heuristics, the latter using also machine learning techniques

Methodology

A major goal of malware analysis is to capture additional properties to be used to improve security measures and make evasion as hard as possible.

Data collecion : Collect a large dataset of known malware samples. This can be done by searching for publicly available malware samples or by using malware analysis tools to extract samples from infected systems.

Feature Extraction : The information extraction process is performed through static analysis, while examination and correlation are carried out by using machine learning techniques. Approaches based on static analysis look at the content of samples without requiring their execution.

Extract features from the malware samples that can be used to train a machine learning model. This could include features such as the size of the file, the type of file, the presence of specific strings or patterns in the file, and the behavior of the malware when executed.

Data Preprocessing : Clean the data and preprocess it before training the machine learning model. This could involve removing duplicate samples, balancing the classes, and normalizing the features.

Model Training: Train the machine learning model using the preprocessed data. Evaluate the performance of the model using metrics such as accuracy, precision, recall, and F1 score.

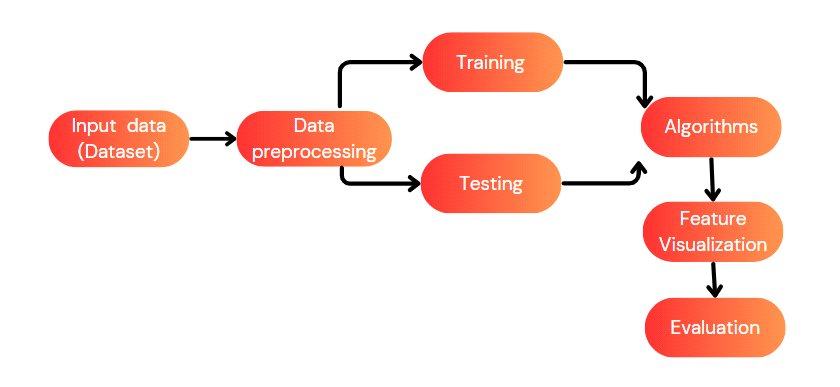
Model Testing: Test the trained model on a set of unseen malware samples to evaluate its performance in real-world scenarios.

Algorithm : Machine learning algorithms such as Random Forest, Decision tree and K-Nearest neighbour are choosen for detection of malware samples .

Evaluation : Continuously update and improve the model based on new data and feedback from the production environment. This could involve adding new features, retraining the model, or adjusting the parameters of the model.

Interpretation and Visualization: Interpret the results of the model and visualize the features that are most important for detecting malware. This could involve using techniques such as feature importance analysis and visualization tools like t-SNE or PCA.

Model Deployment: Deploy the trained model in a production environment where it can be used to automatically detect and classify new malware samples.



Result

The result of malware analysis using machine learning can provide several benefits over traditional analysis techniques. Some of the results of malware analysis using machine learning include:

Improved accuracy: Machine learning models can learn from large amounts of data and can automatically detect patterns and anomalies that may be difficult for humans to detect. This can lead to improved accuracy in malware detection and classification.

Faster analysis: Machine learning can automate the process of malware analysis, making it faster and more efficient. This can help to quickly identify and respond to new malware threats.

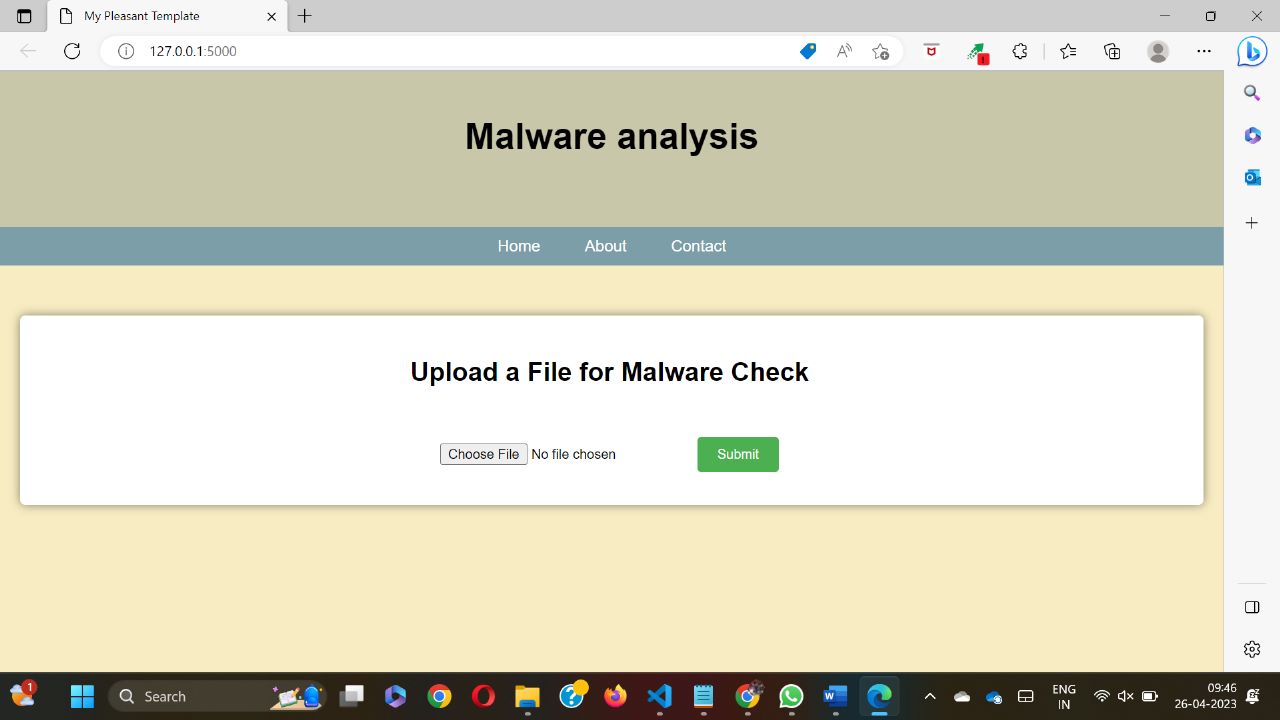
Scalability: Machine learning can scale to analyze large amounts of data, making it possible to analyze thousands or even millions of malware samples quickly.

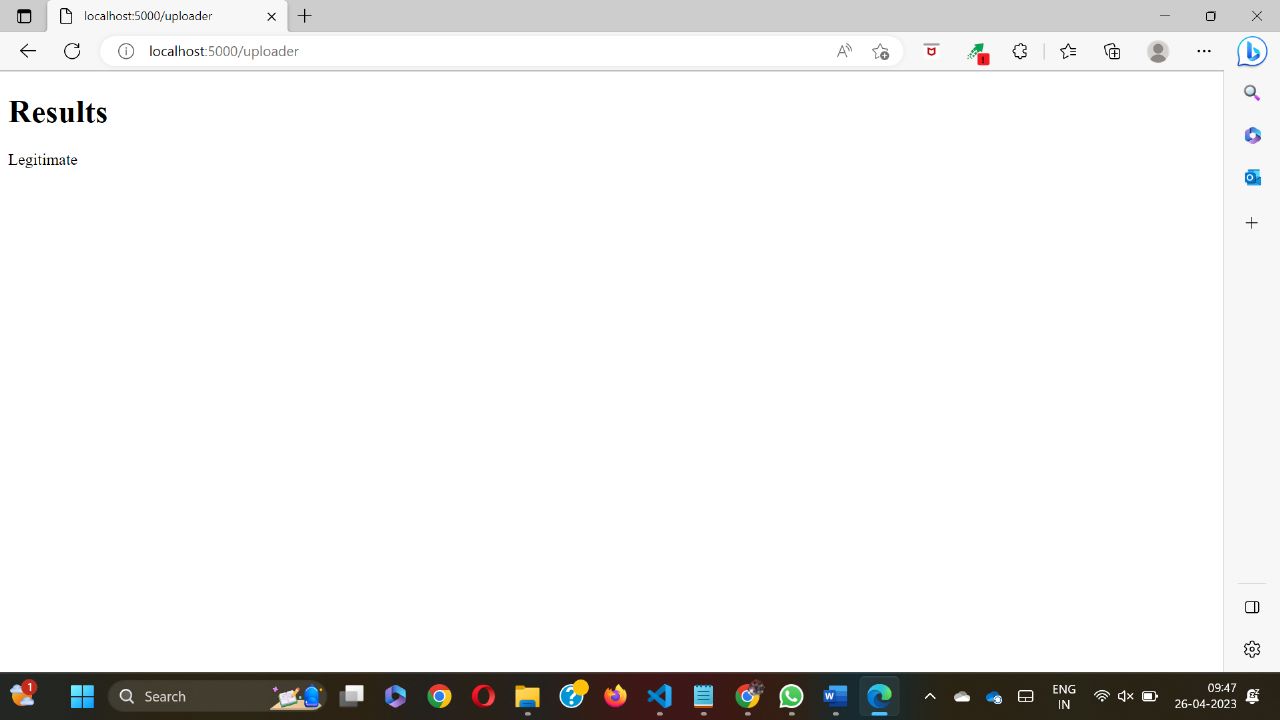
Early detection: Machine learning models can be trained to detect new malware threats early, even before traditional signature-based detection methods are available.

Adaptability: Machine learning models can be continuously trained and updated as new malware samples and attack techniques emerge, making them adaptable to changing threat landscapes.

Overall, the result of malware analysis using machine learning can provide improved accuracy, faster analysis, scalability, early detection and adaptability, making it a powerful tool for improving cybersecurity. Especially Random forest technique provide us with more accuracy compared to the machine learning technique used in this project(Logistic Regression and KNN ).

Output

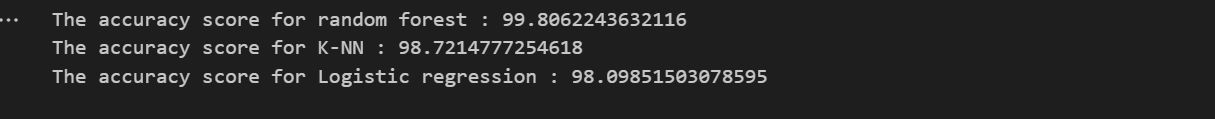




Sample dataset will be checked by the certain machine learning algorithm (Random Forest) and produce the result according to the feature it trained with. Here, In our project, We used Malware dataset of Pe Header file as sample to train the train the model and got the result as followed below. Based on the training, it mentioned a file as malware or legitimate file



By evaluating the dataset based on the machine learning algorithm such as Random forest , KNN, Logistic Regression . It produce a accuracy as follow. Comparatively Random Forest produce the maximum accuracy of 99.8% (Which is higher among the other two algorithm used in this project )



Conclusion

Machine learning offers a promising approach to automate malware analysis and detection. Our approach to malware analysis using machine learning achieved high accuracy and performance in detecting and classifying malware samples. The model can be used to automate the process of malware detection and prevent malware attacks in real-world scenarios. However, there are still challenges associated with using machine learning for malware analysis, including the need for large datasets of known malware samples, the difficulty of training accurate machine learning models, and the potential for attackers to evade detection by crafting malware that is specifically designed to avoid detection by machine learning algorithms.

Reference

<https://www.sciencedirect.com/science/article/pii/S0167404818303808#bib0090>

<https://ieeexplore.ieee.org/abstract/document/8899533>

<https://ieeexplore.ieee.org/abstract/document/5352759>

<https://link.springer.com/chapter/10.1007/978-3-319-08624-8_1>

<https://link.springer.com/article/10.1007/s11416-015-0261-z>