Malware Detection Using Machine Learning (Case Study)

Mr.Vinoth kumar[1], Mrs.Indumathy[2], Karan Krishna Kanth K G[3], Sasi kumar[4]
[1, 2]Asst.Professor, Dept of Information Technology
[3, 4]Dept of Information Technology
[1, 2, 3, 4]Rajiv Gandhi college of Engineering and Technology, Kirumampakam607403.

Abstract- *Currentantivirus software's are effectiveagainst known viruses, if a malware with new signature is introduced then it will be difficult to detect that it is malicious. Signature-based detection is not that effective during zero-day attacks. Till the signature is created for new (unseen) malware, distributed to the systems and added to the anti-malware database, the systems can be exploited by that malware. Research shows that over the last decade, malware has been growing exponentially, causing substantial financial losses to various organizations. Different anti-malware companies have been proposing solutions to defend attacks from these malware. The velocity, volume, and the complexity of malware are posing new challenges to the anti-malware community. Current state-of-the-art research shows that recently, researchers and anti-virus organizations started applying machine learning and deep learning methods for malware analysis and detection. Machine learning methods can be used to create more effective anti malware software which is capable of detecting previously unknown malware, zero-day attack etc. We propose an approach that various machine learning methods such as Support Vector Machine (SVM), Decision tree, Random Forest and XG Boost will be used.*

I. INTRODUCTION

Idealistic hackers attacked computers in the early days because they were eager to prove themselves. Cracking machines, however, is an industry in today's world. Despite recent improvements in software and computer hardware security, both in frequency and sophistication, attacks on computer systems have increased. Regrettably, there are major drawbacks to current methods for detecting and analysing unknown code samples. The Internet is a critical part of our everyday lives today. On the internet, there are many services and they are rising daily as well. Numerous reports indicate that malware's effect is worsening at an alarming pace. Although malware diversity is growing, anti- virus scanners are unable to fulfil security needs, resulting in attacks on millions of hosts. Around 65,63,145 different hosts were targeted, according to Kaspersky Labs, and in 2015, 40,00,000 unique malware artefacts were found. Juniper Research (2016), in particular, projected that by 2019 the cost of data breaches will rise to $2.1 trillion globally. Current studies show that script-kiddies are generating more and more attacks or are automated. To date, attacks on commercial and government organisations, such as ransomware and malware, continue to pose a significant threat and challenge. Such attacks can come in various ways and sizes. An enormous challenge is the ability of the global security community to develop and provide expertise in cybersecurity. There is widespread awareness of the global scarcity of cybersecurity and talent. Cybercrimes, such as financial fraud, child exploitation online and payment fraud, are so common that they demand international 24-hour response and collaboration between multinational law enforcement agencies.

MALWARE DETECTION

In such a way, hackers present malware aimed at persuading people to install it. As it seems legal, users also do not know what the programme is. Usually, we install it thinking that it is secure, but on the contrary, it's a major threat. That's how the malware gets into your system. When on the screen, it disperses and hides in numerous files, making it very difficult to identify. In order to access and record personal or useful information, it may connect directly to the operating system and start encrypting it Detection of malware is defined as the search process for malware files and directories. There are several tools and methods available to detect malware that make it efficient and reliable.

II. NEED FOR MACHINE LEARNING IN MALWARE DETECTION

Machine learning has created a drastic change in many industries, including cybersecurity, over the last decade. Among cybersecurity experts, there is a general belief that AI-powered anti-malware tools can help detect modern malware attacks and boost scanning engines. Proof of this belief is the number of studies on malware detection strategies that exploit machine learning reported in the last few years. The number of research papers released in 2018 is 7720, a 95 percent rise over 2015 and a 476 percent increase over 2010, according to Google Scholar,1. This rise in the number of studies is the

product of several factors, including but not limited to the increase in publicly labelled malware feeds, the increase in computing capacity at the same time as its price decrease, and the evolution of the field of machine learning, which has achieved ground- breaking success in a wide range of tasks such as computer vision and speech recognition. Depending on the type of analysis, conventional machine learning methods can be categorised into two main categories, static and dynamic approaches. The primary difference between them is that static methods extract features from the static malware analysis, while dynamic methods extract features from the dynamic analysis. A third category may be considered, known as hybrid approaches. Hybrid methods incorporate elements of both static and dynamic analysis. In addition, learning features from raw inputs in diverse fields have outshone neural networks. The performance of neural networks in the malware domain is mirrored by recent developments in machine learning for cybersecurity.

INTERFACE REQUIREMENT

The section gives an overview of the functionality of the product. It describe the informal requirement and is used to establish a context a context for the technical requirement specification

HARDWARE REQUIREMENT

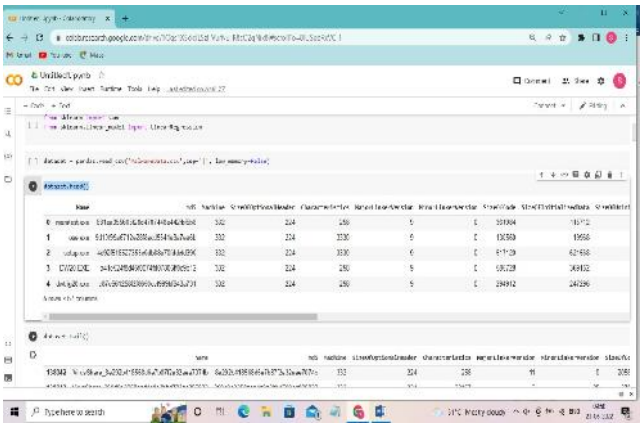
processor type : Intel i5 (10gen)
Speed :4.0ghz
Ram: 8
Hard disk:300 gb
Keyboard:101/102 standard keys
Mouse:Scroll mouse

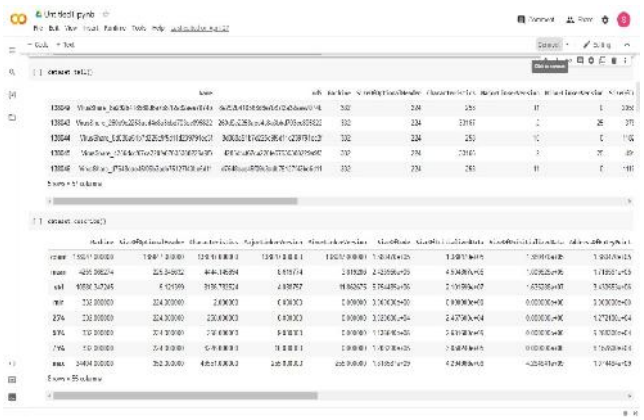
SOFTWARE REQUIREMENTS

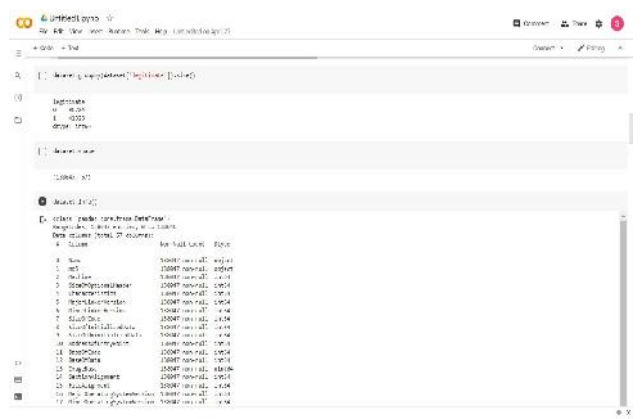
Language	Python
IDE	Anaconda
Packages	Pandas OS Numpy Sklearn

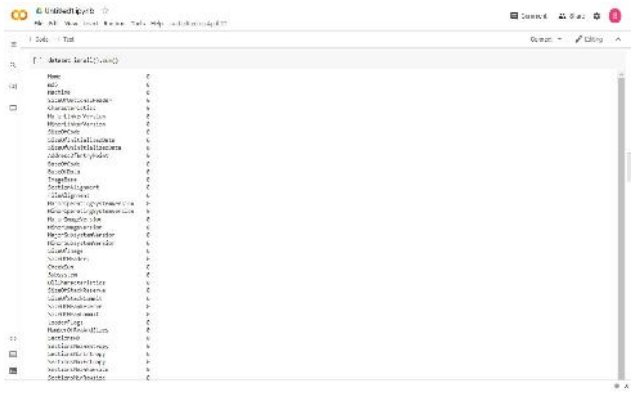
SOURCE CODE

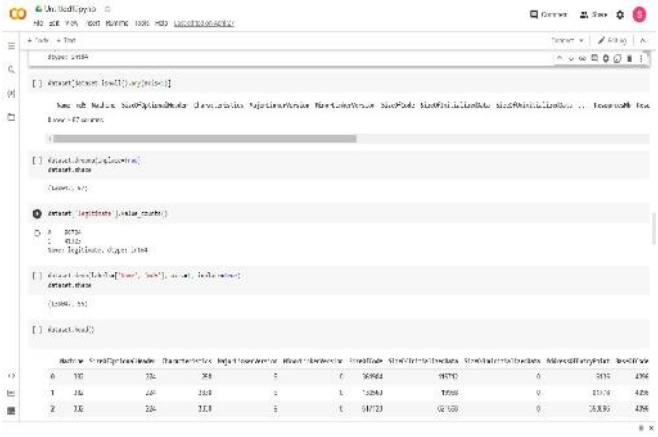
```
importos
import pandas
importnumpy

importsklearn.ensembleasek
fromsklearnimport tree, linear_model
fromsklearn.feature_selectionimportSelectFromModel
fromsklearn.naive_bayesimportGaussianNB
fromsklearn.metricsimportconfusion_matrix
fromsklearn.pipelineimportmake_pipeline
fromsklearnimportpreprocessing
fromsklearnimportsvm
fromsklearn.linear_modelimportLinearRegression
dataset =pandas.read_csv('MalwareData.csv',sep='|', low_memory=False)
```







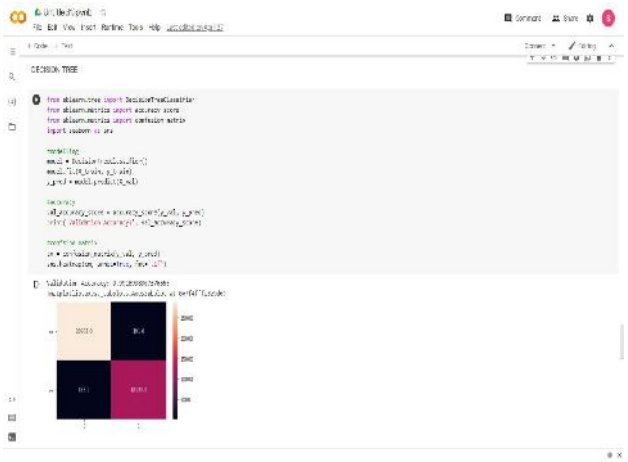




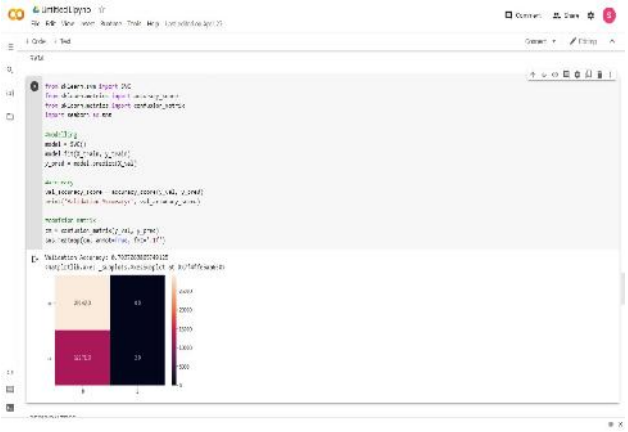
III. PROPOSED SYSTEM

Machine learning can easily identify the malware in the data and datasets Different types of machine learning algorithms are applied such as

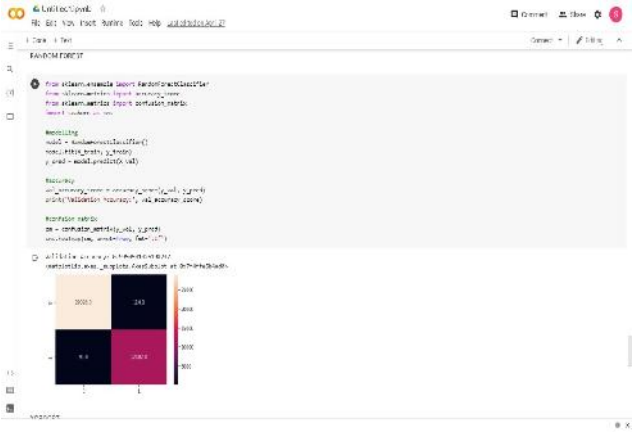
DECISION TREE



SVM



RANDOM FOREST



XG BOOST

IV. CONCLUSION

We have proposed a malware detection module based on advanced data mining and machine learning. While such a method may not be suitable for home users, being very processor heavy, this can be implemented at enterprise gateway level to act as a central antivirus engine to supplement antiviruses present on end user computers. This will not only easily detect known viruses, but act as a knowledge that will detect newer forms of harmful files. While a costly model requires costly infrastructure, it can help in protecting invaluable enterprise data from security threats, and prevent immense financial damage.

This paper presents the survey about existing literature on malware analysis using different machine learning algorithms. Table 1 defines the different literature of existing works with what are the tools used in their work, what are the machine learning algorithms they used in their work, from what sources dataset is collected, what are parameters they consider to reach their goal and the corresponding experimental results and what are the future works are proposed all are listed in the table form. In the discussion, it clearly identifies thatmachine

REFERENCES

[1] http://www.us-cert.gov/control_systems/pdf/undirected_att ack0905.pdf
[2] "Defining Malware :FAQ".http://technet.microsoft.com. Retrieved 2009-09-10.
[3] F-Secure Corporation (DecemberAmount of Malware Grew by 100% 4, 2007). "F-Secureduring 2007". Press Reportsrelease.Retrieved 2007-12-11.
[4] History of Viruses.http://csrc.nist.gov/publications/nistir/threats/subsubsection3_ 3_1_1.html
[5] Landesman, Mary (2009). "What is a Virus Signature?" Retrieved 2009- 06-18.
[6] Christodorescu,M., Jha, S., 2003. Static analysis of executables to detect malicious patterns. In: Proceedings of the 12th USENIX Security Symposium. Washington .pp. 105-120.
[7] Filiol, E.,2005. Computer Viruses: from Theory to Applications. New York, Springer, ISBN 10: 2- 287-23939-1.
[8] Filiol, E., Jacob, G., Liard, M.L., 2007: Evaluation methodology and theoretical model for antiviral behavioral detection strategies. J. Comput. 3, pp 27–37.
[9] H. Witten and E. Frank. 2005. Data mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, ISBN-10: 0120884070.