International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 07 | July - 2022 Impact Factor: 7.185 **ISSN: 2582-3930**

MALWARE DETECTION USING MACHINE LEARNING

[1] Mrs M.Indumathy, [2] Mr R.Vinoth Kumar[3] K.Karl Amarthya, [4] Praveen Raja

[1] Asst. Professor, faculty of Information Technology, Rajiv Gandhi college of Engineering and Technology, Kirumampakam 607403.
[2] Asst. Professor, faculty of Information Technology, Rajiv Gandhi college of Engineering and Technology, Kirumampakam 607403.

[3,4]Department of Information Technology, Rajiv Gandhi college of Engineering and Technology, Kirumampakam 607403.

ABSTRACT:
Current antivirus software's are effective against 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 duringzero-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

Keywords: Malware, Machine Learning, Algorithms

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 servicesand 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 cyber security. There is widespread awareness of the global scarcity of cybersecurity and talent. Cyber crimes,such as financial fraud, child exploitation online and payment fraud, are so common that they demand international 24-hourresponse and collaboration between multinational law enforcement agencies. For single users and organisations, malware defence of computer systems is therefore one of the most critical cybersecurity activities, a seven a single attack may result in compromised data and sufficient losses. Mobile phones have become increasingly important tools in people's daily life, such as mobile payment, instant messaging, online shopping, etc., but the security problem of mobile phones is becoming more and more serious. Due to the open source nature of the Android platform, it is very easy and profitable to write malware using the vulnerabilities and security defects of the Android system. This is the main reason for the rapid increase in the number of malware on the Android system. The malicious behaviors of Android malware generally include sending deduction SMS, consuming traffic, stealing user's private information, downloading a large number of
malicious applications, remote control, etc., threatening the
privacy and property security of mobile phones users. The number of Android malware is growing rapidly; particularly, more and more malicious software use obfuscation technology. Traditional detection methods of manual analysis and signature matching have exposed some problems, such as slow detection speed and low accuracy. In recent years, many researchers

have solved the problems of Android malware detection using machine learning algorithms and had a lot of research results. With the rise of deep learning and the improvement of computer computing power, more and more researchers began to use deep learning models to detect Android malware. This paper proposes an Android malware detection model based on a hybrid deep learning model with deep belief network (DBN) and gate recurrent unit (GRU). The main contributions are as follows:

(i) In order to resist Android malware obfuscation technology, in addition to extracting static features, we also extracted the dynamic features of malware at runtime and constructed a comprehensive feature set to enhance the detection capability of malware.

(ii) A hybrid deep learning model was proposed. According tothe characteristics of static features and dynamic features, two different deep learning algorithms of DBN and GRU are used.

(iii) The detection model was verified, and the detection result is better than traditional machine learning algorithms; it also can effectively detect malware samples using obfuscation technology.

EXISTING SYSTEMS

- Malware detection by using window api sequence and machine learning
- Detecting unknown malicious code by applying classification techniques on oppose patterns
- Detecting scareware by mining variable length instructions sequence
- Accurate adware detection using oppose sequence extraction
- Detection of spyware by mining executable files
- Detection by using neural networks on the malware

PROBLEM IDENTIFIED

Detecting unknown malicious code by applying classifications techniques on oppose pattern : Evaluated number of experiments and found that setting of 2 grams, TF, using 300 features selected by Df measured outperform the perform lacks ML specific techniques
Detecting scareware by Mining variable length instructions sequence: This paper present the static analysis method based on data mining which extends the general heuristic detection techniques using a variable length instructions sequence mining approach for purpose of scareware detection but metrics specific and unsupervised techniques un included can be broken

PROPOSED SOLUTION WITH ALGORITHMS

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

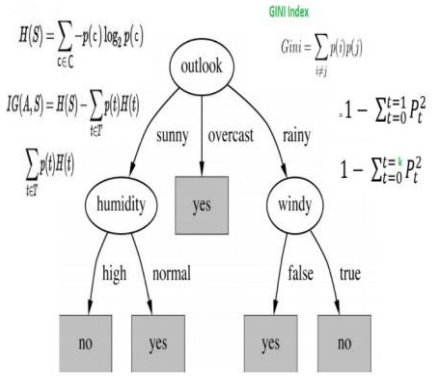
Random forest

XG boost

DECISION TREE

It is a tool that has applications spanning several different areas. Decision trees can be used for classification as well as regression problems.
The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.

In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.



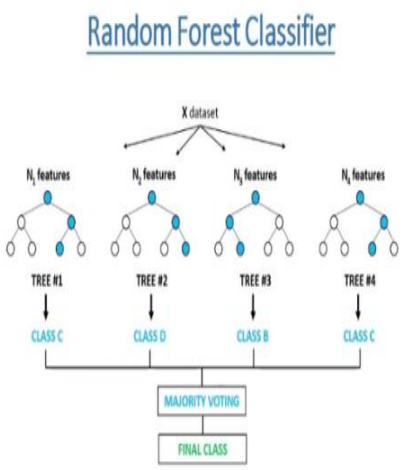
Types of Decision Trees

Types of decision trees are based on the type of target variable we have. It can be of two types:

Categorical Variable Decision Tree: Decision Tree which has a categorical target variable then it called a Categorical variable
decision tree.

Continuous Variable Decision Tree: Decision Tree has a continuous target variable then it is called Continuous Variable Decision
Tree.

Random forest



A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

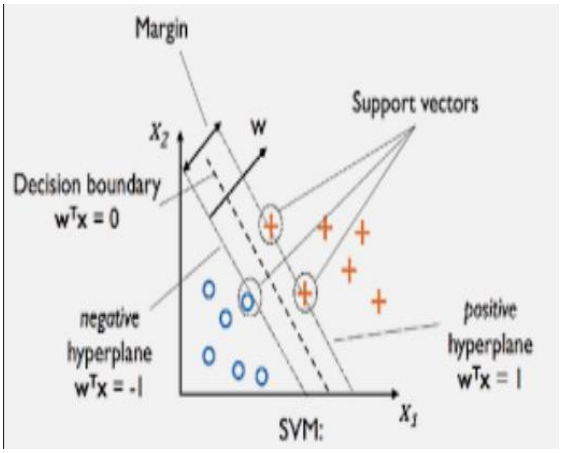
A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating.
Bagging is an ensemble meta-algorithm that improves the accuracy of machine
.

XG boost

XGBoost or extreme gradient boosting is one of the well-known gradient boosting techniques(ensemble) having nhanced performance and speed in tree-based (sequential decision trees) machine learning algorithms. XGBoost was created by Tianqi Chen and initially maintained by the Distributed (Deep) Machine Learning Community (DMLC) group. It is the most common algorithm used for applied machine learning in competitions and has gained popularity through winning solutions in structured and tabular data.

XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data.

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.



SVM :

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.
However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane.
These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. damage caused by them.

CONCLUSION :

A Malware is critical threat to user computer system in terms of stealing confidential information or disabling security. This project present some of the existing machine learning
algorithms directly applied on the data or datasets of malware It explains the how the algorithms will play a role in detecting malware with high accuracy and predictions We are also using data science and data mining techniques to
overcome the drawbacks of existing system

REFERENCES:

[1]http://www.us-cert.gov/control_systems/pdf/undirected_attack0905.pdf

[2] "Defining Malware: FAQ". http://technet.microsoft.com.
Retrieved 2009-09-10.

[3] F-Secure Corporation (December 4, 2007). "F-Secure Reports Amount of Malware Grew by 100% during 2007". Press
release. 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.
[10] J. Kolter and M. Maloof, 2004. Learning to detect malicious
executables in the wild. In Proceedings of KDD'04, pp 470-478.
[11] J. Wang, P. Deng, Y. Fan, L. Jaw, and Y. Liu, 2003.Virus detection using data mining techniques. In Proceedings of IEEE
International Conference on Data Mining.
66
International Journal of Network Security & Its Applications
(IJNSA), Vol.4, No.1, January 2012
[12] Kephart, J., Arnold, W., 1994. Automatic extraction of
computer virus signatures. In: Proceedings of 4th Virus Bulletin
International Conference, pp. 178–184.
[13] L. Adleman, 1990. An abstract theory of computer viruses
(invited talk). CRYPTO '88: Proceedings on Advances in
Cryptology, New York, USA. Springer, pp: 354–374.
[14] Lee, T., Mody, J., 2006.Behavioral classification. In:
Proceedings of European Institute for Computer Antivirus Research (EICAR) Conference.
[15] Lo, R., Levitt, K., Olsson, R., 1995: Mcf: A malicious code
filter. Comput. Secur. 14, pp.541– 566.

[16] M. Schultz, E. Eskin, and E. Zadok, 2001.Data mining
methods for detection of new malicious executables. In Security and Privacy Proceedings IEEE Symposium, pp 38-49.
[17] McGraw, G., Morrisett, G.,2002 : Attacking malicious code,
report to the infosec research council. IEEE Software. pp.
33–41.
[18] P. Szor, 2005.The Art of Computer Virus Research and
Defense. New Jersey, Addison Wesley for Symantec Press.
ISBN-10: 0321304543.
[19] Rabek, J., Khazan, R., Lewandowski, S., Cunningham, R.,
2003. Detection of injected, dynamically generated, and obfuscated malicious code. In: Proceedings of the 2003 ACM
Workshop on Rapid Malcode, pp. 76–82.
[20] S. Hashemi,Y. Yang, D. Zabihzadeh, and M. Kangavari,
2008.Detecting intrusion transactions in databases using data
item dependencies and anomaly analysis. Expert Systems, 25,5,pp 460–473. DOI: 10.1111/j.1468-0394.2008.00467.x
[21] Sung, A., Xu, J., Chavez, P., Mukkamala, S., 2004.Static
analyzer of vicious executables (save). In: Proceedings of the