International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEITCN228618

Malware Detection in Files and URL's using Machine Learning

Prof. Balaji Chaugule, Omkar Chavan, Tushar Kokane, Sagar Hyalij, Dnyaneshwar Bhosale

Department of Information Technology, Zeal College of Engineering and Research Narhe, Pune, Maharashtra, India

ARTICLEINFO

Article History:

Accepted: 02 Jan 2024

Published: 15 Jan 2024

Publication Issue

Volume 10, Issue 1

January-February-2024

Page Number

79-84

ABSTRACT

The use of Support Vector Machine (SVM) to machine learning-based malware detection is the main goal of this study. Since malware threats are always changing, more advanced detection techniques are required. With the help of SVM, a potent classification algorithm, our project is able to precisely identify malware based on a range of attributes. By developing a strong and effective solution that efficiently identifies and reduces malware threats, the aim is to improve cybersecurity and support continuous efforts to protect computer systems and networks. Malware has been posing a serious threat to embedded systems in recent times, and traditional software solutions like antivirus and patching haven't been able to keep up with the sophisticated and constantly changing bad programmes. In this work, we present guardol, a hardware-enhanced architecture designed to identify online malware. Guardol is a hybrid technique that combines FPGA and CPU. Our method seeks to capture malware's malevolent behaviour, or high-level semantics. In order to do this, we first suggest the frequency-centric model for building features out of benign samples and known malware's system call patterns. We next create a machine learning strategy in FPGA to train a classifier with these features, utilising a multilayer perceptron. The trained classifier is applied at runtime to categorise the unknown data as benign or malicious, with In this day and age of advanced technology, the internet has been embraced by most people. And with it has come an increased risk of malevolent cyberattacks by cybercriminals. The attacks are completed.

Keywords :- Malware Prediction, Machine Learning Algorithm, SVM, Malware Dataset

I. INTRODUCTION

The malicious software or malware is the infect of software created machine without the consent or knowledge of the users . It is actually a generic definition for all sorts of Threats that can affect a computer. The malware classification is a simple and it consists of a stand-alone malware and the file infectors. The objectives of a malware could include Accessing private networks, stealing sensitive data, taking over computer systems to Make use of its resources, or disrupting computing or communication operations. The malware analysis main purpose is to obtain the information needed is to provide is to rectify the instruction or network system and exactly what happened to find out is our main goal and all infected machines and files are located. To work predicts a computer driven system's chances of getting Attacked by various malwares in the base level in the time of manufacturing of The System.The system is based on the Machine Learning Algorithm techniques is like Ransomware detection using machine learning algorithms like random forest and trees, and SVM and logistic regression also in data mining projects, important insights to find the classifications and predictions techniques.Malware is like a digital troublemaker created by cyber attackers to mess with computers and software. It's basically a bunch of nasty code designed to cause chaos and sneak into networks without permission. People usually spread it through tricky links or files in emails. Once you click on those links or open the files, the malware springs to life, wreaking havoc and doing things it shouldn't. It's like a sneaky digital vandal up to no good.

Malware detection antiviruses and how machine learning works- Malicious software, or malware, poses a significant threat to computer systems and networks, aiming to cause harm and gain unauthorized access. It is crafted by cyber attackers and often distributed through deceptive links or files in emails. Machine learning, a rapidly evolving field, has proven effective in combating malware. Techniques such as supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning empower algorithms to detect and mitigate threats.In the realm of malware detection, methods can be signature-based or behavior-based, employing both static and dynamic analyses. Static analysis involves examining the malware's source code without execution, while dynamic analysis observes the file's behavior during execution in a virtual environment. The collaboration of these techniques contributes to robust malware detection systems.The training phase of machine learning models for malware detection involves collecting benign and malware binaries. A feature extractor processes these files to extract relevant information, which is then fed into the machine learning model. This model is trained to predict and classify potential threats based on the extracted features.Microsoft Defender exemplifies the integration of machine learning in combating malware. Its real-time protection capabilities leverage machine learning to detect never-before-seen threats. Context-aware detonation systems analyze massive threat intelligence, allowing for real-time attack detection. Cloud security engines receive this information, enhancing the antivirus's ability to defend against a wide range of threats, including Java malware and remote access Trojans.

Organizations, grappling with the enormity of data and the need for continuous monitoring, turn to machine learning for efficient threat identification. Einfochips, a provider of cybersecurity services, employs machine learning techniques such as signature-based algorithms, feature extraction, static analysis, and dynamic analysis to detect and classify malware types. Tools like Virustotal, Process Monitor, and Wireshark are utilized to enhance the classification process, ensuring the deployment of the secure products in the open world Through strategic and transformative operations, eInfochips assists businesses in developing, deploying, and managing security solutions globally, adhering to industry standards and guidelines. This includes compliance with NIST, ENISA, OWASP, MITRE, and IoT Security Foundation, reinforcing the commitment to safeguarding against malicious software and cyber threats.

II. LITERATURE SURVEY

Malware Detection System yanfang ye et al. Was [1] proposed for executable files which using classification techniques of object-oriented associative and which better worked as compared to traditional antivirus software and proposed system tools which used King Soft's antivirus software, [2]the Symbolic Aggregate Approximation (sax), which uses a proposed framework and also uses supervised classification methods. Azizur Rahaman, mozammel Choudhury[3] has developed a machine learning framework and techniques for data mining classification of malware and prediction, and results are obtained better compared to similar works. Doug Jacobson and Michael sgroi.

[4] "Malware Detection and Analysis the Khan mohd hamrah" Malware in Potential If we analyze the result is typically intended, it determines to suspected malware. What can you do when it is in our network, then how to detect it, and how to damage and contain the major files requires full analysis, which once we can identify in our network malware detect the infections and develop signatures its on time .

[5] Yang liv in " predicting and modeling malware which spread between the markets via" securing the android app through markets the mobile devices are recently dominated the android ecosystem .the third party markets and google play including official and markets of android app.trung kien hiroshi sato in "for API sequences the malware classification approaches are based on NLP.dynamic analysis and static alalysis which are two basic types in malware analysis field .the process involved in how functions of particular malware can understanding the functions of malware dynamic analysis using the researchers of malware could collect sequences of call API and these sources are very valuable of sources for information and it malware behavior can be identifying .

[6] Tom oh , toe young kim HO in "the efficiency of malware analysis can be increase the analysis of malware for the android system in the market the products of malware analysis are existing after the survey of comprehensive the result which shown the survey which is tool of assistive and they are needed researchers to help and applications are large number in android malware to predict automatically.

[7] In the realm of malware research, Keita Kishioka and Kouji Hirata delve into predicting malware infection spreading on overlay networks, employing a model based on hosts. Their approach estimates the degree of infection through simulation experiments.

[8] Surya Nepal and Yu Wang address the pressing issue of Android malware, emphasizing the diminishing efficacy of traditional detection methods against evolving threats. They propose A3CM, an Automatic Capability Annotation system, to enhance security and privacy for mobile users.

[9]Meanwhile, Gang Li and Ping Xingo tackle the escalating risk of Android malware. Their focus lies in detection through contrasting permission patterns, highlighting the significance of analyzing required permissions in understanding and combating the growing menace on the Android platform.

In the realm of cybersecurity research, [10]Smita Jangale and Anish Chaudhary delve into the challenges of our data security in their work on "Detecting Malware, Malicious URLs, and Viruses Using Machine Learning and Signature Matching." With electronic devices storing a significant portion of our data, the escalating threat of infections by viruses, malware, worms, Trojans, ransomware, and other unwelcome intruders is a pressing concern. This surge in risks can be attributed to the widespread accessibility of the internet.Shifting focus to the domain of malware analysis,[11] Hiroshi 0Sato and Trung Kien contribute insights in their study, "NLP-based Approaches for Malware Classification from API Sequences." In the intricate landscape of malware analysis, researchers employ static and dynamic analysis as foundational methodologies to comprehend the workings of specific malware. Dynamic analysis, in particular, empowers malware researchers to amass valuable information from API call sequences, offering crucial insights into identifying and understanding malware behavior.

III. METHODOLOGY

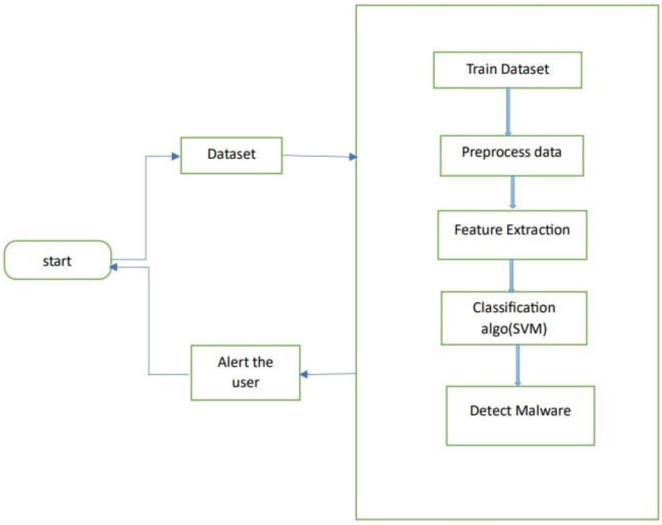
Support Vector Machine (SVM) stands out as a versatile tool in the realm of Supervised Learning, adept at tackling both Classification and Regression challenges. Although its application extends to Regression, SVM primarily finds its prowess in solving Classification problems within the domain of Machine Learning.

The fundamental objective of SVM lies in crafting an optimal decision boundary, often referred to as a hyperplane, capable of effectively partitioning n-dimensional space into distinct classes. This delineation facilitates the seamless categorization of new data points in subsequent instances.

At the core of SVM's methodology is the identification of key points, or support vectors, strategically chosen to define the hyperplane. These support vectors play a pivotal role in shaping the decision boundary and contribute to the algorithm's nomenclature as Support Vector Machine.To illustrate this process, consider the accompanying diagram where two distinct categories are intelligently classified through the imposition of a well-defined decision boundary or hyperplane. SVM, with its reliance on support vectors, emerges as a robust approach for discerning patterns and making informed predictions in various real-world scenarios.

IV. PROPOSED SYSTEM

System architecture



V. PURPOSE

Hence our main goal is to be provide the web framework where the users can upload the files and machine learning classificiation algorithm will test the uploaded files is the malicious or not. and result will be shown at front of users and to prevent the users from various malware attacks.

The primary objective of this system is to assess the likelihood of a Windows machine falling victim to different malware families, utilizing various machine properties. The essential data for predicting malware encounters encompasses any information pertaining to the compromised state of a computer under a malware attack.

VI. CONCLUSION

The better technique of classification is a svm which can be used for malware detection and better generalization needs to attentation and construct better feature of representation.

Support Vector Machine (SVM) endeavors to discover the optimal margin, which is the distance between the separating line and the support vectors representing data points. This pursuit of an ideal margin serves to minimize the risk of errors in classifying the data, making SVM a robust tool for effective classification tasks.

VII. REFERENCES

[1]. StatCounter. (2018). Mobile Operating System Market Share Worldwide. Accessed: Mar. 19, 2018. [Online]. Available:http://gs.statcounter.com/osmarket-share/mobile/worldwide

[2]. G. Suarez-Tangil and G. Stringhini, 'Eight years of rider measurement in the Android malware ecosystem: Evolution and lessons learned,' CoRR, vol. abs/1801.08115, pp. 1–18, Jan. 2018. [Online].Available:http://arxiv.org/abs/1801.08115

[3]. N. Sun, J. Zhang, P. Rimba, S. Gao, Y. Xiang, and L. Y. Zhang, 'Datadriven cybersecurity incident prediction: A survey,' IEEE Commun. Surveys Tuts., vol. 21, no. 2, pp. 1744–1772, 2nd Quart., 2018.

[4]. L. Ma, X. Liu, Q. Pei, and Y. Xiang, 'Privacy-preserving reputation management for edge computing enhanced mobile crowdsensing,' IEEE Trans. Services Comput., vol. 12, no. 5, pp. 786–799, Sep./Oct. 2019.

[5]. J. Qiu, W. Luo, L. Pan, Y. Tai, J. Zhang, and Y. Xiang, 'machine learning predicting the impact of malicious android,' IEEE Access, vol. 7, pp. 66304–66316,2019.doi:10.1109/ACCESS.2019.2914311.

[6]. L. Liu, O. de Vel, Q.-L. Han, J. Zhang, and Y. Xiang, 'Detecting and preventing cyber insider threats: A survey,' IEEE Commun. Surveys Tuts., vol. 20, no. 2, pp. 1397–1417, 2nd Quart., 2018.

[7]. Y. Zhong, H. Yamaki, and H. Takakura, 'A malware classification method based on similarity of function structure,' Applications and the Internet (SAINT), IEEE/IPSJ 12th International Symposium, pp. 256- 261, July 2012.

[8]. U. Pehlivan, N. Baltaci, C. Acartürk, and Baykal, 'The analysis of feature selection methods and classification algorithms in permission based Android malware detection,' Computational Intelligence in Cyber Security (CICS), 2014 IEEE Symposium, pp. 1-8, December 2014.

[9]. Y. Feng, O. Bastani, R. Martins, I. Dillig, and S. Anand, 'Automated synthesis of semantic malware signatures using maximum satisfiability,' in Proc. 24th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS), San Diego, CA, USA, Feb./Mar. 2017, pp. 1–15. [Online]. Available:https://www.ndss,symposium.org/ndss2017/ndss-2017-programme automated-synthesis-semantic-malware-signatures-using maximum satisfiability/.doi:10.14722/ndss.2017.23379

[10]. F. Wei, Y. Li, S. Roy, X. Ou, and W. Zhou, 'analysis of Deep ground truth Android of current malware,' in Proc. Int. Conf. Detection Intrusions Malware, Vulnerability Assessment. Cham, Switzerland: Springer, 2017, pp. 252–276

[11]. Gavrilut D., Cimpoesu M., Anton D., & Ciortuz L., 'Malware of Prediction Using Machine Learning', International Multiconference on Information Technology, 2009 and computer science.

[12]. Rhode, M., Burnap, P., & Jones, K 'Early-stage malware prediction using recurrent neural networks', computers & security, 2018.

[13]. Baset, M, 'Machine Learning For Malware Detection', 2016.

[14]. Yeo, M., Koo, Y., Yoon, Y., Hwang, T., Ryu, J., Song, J., & Park, C., 'Flow-based malware detection using convolutional neural network', 2018 International Conference on Information Networking, 2018.

[15]. 'Features',lightgbm,Documentation. [online]Available:https://lightgbm.readthedocs.io/en/latest /Features.html.

[16]. E. Mariconti, L. Onwuzurike, P. Andriotis, E. De Cristofaro, G. Ross, and G. Stringhini,

'MaMaDroid: Detecting Android malware by building Markov chains of behavioral models,' in Proc. 24th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS), San Diego, CA, USA, Feb./Mar. 2017, pp. 1– 15. [Online]. Available: https://www.ndss-symposium.org/ndss2017/ndss 2017-programme/mamadroid-detecting-android-malware-building,markov-chains-behavioral-models/.doi:10.14722/ndss.2017.23353.

Cite this article as :

Prof. Balaji Chaugule, Omkar Chavan, Tushar Kokane, Sagar Hyalij, Dnyaneshwar Bhosale, "Malware Detection in Files and URL's using Machine Learning ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 10, Issue 1, pp.79-84, January-February-2024. Available at doi : https://doi.org/10.32628/CSEITCN228618
Journal URL : https://ijsrcseit.com/CSEITCN228618