International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 05 | May -2020 ISSN: 2582-3930

Malware Detection System

Naman Mediratta[1], Abhishek Dogra[2]

[1]*Dept. of Computer Science and Engineering,HMR Institute of Technology and Management*
[2]*Dept. of Computer Science and Engineering,HMR Institute of Technology and Management*

Abstract -Malware Detection Engines of present relies on the concept of the using signatures for detection of malware and/or other software which might be determent to security of the system. The problem with them is that the signatures used need a human supervision as well as the need for constant upgradation in the system when the malware morphs into different variants of the malware.Also false positive is another major issue with the system since many a times good software gets identified as malware and real malware slips away

Another problem with this approach is that if the source code of malware is modified slightly the signature detection model fails since the signature changes as source code is altered also encoding the software in different formats also leads to the same problem.Above stated are few of the problems encountered by any tradition antivirus.Our goal in teaching computer more specifically a machine learning algorithm is to detect malware without relying on explicit database of signatures rather feeding the system with multiple malicious files and using them to learn from and detect any variation in the malware to difference between malicious and safe files.The spread of malware in today's computer system has become a major issue and the biggest threat to computer system's worldwide today. Malware analyst are no longer able to keep up with the ever-changing and evolving threat of polymorphic malware that exist today. To solve this problem of polymorphic malware we propose to use various machine learning algorithm. For this we will be working with the PE signature of exe files along with a module for converting the files to PE signatures. We will be using the PE as signatures for our analysis in identifying files as malicious or not.

Key Words:antivirus, PE , malware , machine learning ,viruses

1.INTRODUCTION

Along with the time, the development of computer technology and its evolvement has led to the development of the threats that became more and more prominent along with time in order to obtain sensitive information through the world of computers. This in need of hour lead to development of the anti-viruses that help in protecting our system from such threats. The time period of 1990 to 2000 was the major time period where the development of anti-virus started on a major level. The threat that the computer world faces from virus attacks is being handle by the army of anti-virus that are constantly being developed parallel along with the virus that are resent in the environment.

In start, anti-virus usually focuses on preventing the system from downloading the malicious codes from planting themselves in the system. As time has passed more threats like malware, ransom-ware and even ad-ware started to get developed. This led in the development of the programs for the computer protection's resulting in explosion of antivirus in the market as people began looking for some kind of solution that could help in keeping their pc safe. It wasn't always easy since most antivirus software took u a lot of resources of the system hence affecting the computer performance, resulting in affecting the performance of the system

Antivirus is defined as the special security system that works in providing protection to the system from the type of attacks that are resent outside the system, these attacks are generally written or defined as malware attacks or virus affecting the system, the anti-virus along with the firewall of the windows or mac act as a protective shield for the user system from the outside dangers that keep on luring to attack. Sometimes it is seen that in case the firewall of the system fails, then antivirus act as a big relief package as it protects the system from the malware attacks that were able to penetrate through the firewall defense. Also, in case of a malware or virus which has already affected the system can be debugged with the help of the antivirus as it helps in finding and destroying that file which contains the malware or it tend to completely clean the malicious software from the operating system.

Anti-virus software uses various kind of techniques that are helpful in order to identify the malicious software, that might be resent which often tend to self-protect themselves and hides deep inside in an operating system. These Advanced malware's may tend to use undocumented operating system functionality and obscure techniques in order to persist and avoid being detected in the operating system. Since a large amount of virus attack have started to appear on a large attack bases these days, it has led to the development of antivirus software in such a manner that it is designed in order to deal with all kinds of malicious activities or payloads that are continuously coming from both trusted or untrusted sources.

2 WORKING OF ANTIVIRUS

The body of the paper consists of numbered sections that present the main findings. These sections should be organized to best present the material.

It is often important to refer back (or forward) to specific sections. Such references are made by indicating the section number, for exampleAn antivirus program when installed on your Windows-based computer system, the first thing it does is that it will run in the background along with the operating system while you are using your pc. It will make the files undergo thorough scan the files that you have opened in order to access or use that file, it will keep on looking for any suspicious codes or you can say malware that might be embedded within those file that can be considered or flagged as a virus or a threat that could harm the system in one or the other way by either crashing the system or hacking your personal information. normally a malicious malware or computer bug which might be resent in the file is prevented or not allowed by the computer anti-virus by not allowing you from opening the file, which would may result in the release of the virus that could result in havoc in the system. To fix it, the virus would need to be contained and have to be quarantined and then either destroyed or removed by the antivirus software.

Some malwares are in the form of mails that when opened by the user, may result in malware attacks on your system software programs that can create havoc for the user. In order to counter that the anti-virus software generally tends to scan the email attachments that are obtained by the user and check for any kind of threats that may be resent in it before you open it. Many malwares that go with the name of trojans and worms are generally stopped right there in their tracks by the anti-virus security program, in this manner hence, not letting them loose and giving them opportunity to take control over the user system or getting an access to the user email account.

With the advancements in the strength of the viruses and malwares, as an extra security measure, the programs or files that you normally access while using your system can be manually scanned by the user giving them an extra right and the feeling of sense of security that they can run the test periodically whenever they want or they feel threat and are not only dependent on the anti-viruses software to scan and check for malwares . It will follow the similar procedure that generally an antivirus software does when they run themselves going through all of your files that are being used while continuously checking to see if there is some kind of malware that may has found its way onto the user computer. For those viruses or malware and other threats that it catches, it will remove them completely, preventing them from causing any kind of damage or preventing it from spreading to other files.

Even If a computer does gets infected by a malware even after going all those procedure of scanning by the antivirus, still the antivirus software is capable of removing it from the system or file in which it might be present, most of the time, but still it becomes more and more difficult to do as the malware has already spread and affected the system. These Viruses, malware and other threats can easily spread malicious code to several different places once they are able to affect your system or your pc. This result in a big risk of not able to find all the strains of virus that get spread in the system making it difficult since if even a single strain get remain in the system then you are at a risk of having another attack, as the virus is stills capable of causing damage, especially if it's embedded in your computer's registry keys.

3 MACHINE LEARNING WITH VIRUSES

In today world scenario the growth of the artificial intelligence or commonly called as AI has tend to result in increase in the capabilities of all the things that uses it in its functioning, resulting in more powerful and widespread result of output and effectiveness, due to this we are expecting and currently seeing the growing use of AI systems in the viruses or malware that are made which is leading to the changes in the landscape of threats that have become eminent in the current world, talking about the current possibilities of expansion of the existing threats due to the use of the AI along with the virus. The most prominent that can be seen is in the decrement of the workforce which result in decrease in the costs of attacks as workforce gets cut by the AI which could have required more humans labor to perform, the intelligence and expertise. Naturally effecting area would expand with the use of AI as frequency can increased to carry out particular attacks, the speed at which the rate of these attacks can be carried out can be increased drastically with the help of the use

of an AI, also we can increase the number of the potential targets that can be targeted with the help of an AI.

ML or AI can help in Introduction of new threats that were not possible previously easy to create before. New attacks are possible with the help of the AI, the use of the AI systems result in completion of the tasks that were difficult to complete before, tasks that would be otherwise impractical for humans to do or perform. AI has helped in bringing change to the typical characters of the virus hence resulting in the change in the characters of the threats. The growing use of AI has resulted in increased growth in the effectiveness of the virus attack and the range that it gets of the target that it is going to attack due to this the anti-virus that work without the use of ML and also based on older methods do have a great level of vulnerability faced by them due to the arise of these new kinds of viruses.We believe there is reason to expect attacks enabled by the growing use of AI to be especially effective, target specific, difficult to attribute, and likely to exploit vulnerabilities.

Hence the need has been generated for the requirement of use of AI and ML along with the anti-virus technology in order to cope with the incoming difficulties that are becoming a threat to the computer world along with the current ongoing time. In resent scenario, there has been a lot of suggestion regarding the use of AI technologies like heuristic technique or using data mining techniques, data agent techniques or becoming immune to artificial intelligence or setting up an artificial neural network that is to be believed that it can help in establishing a higher performance gradient for the antivirus detection, which on other hand will result in promoting the production of new AI based algorithm resulting in stronger antivirus system making it fruitful for malware detection.

By integrating AI with antivirus can result in growth in the development of the ability of self-discovering of malware in an anti-virus software. Also, it can become really helpful in analyzing the code in an intelligent way through the use of AI by performing analysis of the code with the use of AI as well as discovering unusual or unopened system calls. In starting the efficiency of the anti-virus alone based on AI can be low but combining it with other traditional method, the efficiency can be increased drastically. Basically, artificial intelligence works like a human brain where it thinks and learn from the experience and based on that it is able to plan and act based on what it has learned from its previous counter.This improvement has resulted in considering the inclusion of artificial intelligence alongside antivirus software in order to obtain higher degree of accuracy and effectiveness in war against the malware and getting a highly effective anti-malware weapon.

4 PROBLEM WITH CURRENT ANTIVIRUS

The number of cyber attacks is constantly growing in today's environment. As the number of cyberattackers is constantly on the rise so is the risk associated with them is increasing at an even faster rate.We are seeing multiple cyber attack daily ranging from bots to automated tools making it impossible for outdated solutions such as legacy antivirus to keep up. With the rise of ransomware attacks like WannaCry and Not-Petya being some of the attacks causing the most damage. One of the major concerns with antivirus is that it is not able to detect the new attack vectors which are constantly changing in today's threat environment. No traditional antivirus can deal with these problems. Consider the following example where let us assume there is a program of 50,000 lines let us assume there are 100 lines among this program which are added by a malicious actor for causing harm and leaking information of whoever is using the said program.Now the way in which a antivirus will detect this is by having a malware researcher go through the working of the program and find out the said piece of malicious code and make an entry in the antivirus database for known malicious code this is usually done in the form of signature of the code the antivirus program later looks for the said signature in all the files it scans to identify if any of them are infected.This works for most of the viruses but the underline problem with this system is that if the hacker modifies the code and rewrites the code by changing some amount of it , it now becomes a whole new malware and we need to add the signature of this malware in the system as well. This again leads to the problem that is the database which the antivirus would have to go through is ever increasing. Modification to preexisting malware to create a new and modified malware is a common practice among hackers which is difficult for antivirus to handle

5 OUR PROPOSAL

The malware detection system rely highly on the constant input by the developers to update the database of signatures for malicious files and to keep them updated with the latest signatures of malware and variations of preexisting malware to perform. We propose a machine learning based solution to this problem by relying on the underline feature of the executables , extracting a group of related features of the executables using the PE file system and performing machine learning analysis on them to get to a model which can determine whether a file is malicious or not

depending upon characteristics found using the PE file system.The characteristics which we used are machine, size of optional header, characteristics, minor link version, image base, major sub system version, sub sytem, dll characteristics, size of stack reserve, sections max entropy, version information size are some of the characteristics of PE file system which we used for the training of model.Along with data for malicious files we also had data for the legitimate / benign files the reason behind this decision being that there are thousands of software's created everyday by multiple companies everyday and some of them can be identified by our system as malicious even when they are not. To prevent the problem with false positives we added the legitimate files to reduce our false positive rate.

This can be used as a static approach to detection of malware rather than the dynamic approach that being said our approach is not the only approach possible and is in now way replacement for the approach used by multiple paid malware detection system in the market.Our approach looks at one of the static method which can be used for static malware detection system.

6 MACHINE LEARNING IN ANTIVIRUS

Malware detection is the best problem which can be solved by a machine learning algorithm. if we see at its core malware detection is a classification problem we can use different approaches to solve this problem. Other than classification we can try using clustering as an approach to solving our problem also almost all machine learning techniques can be used to tackle the problem of malware detection but clustering and classification are the one which produce the best output.

Machine Learning can solve the problem with modification of the code since we can use machine learning to train our model by feeding into it the characteristic of the malicious software and/or code and train our model on that instead of adding multiple signatures for the variation of the same malware. As we increase the data being feed to out model it will get more accurate with time and able to predict threats not seen before or being seen for the first time.

For machine learning based malware detection system more data will produce better results so it is safe to say it is better to go for data driven approach.A model depends on the data it has during the training phase to get to a stage where it can predict accurately. Randomizing the data is going to be another main issue for this problem since we need to get enough random data to ensure our model is not biased towards certain results.The rate of false positive must be low since all positive cases in malware detection need to be removed from the system by the method of permanet deletion.If benign files are identified as malware then it may result in inconvenience in the least and financial losses in the worst case scenarios.Detection system has to take into consideration the fact that our data is not fixed and we will be adding more malware signatures to the system since the nature of malware is that it is ever-evolving and we will see new malware coming out everyday in the wild.

Along with malware signatures we also need to train our model on the safe software since thousand of companies produce software everyday and our system need to detect them as safe.Dealing with polymorphic viruses for the changing hashes is another problem for the system since every change in the malware code will result in production of a new hash value and the legacy approach of verifying signatures of a file against a database of know malicious file may not work since the database's are too huge and it takes too long for the file to be checked, increase in time for checking each file results in a exponential increase in scan time since each file will need to be checked against the complete database.

7 PE FILE SYSTEM

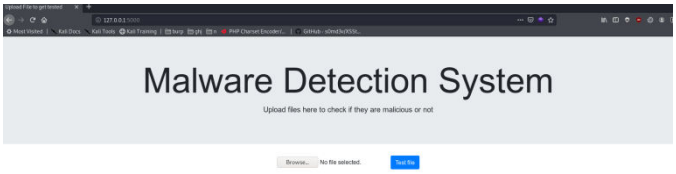
PE stands for Portable Executable file system this format is a file format for executables used commonly in Windows operating system.It was first introduced in windows 1993 operating system more commonly known as Windows N.T. 3.1 Operating System.The word portable means that it can be used for various operating system environments and architecture. The format contains data which is used by the windows OS to get library files, API call details etc.

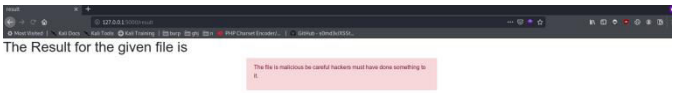
A File SystemPE file system is made up of two parts namely header and section that tell the linker on the method of mapping the file in the memory of system. The header is further divided into subsections. The header generally contains information about the file structure of the executable and location and its sizes.The section part is the main part of the PE system. It contains data, resources,code,and executable files. The header and sections are further divided into many parts mentioned ahead.

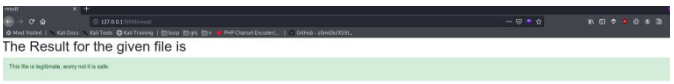
A PE file starts with a header which tells whether the file is a PE file or not followed by the magic number which usually tells if the file is compatible with the dos system or not. The next part of the header section is the DOS stub which is used for error and contains the line

for "The program cannot be run in DOS mode".The next major part of the header section is the PE header which is split into three parts namely Signature, File header, Optional Header. Signature contains the value DWORD which is used to tell that the actual information for the executable will start from here, the information before was for the operating system know how to handle the said file.File header contains file characteristics namely Machine, Number of Sections , Time Date Stamp, Pointer to Symbol, Number of Symbols, Size of optional header, and lastly characteristics. The optional header contains the characteristics magic, Address of Entry Point, Section Alignment, Size of Image, File Alignment, and finally data directories.The table after the optional header is the Section Table which is the second part of the PE file system. It contains Name1, Virtual Size, Size of Raw data, Pointer to Raw Data, and lastly characteristics.The table after it is the PE file section which contains the actual content of the executable and includes .text which is known as code is the place where all the instructions reside these instructions are the one's which are executed by the CPU when the executable is executed, the second section is rdata or read data this contains the read only data like the literals and constants, the next section is data which contains the data of the program which can be accessed anywhere from the program and finally the rsrc or the resources section which contains images, icons etc. typically the resources required by the program.

8 RESULT







- We used around 140,000 samples of PE file system gathered from various sources to train our model.
- We were successfully able to design a malware detection system for exe files using the PE file system and machine learning.
- Machine learning methodology is the best way to go for making static analysis tool for malware detection.
- Random Forrest Classification Algorithm proved to give the maximum accuracy (~90%) in our research.

9 CONCLUSIONS AND FUTURE SCOPE

The system we created can be used for detection of malware for any exe file. The system works only for exe as of right now but efforts can be made to add other functionality to it as well which might include the options for scanning other files like pdf , docx , xlsx etc.Malware and antivirus is the arms race of 21'st century just like how nuclear weapons were for the 20'th century. Development for malware will never stop and so will be the need for making better antivirus solutions.

ACKNOWLEDGEMENT

We would like to thank our project guide Mr. Anil Vats for his support, encouragement and guidance during the period of our training and helping us to make "Malware Detection System". He taught us a lot about python, machine learning, PE file system and guided us through every step of our project. He gave

us lot of inspiration during the training which helped us a lot and due to his supportive feedback, we were able to complete our project named as Malware Detection System.

REFERENCES

1. Yanfang Ye · Dinging Wang · Tao Li · Dongyi Ye ·Qingshang Jiang"An intelligent PE-malware detection system based on association mining".ResearchGate. Journal of Computer Virology
2. Dragos Gavrilut . "Malware Detection Using Machine Learning". International Multiconference on Computer Science and Information Technology
3. Priyanka Singhal. Narassha Raul "Malware Detection Module using Machine Learning Algorithm to Assist in Centralized Security in Enterprise Networks". Research Gate.
4. Kaspersky Labs. Machine Learning Methods for Malware Detection. Enterprise Security by Kaspersky labs whitepaper
5. Computer Hope. "virus signatures".https://www.computerhope.com/jargon/v/virus-signature.htm
6. https://blog.lucideus.com/2019/09/portable-executable-file.ht