European Journal of Computer Science and Information Technology

Vol.9, No.1, pp.48-61, 2021

Print ISSN: 2054-0957 (Print), Online ISSN: 2054-0965 (Online)

THE IMPORTANCE OF MACHINE LEARNING TECHNIQUES IN MALWARE DETECTION: A SURVEY

Maryam Shabbir
Bahria University, Karachi, Pakistan

ABSTRACT*: In the current age, keeping pace with the evolution of malware is becoming immensely challenging each day. In order to keep up with the unconventional trend in the development of malware, it is imperative to develop intelligent malware detection methods that accurately identify malicious files from real world data samples. The sheer complexity and volume of malware attacks on a day-to-day basis has given rise to the need of utilising machine learning techniques for dynamic analysis of files and data. In this paper, types of malware are described to understand the scope of the problem and the traditional techniques that are used for malware detection. Dynamic and behaviour-based detection methods coupled with machine learning techniques are considered to be at the core of future research and progress. Unfortunately, there are still a plethora of problems and challenges to overcome like polymorphic malware, black-box models of machine learning algorithms, reverse engineering, theoretical and practical research gaps that limit our progress and success. It is crucial to find solutions as malware experts are also exploring and exploiting the concepts of machine learning for advanced malware development and better elusive techniques. Additionally, it is required to bridge the gap between malware and machine learning experts. Their combined expertise can secure better results. In conclusion, future research direction in the field of malware detection is presented.*

KEYWORDs: Behaviour-based Detection, Dynamic Malware Analysis, Machine Learning, Pattern Recognition, Signature-based detection, Static Malware Analysis.

INTRODUCTION

Since the dawn of computing, computer programmers have shared a frontline with malware developers and with each day and age it has grown in its innovation and complexity. The term Malware is short for malicious software and It encapsulates all computer viruses and harmful computer programs that are used for disruption and damage of other computers. Malware can be directed towards a single computer, a network of computers, a server, or a network of servers, depending on what it is being created for and what the intended purpose is.

The beginning of the 21st century has seen profound technological advancements. Technology has revolutionized every person's life. Most, if not all, of the worlds population is some how reliant or connected in this technological boom. The recent advancements in the computing industry and new technologies such as IoT (Internet of Things), cloud computing and big data and developments in communication industry has vastly increased global connectivity (mostly via the Internet). Consequently, this has made it easier to inject and distribute malware on a much larger scale.

Although having humble beginnings as benign computer viruses, the malware industry has grown to gigantic proportions. Cyber security has become extremely difficult, presenting

cyber-security analysts with new challenges day by day. Malware threats have continued to expand on all new technologies. Targets can be in any public or private sector corporate companies, government organizations, and even individuals. Malware can monitor their behavior, sabotage their activities, steal sensitive data, send spam messages or threats, steal personal information, snip bank passwords, transfer information etc. In order to cope with this rapidly growing industry, there is a constant need to improve cyber-security defense strategies. It is also helpful to understand the types of existing malware which may pose a serious threat to individuals and industries alike. Considering the advances of the malware industry over the years, these threats will only intensify and increase in complexity.

In this paper, Section 2 presents a brief history of malware and describes the types of existing malware. Section 3 discusses in detail the different types of malware detection methods and machine learning techniques. In Section 4, we present an overview of the existing problems and new challenges faced by programmers and security analysts in malware detection and machine leaning techniques that can help in fighting new kinds of malware. In Section 5, a conclusion highlights the importance of the need of intelligent malware detection techniques and offers a future research direction.

Background Research

Brief History of Malware

The history of malware can be dated back to the 1940's when the scientist John von Neumann explained the concept of a self-replicating program in his book "Theory and Organization of Complicated Automata" and later a paper published by Arthur W. Brooks "Theory of Self-Reproducing Automata" in 1966. However, the earliest known computer virus appeared in the 1970's and went by the name of ‘Creeper Worm'. It was developed by Bob Thomas as a harmless self-duplicating program that displayed a message saying *'I'm the creeper, catch me if you can'*. Its original purpose of design was to test if self-replicating programs were realistically conceivable [6]. There was another virus created in the same decade called 'The Rabbit Virus'. This virus adversely affected the computers by producing numerous copies of itself, causing serious damage to the computer systems. The second most documented computer virus was called the 'Elk Cloner' made in 1982. It spread itself through the only medium of transfer available at that time, the floppy disk. The Elk Cloner virus infected computers with Apple II Operating systems and was created as a prank by a 15-year-old high school student 'Rick Skrenta'. This became the fist known virus to be spread out of the realms of where it was created. Around the same time, another virus known as 'Brain' was introduced, it was the first virus to infect IBM PCs and the first stealth virus that spread all over the world undetected. It was created by two Pakistani brothers, 'Amjad Farooq Alvi' and 'Basit Farooq Alvi' who wanted to save their program from plagiarized copies. It contained a hidden copyright message and the virus overwrote the boot sector of the floppy disk without corrupting any data. At the time when these viruses were created, the scientific community was unaware that this would go on to become the billion-dollar malware industry that it is today and be on of the biggest security threats in the world.

Types of Malware

Although these viruses marked the beginning of Malware, they are not like the ones we encounter today. Nowadays malware is used for a variety of malicious activities including deleting or encrypting files, stealing secret data, extortion, demanding ransom via multiple methods, hacking into core computer files to alter or damage them, spying activity on a

computer, damaging data, etc. These activities can cause severe damages to the machines and can trigger huge financial losses.

Based on the behaviors and purposes of malware, they are divided into a number of distinct, but not mutually exclusive, categories. In order to understand how malware detectors must work to accomplish their tasks, it is important to analyze the domain/scope of the detector. An overview of the taxonomy of malware is given is given below [1] [2] [7].Viruses: This is the most common term used when referring to anything related to malware. Thankfully, viruses now comprise less than a quarter of the entire malware industry. A Virus has the ability to latch itself to another program and execute when the program is run. It makes multiple copies of itself, infecting other files and data to perform a malicious task.

Worms: Worms appeared even before viruses. Worms are the most notorious type of malware as they can spread and replicate themselves from one machine to another without any human intervention. They do not even require latching themselves to another computer program to spread. A Worm relies on the security flaws in the network or internet connections for its propagation. Worms can cause harm to the machines and network, they can disrupt workflows and network traffic etc.

Trojan Horses: Trojan Horses (Trojans) penetrate computers as real programs with malicious code concealed within. They propagate via fake emails and infected websites and scams users into opening them. One of the worst types of Trojans is the 'Anti-malware hoax' that convinces users to run it by showing them that their computer is already infected. This type of malware is not easy to stop as it tricks the users themselves into executing it. Trojans are not a self-replicating malware. They are used to attain privilege access, affect the computer's performance, delete or infect files, etc.

Backdoors: This malware was first devised as a helping tool for hardware developers to check their product after it had been assembled. Based on that, it has become one the most dangerous kind of malware that provide the cyber criminal the root access of the system or network. It can harm its software, hardware, files, data and even other machines linked to the target computer.

Bot and Botnets: A Bot is a computer that has been deliberately infected with a malicious file to create a Botnet. The creation of a bot only takes the carriage of the infected file disguised as a virus, worm, or Trojan, sent via a network or the internet [2]. A Botnet is a collection of infected computers controlled by one person or a team of people to launch a coordinated attack on one target such as the Distributed Denial of Service attacks. Bots are also referred to as zombie computers, since the person who controls the bots can use the machine to carry out any sort of malicious task.

Phishing: Phishing is a method of spreading malware that involves sending spam emails or creating deceptive emails from known and trustworthy sources of the target, such as their workplace, banks, job offers, lottery wins, etc., to attain their personal information. They trick targets into believing the email is legitimate and send malware with its attachments.

Ransomware: These programs encrypt data or lock screens to demand a ransom amount. The process involves encrypting and decrypting data. Sometimes even after paying the ransom,

targets do not get their data back. These programs can appear through Phishing emails or Trojans and get their targets to clicking or downloading data that will infect their computer.

Spyware: Spywares are activity monitoring programs. Spyware attacks can steal sensitive information of the user or gather information about them and send it to other servers. Although they may seem harmless at first, but they work on the same principles as any type of malware and are a red flag for detectors, informing that vulnerabilities in the system exist. Malware such as keyloggers that record every keystroke of the user and keeps track of it is also a form od spyware that can save your data. The hacker only needs another technique foe accessing the computer to get the required information.

Fileless malware: This may not be distinguished as a different type of malware, rather it is a method of malware propagation. They target processes and launch as a sub process or misuse other key OS objects and replicate through that. Since they do not exploit the file system, fileless malware are harder to detect and Remove.

Malvertising: This method of spreading malware uses legitimate advertising companies' websites and links, and places malicious code in them. The code may execute itself with or without any human intervention once the page is open. They are harder to detect as sometimes people actually pay a company for putting up an ad that is basically malware. Malvertising is used to carry and distribute all kinds of malware.

Rootkits: Rootkits are used to conceal the presence of a malicious program in the system, they modify the operating system to hide and avoid detection of malware. Rootkits can also be controlled to gain privileged access or admin access and cause harm to the system.

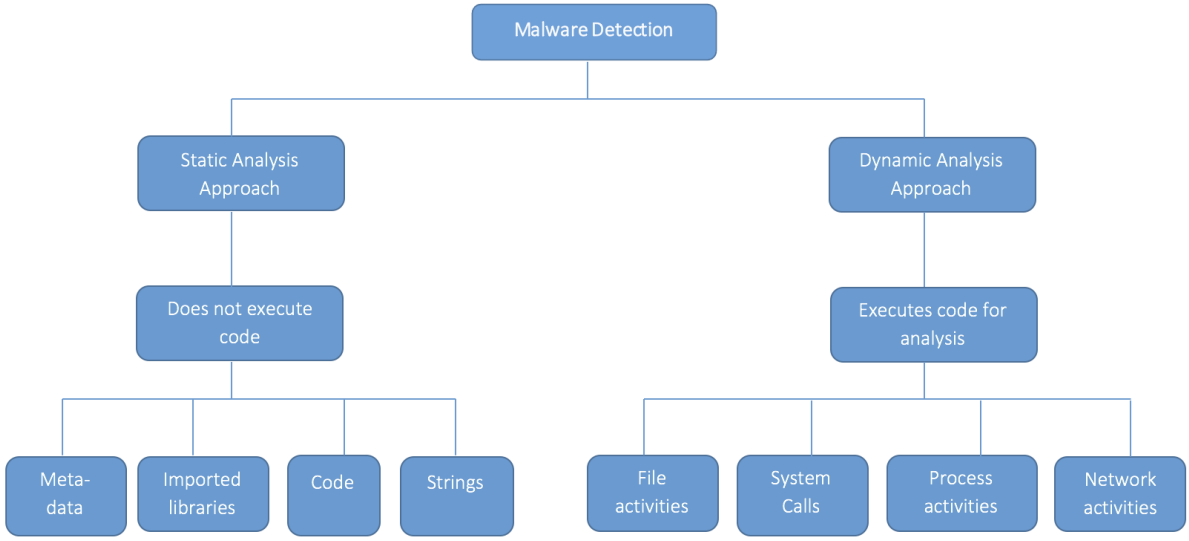

Figure 1: Static and Dynamic Malware Detection Techniques

Malware Detection Techniques

The scope of malware detection is vast and diverse, even though the medium of propagation is nearly the same for all i.e. internet or some form of network. Malware designers are highly dependent on the naivety of their targets to reach their goal. The goal of Anti-malware developers and experts is to maintain the security of the system and make their users aware of

the types of existing threats. The programmers job is to make sure none of these threats get through. Explained below are the kinds of methods or frameworks that are used for detecting malware from the real world data to prevent them from infecting any systems or machines [8] [9]. Figure 1 shows a representation of the Static and Dynamic analysis approaches of malware detection.

Static Analysis Approach

The static analysis approach is the identification process of malware that does not require the execution of the code. They only thoroughly examine the code itself, imported libraries and other binary data properties. There are many softwares available online that help provide this type of analysis. These can be used to locate static features from the file and use them as classification features in machine learning models. The major advantage of this technique is that it provides a very thorough analysis. It is the safest option as the malicious file is not executed, which provides no chance of accidental infection.

Dynamic Analysis Approach

The dynamic analysis approach observes and analyzes the performance, behavior and execution pattern of the code while it is being executed. There is an abundance of dynamic analysis malware detection tools online that extract the dynamic features of the code such as running processes, OS system calls and usage of the file structure, interaction with any network links. Ideally, the code is run in an isolated virtual environment to quarantine its affects. It can also be virtually run on different platforms, softwares and network settings to identify its triggers. These features can help us understand the full functionality and the intended purpose of the code to deduce a correct output from the models. The quality of the output is highly dependent on the quality of the data sample provided to the machine learning models.

Signature-based Detection Methods

Signature-based methods are based on the static analysis technique of malware detection. As suggested by their name, each malicious file or object has its own specific structure and unique signature. The programmers of the anti-malware software can use multiple algorithms to scan the code (without executing it) to detect its nature. When a malicious file is detected, the signature of a new file is added to a database of known signature files. After this process is complete, it is easy to compare other files to the signatures in database and find out if the file is malicious or not. However, if the file appears to be clean, it will still require to be checked via other analysis methods to be sure that it is not a new kind, or some variation of an already existing malicious file.

The advantages of this method are:

- It is easy to understand and execute.
- It Provides fast recognition of malicious files.
- Gives thorough information about the malicious files.
- For machine learning algorithms, it works on pattern recognition which decreases the required time and computing power required for predictive analysis.

The disadvantages of this method are:

- As it is dependent on unique signatures for detection, it fails to recognize even slightest variations from an already recognized malicious file.

- All variations of malicious files need to be saved in the database (increasing file sizes dramatically).
- For machine learning algorithms, it does not support feature selection.

This, in principle, works like any machine learning algorithm training concept. If you look at Figure 2, you will see that the process of training a model in machine learning is similar to the process of this analysis. First, you need a large sample data, with or without class labels that will be given to the algorithms for processing. After the processing is complete, we will have a trained data model that will be able to categorize new files based on what it has learnt. In our detection model, our class labels will be malicious or benign, and data samples will comprise of both file types. The assembly and binary features are two different ways when applying the signature-based approach for detection in machine learning algorithms.

Behavior-based Detection Methods

The behavior-based method is based on the dynamic analysis technique of malware detection. It works by analyzing and observing the behavior of the malicious file by executing the code in a virtual environment or imitating its execution patterns. Executing a program and observing its behavior in run time provides a superior understanding of how the malware is created and deployed. Any sort of anomaly or unconventional behavior can alert the analyst that the file is malicious or at the very least requires further investigation. In behavior-based methods, the assembly features and API calls are the two different ways for detection in machine learning algorithms. Note that the assembly features are also a method in the signature-based detection analysis.

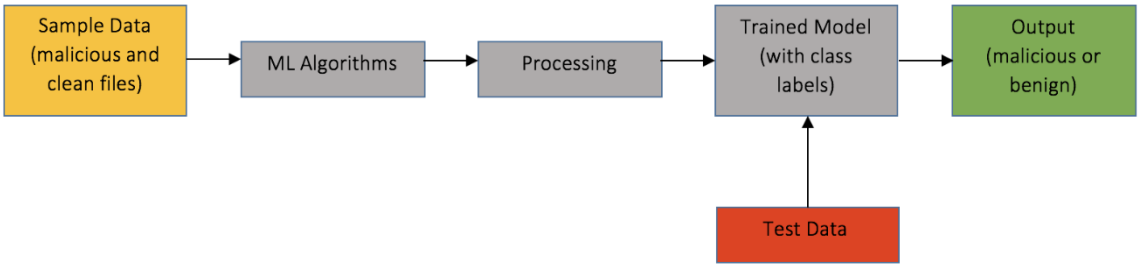

Figure 2: Workflow of Machine Learning detection model

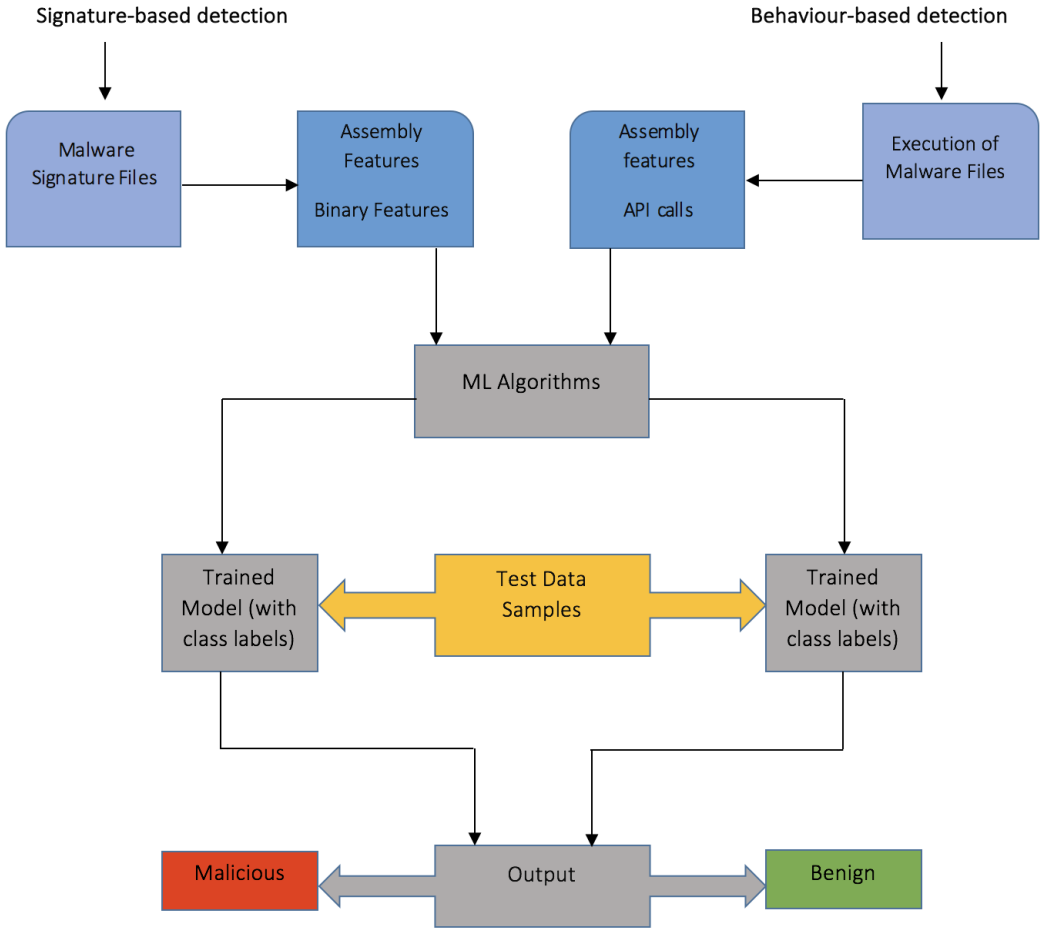

Figure 3: Signature-based and Behavior-based detection frameworks

The advantages of this method are:

- Detecting all variations of a malicious file.
- Observing data flow patterns and deducing the output as malicious or benign.
- Identifying new types of malware.

The disadvantages of this method are:

- The time complexity of this type of analysis is high as it requires executing the codes and analyzing behavioral patterns at runtime.
- The high storage complexity of behavioral patterns.

While this method has its drawbacks, unusual developments in the malware industry, rise of complex techniques for propagating malware and eluding detection have indicated the importance of behavior based techniques coupled with dynamic analysis of malicious files and objects. There are many machine learning models that have been used so far in malware detection. They include supervised and unsupervised algorithms as Support Vector Machines (SVMs), Random Forests (RFs), and Decision Trees (DTs), the Artificial Neural Networks (ANNs) from deep learning, and other meta-heuristic algorithms [10] [11].

Existing Issues and Challenges

Threats of malware have been present since the advent of computer programming and commercial platforms. They are the greatest challenge in cyber security, and will remain so in

the future. After discussing the types of techniques for malware detection, now we will investigate the challenges and problems faced by the security analysts today [12].

Current Malware Threats

Each day there are around 350,000 new, unseen malwares are identified by the existing detection methodologies [3]. Trends in the malware industry come in waves. The year 2017 can be seen as the year ransomware was most common. For e.g. the 'WannaCry' ransomware program was a worm, used to encrypt the target computer's files and then ask for a ransom. Although, security analysts figured out the decryption process, they only managed to recover a part of the data. The worm did not inflict much monetary damage, but the data losses were huge. Another popular form of ransomware is the in-memory ransomware. Some companies store their data in an in house memory before shifting it to a permanent storage server for higher performance and accessibility. The hackers encrypt this valuable data and demand ransoms to release the data. Hackers can even extort companies threatening to leak sensitive data. The impact of damaged or leaked data has far worse implications than the financial losses for companies.

In 2018, there was a surge in botnets and spyware. As bitcoin became more valuable there was also a trend in creating programs that would steal the harnessing power of machines in order to mine crypto-currencies. Normally these programs appeared as Trojans or as third-party softwares that used their target computers with or without explicit authorization. For example, the new trend in tricking consumers is also to write technical and long 'Terms & Conditions' that are often not read, to finish their intended task. This gives the cyber criminals a window to make their activities seem permissible. Some spywares install themselves disguised as cookies for a website.

In 2019 and 2020, Banking sectors have been targeted persistently. As e-commerce and online shopping activities have become mainstream during the CoVid-19 pandemic, it has been much easier for criminals to attain personal information. They inject programs such as Trojans, spyware, spam emails and phishing. There have been reports of people receiving emails from their own banks or WHO, and other health organizations pretending to collect personal data in regards to the pandemic. Most of the programs are designed for the user to enter the data themselves. Then these programs steal data by targeting text and document files on the users' computer, as important information and other credentials are usually stored on a computer in a document file.

Overall, in the next few years, the trend is rising towards ransomware and fileless attacks. These types are most popularly disguised as Trojans. Being the most versatile, fileless malware manipulates the Operating Systems utilities and the pre-installed libraries of softwares to carry out their attacks. Fileless malware has recently been used to inject ransomware, target distributed or decentralized information channels, and create bots. This anticipation stems from the records of previous years, showing an increasing trend in these types of assaults. Figure 4 shows the growth rate of total malware over a period of the last decade (in millions), while the total number of attacks recorded in the past 5 years remain in the bracket of approximately 8-9 billion each year. Figure 5 shows the distribution of these attacks as per the statistics of the past year i.e. 2019.

These techniques are surpassing the current static malware analysis approach. For instance, in a fileless attack, the program does not 'drop a payload'. If there is no executable file, there will be no medium of analysis. Hence, this type of malware needs to be identified and prevented in the wild before it hits the target. Machine learning approaches have the ability to learn from real-time data and all the deep learning algorithms of neural networks are capable of analyzing the data to draw their own patterns of detection. This can be very advantageous when dealing with such a large number of new attacks everyday.

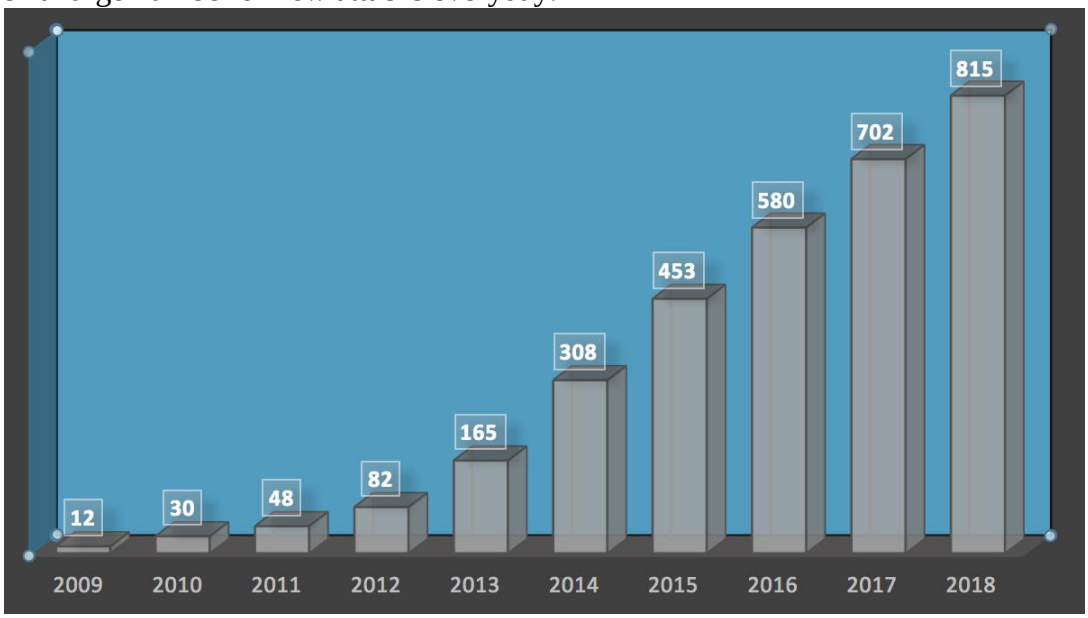

Figure 4: Growth rate of new Malware Threats in millions

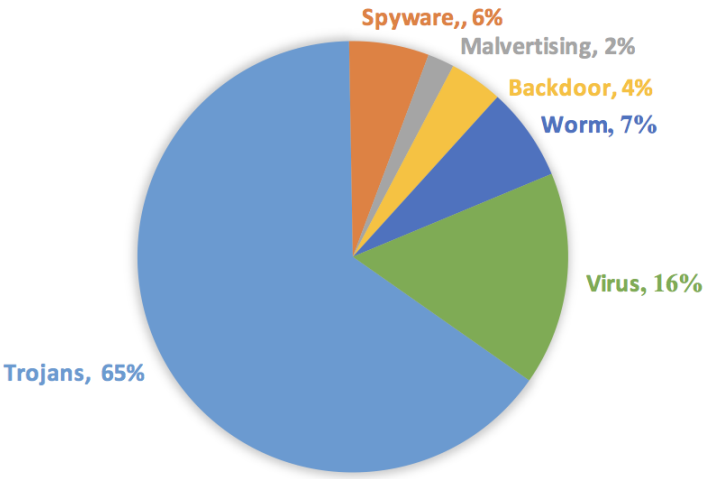

Figure 5: Distribution of different types of Malware Attacks

4.2 New Elusive Mediums of Propagation

Cyber-criminals are increasing in the world day-by-day, these are geniuses hidden behind computer screens wreaking havoc in peoples lives. Recently, there have been advancements in the distribution and targeting strategies of all malware.

A distribution method that has become common is to identify small businesses such as less secure third-party vendors attached to larger businesses and their websites. Cyber-criminals compromise the security of these smaller businesses in order to reach the larger ones. Magecart is an example of this kind of propagation, Magecart is a term used for malware attacks that steal credit/debit card information from e-commerce websites. These attacks specialize in targeting shopping cart systems with e-payments, by using card skimming scripts.

In mobile malware, the most common trend is a self-replicating virus, that sends malicious links with convincing taglines from known numbers. They reach victims via third-party software programs. For eg. when you install an app, it shows dependencies on other apps and forces the user to install it. It may even download itself as an upgrade, or ask the user for payments after the trial period is over. With accessibility to mobile phones growing every day these attacks are increasing day by day. In the last year alone there was a 54% increase in mobile malware [4].

There are new variations of existing malware that can elude traditional static methods. It is impossible to keep all variants of discovered malware in a database for comparison. Dynamic approaches can identify malicious files after analyzing their execution, but the volume is too high. It is not possible to run each file in a sandbox and block all the identified malware. However, with machine learning dynamic approaches can be modified to learn features that make a file malicious. This method can predict malware based on its learning from real-time data and then block these files. A study published by Ihab *et al.* [11] from Jordan, applied the Decision Trees, Random Forests and Naive Bayes (NB) algorithms for binary and multi-class classification of malware on a benchmark dataset. They pointed out the drawbacks of static and signature based methods, and stated that dynamic analysis can be extremely resource intensive when dealing with polymorphic malware. Even their least performance of NB algorithm had an accuracy of 91% for binary and 81.8% for multi-class classification. Liu *et al*. [13], worked on more than 20,000 gray scale images of malicious files. They applied the Shared Nearest Neighbor (SNN) algorithm for identifying variants of malware. Their experimental results show an accuracy of 86.7% successful detection rate of new malware.
The problem has become increasingly severe with the introduction of commercial malware. Expert malware developers can make money by selling their malware. Commercial malware has attracted a more diverse development which is very difficult to analyze. This is a fairly new and unknown research area and a lot of work is being done in this field, but the challenges remain immense.

Theoretical and Real-world Challenges

According to the figures on Google scholar, over the past 5 years, there has been a 95% increase in the amount of papers published related to machine learning techniques in malware detection. This shows that the research community is working enthusiastically to find new unassailable methodologies using the concepts of artificial intelligence. Most of the papers have proposed algorithms with promising results. However, the challenge that we face is that there is a huge gap between the accuracies of these algorithms in the test environment and the real world environment. Some of the detectors face a setback of about a half of the percentage of true positives and false positives are alarmingly high. The largest contributing factor is the variation in data. The data sets used for training are not yet capable of matching the structures that exist in reality.

Additionally, advances in machine learning and deep learning are also empowering malware developers. It has opened up new opportunities for them to enhance malware eluding techniques. Reverse engineering the algorithms of machine learning and understanding their process of detection can assist them in modifying the files to appear as benign. Most common instance is network traffic is disguised to seem like normal internet traffic. An experiment conducted by a group of researchers showed that about 60 percent of the malicious files bypassed as normal files and did not trigger the detectors [14]. In another study, the researchers explained how benign files can be paired with conventional malicious files. A Convolutional Neural Network (CNN) was designed to devise such hybrid files and locate its target i.e. combining WannaCry with an apparent benign file [15]. It was developed in a fashion that it appeared as a harmless file to all detectors, programmed to release the malware only upon reaching its target.

This discussion highlights the importance of development in machine learning based detection models [11]. If cyber criminals succeed in exploiting the concepts of artificial intelligence, they will create a new breed of malware that will easily bypass all traditional detection approaches. The machine learning concepts utilizing the dynamic and behavioral analysis techniques are a promising solution for this new generation of malware. Therefore, it is important to overcome the presented challenges.

The hurdle with machine learning models is the cost of training the algorithms. Malware detection is not as simple as other domains. Malware is evolving drastically and the scope of the problem can change dramatically over a short period of time even a day. This requires either a very robustly trained model or a reinforced learning algorithm model. In reinforced learning, the model is retrained to align its goal of detection each time the dataset is varied. In the domain of malware detection, this could mean every day. Even little variations of existing malware can bypass the already trained detectors if coded intelligently. Consequently, the space complexity, time complexity and cost of training the detectors repeatedly is a very big overhead.

Another barrier is the gap between the expertise of security analysts and machine learning experts. The developers who design the model are aware of the goal, whereas the security analysts understand the complexity of reaching and maintaining that goal. It is important for the analyst to understand the mechanism of the model to interpret its functions. For example, if the model gives a false positive of a benign file, the analyst must be able to comprehend how it happened to fix the problem. So, one problem of reinforcement learning can be fixed by implementing Artificial Neural Networks (ANNs), programming them to self-learn from the changing environment. However, ANNs are quite complex and their internal mechanisms are difficult to interpret. There is a concept associated with ANNs known as the black box model. It means input goes into the network, and output comes out, but it is not humanly possible to trace how the output was generated.

ANNs mechanism works based on an ingenuous strategy. It consists of a layer or infinite amount of layers built up of nodes (neurons). Each node in one layer is connected to all the nodes in the next layer. Every link is associated with a weight. The output of each layer is calculated by the following formula:

$$Y = f\left(\sum_{i=0}^{n}(Wi * Xi) + B\right) \quad (1)$$

Where Y is the output, f is the activation function that is used to introduce non-linearity in the network. Most common activation function are tanh, sin, ReLU (Rectified Linear Unit which is an identity function), sigmoid (recalculates values between 0 and 1), n are the number of layers, W are the weights, X is the input to nodes and B are the biases. Bias function are used to keep the overfitting and underfitting in check. This equation is for the least complicated feed-forward neural networks. It trains itself by reducing the value of the cost function which is the mean of squared errors (MSE).

$$Mean\ Squared\ Error = \frac{1}{n}\sum_{i=1}^{n}(Yi - \hat{Y}i)^2 \quad (2)$$

where n is the number of input values, Y_i are the actual values and $\hat{Y}_i$ are the predicted values of the algorithm. The model is said to be optimally trained when the value of the MSE is the least or nearly zero. The weights and biases are adjusted after each iteration to reach the optimal goal. It also shows how power intensive these algorithms can be. If you imagine this on one hundred hidden layers, it is not possible for humans to trace all mathematical computations done by the computer to reach its final conclusion. These hidden layers can also be in thousands, depending on how much processing the hardware can bear. This is known as the black-box problem in machine learning. This domain of machine learning is a separate study of interest, where experts try to interpret the working of ANNs. When applying these techniques for malware detection, it is difficult to impart this knowledge to the security analysts, as its full functioning is not even wholly understood by the machine learning developers.

CONCLUSION

In this paper, I have reviewed the most common types of malware to understand the mechanisms of the malware industry. The paper gives an insight on the kind of threats that one can anticipate. Then I have given an overview of the types of methods for malware detection and how can they work alongside machine learning algorithms. After discussing these different techniques, a few of the problems and challenges that the security analysts and machine learning experts face have been listed. Although the challenges are present, there are solutions to overcome them. With additional effort, we can soon produce resilient and powerful malware detectors. These detectors are expected to identify polymorphic malware, new methods of injection, discover disguises, reduce spamming, etc. These malware detectors can be expanded to different platforms and technologies like the IoT, big data, cloud computing, social media, and most importantly, the corporate and banking sector.

According to current research trends, it is believed by most scholars that the future of malware detection industry lies in the development of tools and softwares driven by machine learning algorithms [9] [12]. The sheer quantity of papers that are being published to support this idea are an evidence of this theory. In the literature, many machine learning algorithms used for malware detection have shown excellent effectiveness. A gap may exist between real time data

analysis and experimental setup analysis, but with time and effort it can show promising results in the real world as well.

References

[1] 9 types of malware and how to recognize them. (2020, November 17) [Online]. Retrieved from https://www.csoonline.com/article/2615925/security-your-quick-guide-to-malware-types.html

[2] What are malware bots and how to get rid of them? (2019, September 23) [Online]. Retrieved from https://www.thenetworkpro.net/what-are-malware-bots-and-how-to-get-rid-of-them/

[3] AV-TEST. The independent IT-Security Institute, Malware Statistics [Online]. Retrieved from https://www.av-test.org/en/statistics/malware/

[4] 2020 Cyber Security Statistics the ultimate list of Stats, Data & Trends. [Online]. Retrieved from https://purplesec.us/resources/cyber-security-statistics/

[5] The worm has turned: What are the trends in malware today? (2020) [Blogpost]. Retrieved from https://www.kaspersky.com/blog/secure-futures-magazine/malware-trends-2019/28098/

[6] D. Gibert, C. Mateu, and J. Planes, "The rise of Machine Learning for Detection and Classification of Malware: Research Developments, Trends and Challenges", Journal of Network and Computer Applications 153 (2020) 102526.

[7] Y. Ye, T. Li, D. Adjeroh, and S. S. Iyengar, "A Survey on Malware Detection Using Data Mining Techniques", ACM Computing Surveys, Vol. 50, No. 3, Article 41, Publication date: June 2017.

[8] P. HarshaLatha, and R. Mohanasundaram, "Classification of Malware Detection Using Machine Learning Algorithms: A Survey", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 02, FEBRUARY 2020.

[9] A. Souri, and R. Hosseini, "A state-of-the-art survey of malware detection approaches using data mining techniques". Souri and Hosseini *Hum. Cent. Comput. Inf. Sci. (2018) 8:3.* https://doi.org/10.1186/s13673-018-0125-x.

[10] H. E. Merabet, and A. Hajraoui, "A Survey of Malware Detection Techniques based on Machine Learning", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 1, 2019.

[11] I. Shhadat, B. Bataineh, A. Hayajneh, and Z. A. Al-Sharif, "The Use of Machine Learning Techniques to Advance the Detection and Classification of Unknown Malware and Classification of Unknown Malware". International Workshop on Data-Driven Security (DDSW 2020), April 6 - 9, 2020, Warsaw, Poland. Available online at www.sciencedirect.com.

[12] S. Saad, W. Briguglio, and H. Elmiligi, "The Curious Case of Machine Learning In Malware Detection".

[13] L. Liu, B. Wang, B. Yu, and Q. Zhong, "Automatic malware classification and new malware detection using machine learning", *Liu et al. / Front Inform Technol Electron Eng 2017 18(9):1336-1347.*

[14] M. Rigaki, and S. Garcia, "Bringing a GAN to a Knife-Fight: Adapting Malware communication to Avoid Detection". In *2018 IEEE Security and Privacy Work- shops (SPW)*, pages 70–75.

[15] D. Kirat, J. Jiyong, and M. Stoecklin, "Deeplocker - Concealing Targeted attacks with AI Locksmithing".

[16] R. Komatwar, and M. Kokare, "A Survey on Malware Detection and Classification", JOURNAL OF APPLIED SECURITY RESEARCH https://doi.org/10.1080/19361610.2020.1796162

[17] E. B. Karbab, and M. Debbabi, "MalDy: Portable, data-driven malware detection using natural language processing and machine learning techniques on behavioral analysis reports", https://doi.org/10.1016/j.diin.2019.01.017 1742-2876/© 2019 The Author(s). Published by Elsevier Ltd on behalf of DFRWS.

[18] J. Ming, Zhi Xin, P. Lan, D. Wu, P. Liu, and B. Mao, "Impeding behavior-based malware analysis via replacement attacks to malware specifications", Received: 13 September 2015 / Accepted: 11 May 2016 © Springer-Verlag France 2016.