Malware: Detection, Classification and Protection

Dhanashree Paste[1], Trupti Wadkar[2]

[1]Dhanashree Paste, Dept. of Information technology, B. K. Birla College of Kalyan, Maharashtra, India
[2]Trupti Wadkar, Dept. of Information technology, B. K. Birla College of Kalyan, Maharashtra, India

---

Abstract - *Nowadays cyber threats are the most hazardous risks by top management of enterprises and businesses. Exponential growth rate of number of interconnected devices and lack of security awareness maturation of the cyber results in increase of malware. There are incipient families of malware developed and released on day-to-day basis. The existing anti-malware systems like anti-virus and anti-malware packages need to update frequently. There are chances that new malware families are not detected by the old packages because new generation malware is complex and sophisticated which are hard to detect and remove. Hence there is a need for an efficient anti-malware system which will provide defense against malware and viruses. This paper first presents about the malware attacks happened in the last decade and then systematic classification and analysis of malware. Analysis of the malware will help to determine which component of system need to protect and which will further reduce risk in data loss. It also provides an efficient and fast alternate way to detect and prevent malware which are sophisticated. This paper provides a vision to understand more precise and robust techniques for analysis and detection which are timesaving, flexible and gives high accuracy. This will disable attackers to pierce to system and show high-efficiency.*

Key Words: Machine learning, Heuristic, Cloud computing, Deep learning, Artificial intelligence

1.INTRODUCTION

In this digital era, there are lots of software applications available within the internet without spending a dime of cost. There are lots of dangers concerned in downloading those software packages. The chances of getting a malicious software program or applications are very high. The malicious software which got downloaded could be a malware or a virus. Malware is the computer program that infiltrates and damages a computer without the user's consent. (Yuan et al., 2020) As of March 2020, the total number of new Android malware samples amounted to 482,579 per month. According to AV-Test, trojans were the most common type of malware affecting Android devices. In 2019, trojans accounted for 93.93 percent of all malware attacks on Android systems. Ransomware ranked second, with 2.47 percent of Android malware samples involving this variant.

1.1 Evolution of Malware

Malware is a standard cyber-attack that contains malicious content installed by the sender or developer. Scammers have won using various methods to detect malware on as many computers as possible over a long period of time. The first computer virus, called the "Elk Cloner", was the first PC-based malware, known as Brain, to released. In the late 1980s, the most vicious programs were simple boot sections and file infector spread via floppy disk. As technology advances, certain types of malwares proliferate. Macro viruses that exploit Microsoft office products receive distribution through email expansion. In the late 1990's, viruses began to infect domestic users, and email distribution increased. As the increase in the use of utility kits led to the explosion of malware that came online in the 2000s. Since then, the number of malware attacks has apparently doubled or doubled each year. Throughout 2002 and 2003, Internet users suffered from unruly pop ups and other JavaScript bombs. Phishing scams and other credit card cases have also begun this season, with well-known online threats such as blaster and slammer. Slammer was released in 2003, prompted a ban on service attacks from other online administrators and reduced internet traffic. In 2004, a worm war broke out between the writers of MyDoom, Bagle, and Netsky. The discovery and disclosure of various financial scams, 419 Nigerian scams, Sony rootkit led to root fraud, sensitive identity theft, and lottery scams were rampant between 2005 and 2006. By the end of 2007, SQL injection attacks are on rise, the victims include the popular websites such as Cute Overload and IKEA. In June 2008, the Asprox botnet simplified SQL's automatic injection, claiming that Walmart was one of its victims. In early 2009, Gumblar emerged, introducing new Windows operating systems. Its method was quickly adopted by other invaders, which led to botnets being hard to find.

The number of cyber-attacks are increasing vigorously during 21st Century. Some malwares of 21st are mentioned below.

Attack	Year	Description
SQL injection attack	2020	SQL injection attack targeted Flaticon
TJX	2013	40 million card numbers and 70 million personal records stole
Zero-day attack	2019	Attack on micro-soft windows
Ransomware	2017	WannaCry ransomware attack was a worldwide cyberattack
Man in the middle	2017	Equifax withdrew its mobile phone apps following concern about MITM vulnerabilities
XSS attack	2018	The breach affected 380,000 booking transactions
DoS State exhaustion attack	2018	NETSCOUT reported that one of their customers was targeted by a 1.7 Tbps DDoS attack.
DDoS	2020	Amazon Web Services, the 800-pound gorilla of everything cloud computing, was hit by a gigantic DDoS attack

2. MALWARE CLASSIFICATION

Malware is abbreviation for malicious software, cyber criminals design malware to compromise computer functions, steal data, bypass access controls and otherwise cause harm to host computer, its applications or data. Malwares are evolving day by day because new generation of malware is using new techniques to disguise. We classify malware based on delivery method or attack methodology or vulnerability that malware exploits. A basic classification of malware is virus, worm, trojan, spyware, ransomware, scareware, diallerware, bot, bootkit/rootkit, backdoor, downloader. Further classification of virus is file infector, macro-virus, browser hijacker, web scripting, polymorphic, boot sector, multipartite, spacefiller, resident, overwrite. While spyware is classified as adware, keylogger, trackware, cookie, risware, sniffer. Bot is classified as spamware and reverse shell.

3.MALWARE ANALYSIS

Malware analysis helps us to understand behavior and motive of suspicious file. Also provides a fast and accurate approach that reduces your total cost of malware processing while increasing the accuracy of detecting malware. It will Analyze malware, create a mitigation strategy and detect unsigned variants of the same malware. Types of malware analysis are as follows:

3.1 Static Analysis

It is a technique of collecting information about the malicious application without running it. Static Analysis is usually done by performing analysis of binary file from different resources without executing it and studying the components. The binary file can also be disassembled using a disassembler. Static analysis uses a signature-based approach. Malicious software analysis involves several stages such as - • Manual Code Reversing

- file fingerprinting
- virus scanning
- memory dumping
- packer detection
- debugging
- Interactive Behavior Analysis
- Static Properties Analysis
- Fully-Automated Analysis

3.2 Dynamic Analysis

Dynamic analysis uses a behavior-based approach. Dynamic analysis analyzes malware in a sandbox environment to protect other systems from malware. During dynamic analysis the proposed program is actually run. However, this is done in a virtual sandbox environment so that your actual systems remain unaffected and safe. This allows us to detect potential malware and determine if its behavior is actually a malware or not. Powerful analysis is done by looking at the performance of the malware while it is running on the hosting system.

3.3 Hybrid Analysis

Hybrid analysis includes strategies from both approaches to cover each other's shortcomings. Certain actions that can be hidden during startup can be found when downloading binary files or viewing them with a meeting code. Similarly, the available opcode can be displayed when in use, and actions or results can be found live.

4.MALWARE DETECTION TECHNIQUES

Malware detection is the process of scanning malware in your computer/smartphone. If our desktop is infected the we need to detect the malware before this malware can destroy your whole device. Problems during shutting down or restarting, Frequent system crashes or error messages, Emails that send autonomously from your account, Security solution is disabled, Suspicious shortcut files, Battery drains faster than expected, Unexplained data usage, Popup ads start popping up everywhere in browser, etc. this are the basic signs which indicates that your device is infected with malware. Following are the developing techniques of malware detection which are useful for big data such as businesses which are infected by malware.

4..1 Machine Learning

This method has two machine learning aided approaches (classification and clustering) based on app permissions and source code analysis to detect malware on Android devices. The great advantage of these methods is that the use of machine learning tools enables them to detect invisible families of malware with very high precision and recall. The source code-based classification achieved a F-score of 95.1%, while the approach that used permission names only performed with F-measure of 89%. This method provides a way for automated static code analysis and malware detection with high accuracy and reduces the time required for malware analysis of smartphone. However, static analysis with the help of machine learning could help detect new, zero-day malware with relatively high precision and recall. The permission-based method was able to distinguish malware from good material in 89% of cases while the performance of source code analysis classification is more than 95%.(Milosevic et al., 2017)

4.2 Multi-Layer Perceptron

This is another way to detect and differentiate malware using Multi-Layer Perceptron (MLP). MLP is the ANN (Artificial Neural Network) feed phase. MLP contains many layers such as Input Layer, Hidden Layer and Output Layer. MLP is widely used in supervised learning along with backpropogation algorithm for network training. The backpropogation algorithm combines a gradient to find the correct amount of weight to update the network model. The algorithm contains the task of loss function for classification. The output layer uses ReLU as the activation function. Depending on the feature set released using DCV, the number of installation layers is determined. Malware are the binary files. The Binary files are converted to gray scale images. And the total number of output layers has been made to 25 to divide malware images into 25 categories. The converted images are classified using MLP-DCV. The number hidden nodes were selected based on learning performance, loss function and the number of square error values. The MLP is periodically trained. After training the model is used to test images that are not part of the training dataset. The result shows a classification accuracy of 93% when MLP is used with DCV. The Malimg dataset contains malware images used for comparison to study malware detection. MLP-DCV and RBF-SVM network models were used for classification. According to the results, both data models produced 92% classification accuracy. From the result the classification accuracy in stages is 92.63% better when DCV is used with MLP. (Balamurugan, 2021)

4.3 Block-Chain Technology

This approach has demonstrated the formation of a meaningful and effective data exchange on the basis of a blockchain and a community detection framework. To detect false data injection attacks (FDIA) on the MG system, the Hilbert-Huang transform methodology and blockchain-based ledger technology is used to strengthen security on smart DC-MGs by analyzing voltage and current signals on smart sensors and controllers by extracting data signal. FDIA violates the concordance protocols applied to cyber-physical smart DC-MGs. The FDIA detection method was introduced based on Hilbert-Huang's modification to detect malicious attacks in the sensors and controller. This method can detect various FDIA in current voltage and sensors and controller of the converters be defining a threshold. For secure and efficient data sharing, four phases are introduced, involving initialization, identity authentication, signature/verification, and information exchanging phases. The community detection server considers it the key to information exchange layout. In the system, the community detection server detects and analyzes the complete customer database datum, examining the public for cosine similarity. By accessing data exchanged between agents on a smart DC-microgrid, the attacker is unable to break into the systems and make the system more reliable and stable. The simulation results in the test system show very good efficiency and benefit of the proposed method, especially in the presence of cyber-attacks where the information is not available to unauthorized members out of the system. The main reason is that the HAs are converted to any iteration. (Ghiasi et al., 2021)

4.4 Markov Image and Deep Learning This method of byte-level malware classification based on markov images and deep learning is called MDMC. A major step in MDMC is converting malware binaries into markov markers by switching byte transfer probability matrix. Thereafter a deep convolutional neural network is used for the classification of markov images. Tests are performed on two malware datasets, the Microsoft dataset and the Drebin dataset. The average accuracy rates of MDMC are respectively 99.264% and 97.364% in the two datasets. Only malware binaries were used without reverse analysis and dynamic analysis. MDMC can work on various applications such as windows and android. Additional tests with various training dataset and testing datasets also show that MDMC has better performance than GDMC. Because static reverse analysis and dynamic analysis in sequence have their limitations, traditional machine learning algorithms are often difficult to process large unknown anonymous samples of malware. (Yuan et al., 2020)

4.5 Spam E-mail Filtering

This method introduced a secure way to filter out spam and malware in the email system, including standard layers of protocols and policies. An experimental testbed is established to evaluate the effectiveness of the methodology and was tested with spam and malware e-mails. The results showed 95% accuracy, compared to the standard e-mail configuration system. The main purpose and main effect of this approach was to protect email infrastructure from malicious email attacks, phishing e-mail, email scams, and other cyber threats. It also provides the ability to protect the domain from unauthorized use, commonly known as e-mail spoofing, thus providing complete protection of the email server. The results showed better accuracy, compared to the standard email program configuration.(Adiwal et al., 2021)

4.6 Cloud Computing

This method introduced a cloud-based malware detection framework, which uses a hybrid method to detect malware. Cloud computing plays an important role in all aspects of information storage and online services. It brings many benefits beyond the traditional storage and sharing scheme such as easy access, request storage, discount and reduced costs. Utilizing its rapidly evolving technology can bring many benefits to the security of Internet of Things (IoT), Cyber-Physical Systems (CPS) from various types of cyber-attacks, where IoT, CPS provides services to people in their daily lives. The cloud-based detection approach brings more benefits than traditional methods. The cloud environment provides a lot of computational power and very large details of malware detection. It also improves the performance of personal detection equipment, mobile devices and CPS. However, there are other issues on the cloud side such as loss of data control, lack of real-time monitoring, limited use of infrastructure for various clients, and more between the client and the server. (Aslan et al., 2021)

4.7 Smash Method

This is dynamic detection method of malware based on a multi-feature ensemble learning. First, this approach utilizes a combination of software features such as API call sequences for high detection accuracy and low-level hardware features such as resistance to avoid memory dump grayscale and hardware performance tools. Second, it will select a high-quality classifier model to improve the detection of a single feature. Finally, it will set up an integrated learning algorithm with multiple classification detecting malware detection, many features that can explain malware performance from multiple dimensions to improve detection performance. Here is a large dataset of malware sample used for experiments, and the results show that this detection method can get a good detection precision rate, and is better than other recently proposed methods of gaining strength in anti-evasion performance. By improving the detector model for each feature and using a ensemble learning method, malware detection accuracy can be adjusted, and detection accuracy can reach 97.8%. In the experiment, the accuracy of detection decreased by no more than 3%, and the effectiveness of the evasion attack is much better than in other recent studies. (Dai et al., 2019)

4.8 Deep Belief

The approach focuses on developing an efficient computational framework based on Deep Belief Networks for malware detection. This framework merges high level static analysis, dynamic analysis and system calls in feature extraction in order to achieve the highest accuracy. The evaluation compares the most familiar machine learning approaches that were applied in malware detection with this framework. The obtained results demonstrate that Deep Belief Networks technique can realize 99.1% accuracy with the presented dataset. There is a complete static analysis jar which adapts different efficient methods in an attempt to facilitate and speed up the static analysis by handling all the Android applications in only one step rather than considering one application at a time. (Saif et al., 2018)

4.9 Artificial Intelligence

This approach determines the effectiveness of artificial intelligence strategies against cyber security risks, the Researcher has chosen the method of multidisciplinary research and key data. The researcher collected data from employees working in the IT industry. The sample size of this study was 468 and confirmatory factor analysis, discriminant validity, basic analysis of model and lastly, hypothesis assessment was performed. The P-values of all variables were found to be significant except for a professional program that does not have a significant relationship with artificial intelligence and cyber security. Geographical area, sample size, less variables and accessibility were major issues. Altogether, it can be said that the researcher conducted quantitative research on key data collected from staff working in the IT sector of Iraq. In the hypothesis, it was found that there is a significant influence of intelligence personnel and neural networks on AI. Technological advancement has led to an increase in data storage that requires more data security. Model analysis of this article includes an independent variant that is an expert system, neural networks and intelligence agents as well as flexible AI interventions and dependent variables that were cyber security. Complete research results have shown that AI has become one of the main assets of firms to improve their performance in terms of cyber security.

(Alhayani et al., 2021)

4.10 SCIRAS Model

This method introduced a mathematical model to simulate high-level malware. Specifically, it is a worldclass model of SCIRAS (Susceptible- Carrier- Infectious- Recovered- Attacked- Susceptible) in which susceptible, carrier, contagious, acquired and attacked devices are considered. The local and global stability of its equilibrium points are studied and the basic reproductive number is calculated. From the analysis of this epidemiological threshold, the most effective preventive measures are found. This is a fragmented, decisive and global model whose power is based on a standard measurement system.

Consequently, a quality assay for different ratings can be used to study the effectiveness of solutions. In this sense, two types of stable conditions can be achieved: a disease-free steady state in which malware disappears from the network, and a endemic steady state in which there will always be infectious devices. The basic reproductive number is calculated and it is shown that this threshold parameters determines the behaviors of the system depending on whether its numerical value is greater or less than 1. Analysis of this coefficient was performed to find the most effective control methods in which one or two epidemiological coefficients differed. (Guillen et al., 2019)

4.11 Heuristic Detection

This method is performed using an API call network with a heuristic detection method. This is intended to identify the performance of malware that attacks the network. To check for malware can use several environments that work like sandboxes. In this study the Windows operating system is used on VMware which works as a sandbox so that the operating system is not infected. The Windows operating system is used as a target in analysis using the Cuckoo Sandbox. The research results are based on test results from Malware bytes that can detect malwarecausing programs in a heuristic way and detect malware string on show string. Malware with a network API will attack the operating system registration key and have a program that could create spyware or adware that could interfere with the user's work while using a computer device. This method will show tips for protecting computer systems such as using antivirus or antimalware, not installing unauthorized programs, accessing unsafe websites and you do not need to install other unwanted programs when installing the application. For these results, there may also be actions that a user can take to protect his or her computer device.(Suryati and Budiono, 2020)

4.12 Honeypot with Machine Learning This is another way to get malware to use honeypot with machine learning. Honeypot can be used as a trap for suspected packages while machine learning can detect malware by classifying classes. This structure is suggested for detecting malware. The classification in this study uses the Support Vector Machine (SVM) algorithm and the Decision Tree algorithm so that this algorithm produces high accuracy and very effective results. In addition, testing methods have been introduced. The segmentation is determined by 90:10 of each training data and test data to produce the highest accuracy. Verification test is determined by 10 trials. In this test, monitored devices with labeled

datasets are used.(Matin and Rahardjo, 2019)

4.13 Multi-Layered Security

This approach is based on a multi-layered security program for the defence and protection against ransomware. This approach introduces antimalware software to local machines, well-configured firewalls, effective DNS / Web filtering, email security, backups and staff training. With the implementation of these layered defence and protections, the effort can be seized and stopped in many areas in the event of a ransomware attack. If the attack is successful, the layer protection provides the opportunity to retrieve the affected data without paying a ransom amount to hacker. (Pagán and Elleithy, 2021)

5.PROTECTION TECHNIQUES

- Regularly update your operating system and always check your browser's security settings.
- Always backup your files to access the data even if they are ransomed or corrupted.
- Always scan your device and data. It can catch the system vulnerabilities.
- Use unique, complex and different passwords for each account. It will protect your account from hackers. Also change the passwords after a definite period of time.
- Avoid clicking on pop-ups during internet surfing.
- Read the privacy policies, permissions and agreements on the apps. It will help you to know how much personal information is open to developer.
- Don't wait until your whole device gets infected. It's your priority to protect your device.
- You should think twice about browsing or entering your personal information. Your browser may not warn you, but you have to be careful.
- Avoid downloading emails and attachments from unknown sources. Always recheck the actual mail. Sometimes hackers will use friends name or a popular e-mail.
- Be careful when downloading software. Only download programs, movies and music from official websites or services. Most people connect using public Wi-Fi at good restaurants, restaurants or stores, but these insecure networks can leave your phone, tablet or computer infected.
- Using internet is a better option than using a personal hotspot. If you don't have an option about it then use a VPN to better protect your device and data.
- Use good antivirus software which can helps you to protect your desktop, tablets, and other android devices. This will block the malware chain and avoid spread of malware to other devices.

6.CONCLUSION

Day by day Malware is getting challenging. Earlier anti-malware techniques require more detection time, low accuracy, less flexibility and expensive. And they can only detect known malwares. While Hardware based detection suffers from imitation attack. Due to advance packing techniques malware detection is being time consuming and risky. Providing education of protection from malware can be a best solution. It is important to familiarize yourself with today's methods, so that you can better adapt to the ever-changing future. An antimalware which contains proactive exploit protection and prevention system can overcome some issues. While layered anti-malware in which layers having different functions but working simultaneously. Nextgeneration of antivirus software extends to traditional antivirus software. It does this by including machine learning features, behavioral detection, and exploit mitigation. These feature enables Next generation of antivirus to detect malware even when there is no known signature or file hash. In addition, these solutions are often cloud-based, allowing you to deploy tools separately and at scale. This helps ensure that all of your devices are protected and that protection remains active even if devices are affected. These new techniques can reduce time consumption and provide more accuracy. This helps ensure that all of your devices are protected.

REFERENCES

1) Adiwal, S., Gupta, A., Rajendran, B., Bindhumadhava, B.S., 2021. A Secure Methodology for Filtering Spam & Malware in E-mail System and Secure E-mail Testbed Setup. Int. J. 10.

2) Aldwairi, M., Flaifel, Y., Mhaidat, K., 2020. Efficient wu-manber pattern matching hardware for intrusion and malware detection. ArXiv Prepr. ArXiv200300405.

3) Alhayani, B., Mohammed, H.J., Chaloob, I.Z., Ahmed, J.S., 2021. Effectiveness of artificial intelligence techniques against cyber security risks apply of IT industry. Mater. Today Proc. 4) Aslan, Ö., Ozkan-Okay, M., Gupta, D., 2021. A Review of Cloud-Based Malware Detection System: Opportunities, Advances and Challenges. Eur. J. Eng. Technol. Res. 6, 1–8.

4) Balamurugan, P., 2021. An Efficient AntiMalware System With Multi Layer Perceptron And Discriminative Common Vector. Turk. J. Comput. Math. Educ. TURCOMAT 12, 929–937.

5) Dai, Y., Li, H., Qian, Y., Yang, R., Zheng, M., 2019. SMASH: A malware detection method based on multi-feature ensemble learning. IEEE Access 7, 112588–112597.

6) Datta, A., Kumar, K.A., n.d. An Emerging Malware Analysis Techniques and Tools: A Comparative Analysis. Int. J. Eng. Res. 10, 5.

7) Dhalaria, M., Gandotra, E., 2021. Android malware detection techniques: a literature review. Recent Pat. Eng. 15, 225–245.

8) Ghiasi, M., Dehghani, M., Niknam, T., KavousiFard, A., Siano, P., Alhelou, H.H., 2021. Cyberattack detection and cyber-security enhancement in smart DC-microgrid based on blockchain technology and Hilbert Huang transform. IEEE Access 9, 29429–29440.

9) Guillen, J.H., Del Rey, A.M., Casado-Vara, R., 2019. Security countermeasures of a SCIRAS model for advanced malware propagation. IEEE Access 7, 135472–135478.

10) He, D., Chan, S., Guizani, M., 2015. Mobile application security: malware threats and defenses. IEEE Wirel. Commun. 22, 138–144.

11) Kong, F., 2016. Research on Security Technology based on WEB Application.

12) Matin, I.M.M., Rahardjo, B., 2019. Malware detection using honeypot and machine learning, in: 2019 7th International Conference on Cyber and IT Service Management (CITSM). IEEE, pp. 1–4.

13) Milosevic, N., Dehghantanha, A., Choo, K.-K.R., 2017. Machine learning aided Android malware classification. Comput. Electr. Eng. 61, 266–274.

14) Pagán, A., Elleithy, K., 2021. A Multi-Layered Defense Approach to Safeguard Against Ransomware, in: 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, pp. 0942–0947.

15) Saif, D., El-Gokhy, S.M., Sallam, E., 2018. Deep Belief Networks-based framework for malware detection in Android systems. Alex. Eng. J. 57, 4049–4057.

16) Suryati, O.T., Budiono, A., 2020. Impact Analysis of Malware Based on Call Network API with Heuristic Detection Method. Int. J. Adv. Data Inf. Syst. 1, 1–8.

17) 18. Uuganbayar, G., Yautsiukhin, A., Martinelli, F., Massacci, F., 2021. Optimisation of cyber insurance coverage with selection of cost

effective security controls. Comput. Secur. 101, 102121.

18) Verma, R.M., Zeng, V., Faridi, H., 2019. Data quality for security challenges: Case studies of phishing, malware and intrusion detection datasets, in: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. pp. 2605–2607.

19) Yuan, B., Wang, J., Liu, D., Guo, W., Wu, P., Bao, X., 2020. Byte-level malware classification based on markov images and deep learning. Comput. Secur. 92, 101740.