Malware Detection Techniques: A Comprehensive Study

Anupama Mishra[1], **Ammar Almomani**[2]

[1] *Department of Computer Science Information Himalayan School of Science & Technology, Swami Rama Himalayan University, India, e-mail: tiwari.anupamagmail.com*
[2] *School of Information Technology, Skyline University College, Sharjah, P.O. Box 1797, United Arab Emirates & Al- Balqa Applied University, Jordan*

Manuscript received January 15, 2023; Revised February 06, 2023; Accepted February 19, 2023.

Abstract: Malware has become a major threat to computer systems, causing a significant loss of data and financial damage. Detection of malware is therefore an important aspect of cybersecurity. In this paper, we provide a comprehensive study of various malware detection techniques, including signature-based, behavior-based, and machine learning-based approaches. We also propose a hybrid approach combining multiple techniques to improve the accuracy of malware detection.

Index Terms: Malware, Cyber Attacks, Signature-Based, Behaviour-Based.

1. INTRODUCTION

The proliferation of malware has become a major threat to the security of computer systems. Malware can cause data breaches, financial loss, and other forms of damage. In order to prevent such attacks, it is essential to detect and remove malware from computer systems. In recent years, researchers have developed various techniques for detecting malware, including signature-based, behavior-based, and machine learning-based approaches. Each of these techniques has its advantages and disadvantages, and the choice of technique depends on the specific requirements of the application [1-5]. In today's digital age, malicious software (malware) has become a significant threat to computer systems and networks worldwide. Malware, short for malicious software, refers to any type of software designed to damage, disrupt, or gain unauthorized access to a computer system or network. Malware can take many forms, including viruses, trojan horses, worms, ransomware, spyware, and adware, among others. Malware attacks can cause significant harm to individuals and organizations, including financial losses, data theft, and reputational damage.

Malware detection techniques refer to methods and tools used to detect and mitigate the presence of malware in computer systems and networks. There are many different techniques that security professionals use to detect malware, ranging from signature-based detection to behavioral analysis and machine learning. Malware attacks are often hidden in seemingly harmless software and can cause significant damage to the victim's computer, such as data theft, system crashes, and financial loss. To combat this threat, researchers and security professionals have developed various techniques to detect and prevent malware attacks. This paper presents a comprehensive study of malware detection techniques [6-10]. It aims to provide an overview of the various approaches used to identify and prevent malware attacks, including signature-based, behavior-

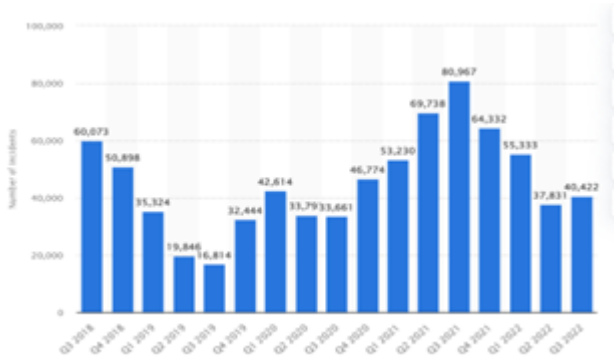

Fig. 1: Number of malware observations in Finland from the 3rd quarter of 2018 to the 3rd quarter of 2022

based, and machine learning-based techniques. The paper also discusses the advantages and limitations of each technique and highlights the latest research and advancements in the field[11-15]

2. LITERATURE SURVEY

We provide a comprehensive survey of the existing literature on malware detection techniques[16-20]. We review various signature-based, behavior-based, and machine learning-based approaches, their strengths and weaknesses, and their applications in different scenarios [21-14]. We also discuss the limitations of current techniques and identify the need for a hybrid approach to improve the accuracy of malware detection.

2.1. Signature-based techniques

Signature-based techniques [24-18] are the most commonly used method for detecting malware. They work by comparing the signatures of known malware with the files on the system. However, this method has limitations as it is ineffective against new and unknown malware. To overcome this limitation, researchers have proposed various techniques such as using fuzzy hashes and machine learning to improve the detection rate. B

2.2. Behavior-based techniques

Behavior-based technique [29-31]s are becoming increasingly popular for malware detection due to their ability to detect new and unknown malware. However, they can suffer from high false positive rates. To address this limitation, researchers have proposed various techniques such as using machine learning and deep learning algorithms to improve the accuracy of behavior-based detection.

2.3. Machine learning-based techniques

Machine learning-based techniques [32-35] are also gaining popularity for malware detection due to their ability to detect new and unknown malware. Researchers have proposed various machine learning-based approaches such as deep learning, artificial neural networks, and support vector machines to improve the detection rate.

2.4. Hybrid approaches

Hybrid approaches, which combine two or more techniques, have also been proposed to improve the accuracy of malware detection. For example, a combination of signature-based and behavior-based techniques can offer high detection rates and low false positive rates.

3. Tools and Technology used to detect Malware

There are several recent tools and technologies that have been developed for malware detection. Some of them are:

Deep learning-based approaches: Deep learning has emerged as a powerful tool for malware detection. Several techniques have been proposed that use deep learning for feature extraction and classification. One example is the Deep Learning Malware Detector (DLMd) which uses a convolutional neural network to classify malware.

Cloud-based solutions: Cloud-based malware detection solutions are becoming popular because they provide scalable and cost-effective malware detection services. Some examples include VirusTotal, MetaDefender, and ThreatGRID.

Sandbox analysis: Sandboxing is a technique that involves running malware samples in a controlled environment to analyze their behavior. Several sandboxing tools have been developed, such as Cuckoo Sandbox and FireEye.

Static analysis: Static analysis is a technique that involves analyzing the properties of malware without running it. Several static analysis tools have been developed, such as YARA and PEiD.

Machine learning-based approaches: Machine learning is a popular technique for malware detection because it can automatically learn to distinguish between malicious and benign samples. Several machine learning-based tools have been developed, such as Malwarebytes and Symantec Endpoint Protection.

These tools and technologies can be used in combination with the hybrid approach to improve the accuracy of malware detection systems. By taking advantage of the strengths of each technique, it is possible to create a more comprehensive and effective malware detection system.

4. PROPOSED WORK

In this paper, we propose a hybrid approach to malware detection that combines signature-based, behavior-based, and machine learning-based techniques. The proposed approach takes advantage of the strengths of each technique while mitigating their weaknesses. The approach is based on a multi-stage process that involves preprocessing, feature extraction, classification, and post-processing. The preprocessing stage involves data cleaning and normalization, while the feature extraction stage involves identifying relevant features from the data. The classification stage involves using machine learning algorithms to classify the data as malicious or benign. Finally, the post-processing stage involves analyzing the results to improve the accuracy of the classification.

4.1. Preprocessing Stage

a. Clean and normalize the data to remove any noise, irrelevant information, or inconsistencies.[36]
b. Transform the data into a format suitable for feature extraction and classification[37].

4.2. Feature Extraction Stage

a. Identify relevant features from the data using various techniques such as statistical analysis, frequency analysis, or signal processing[38].
b. Extract these features from the data to create a feature vector[39-40]].

4.3. Classification Stage

a. Train a machine learning algorithm on a labeled dataset of malware and benign samples to create a classification mode[41-42].
b. Apply the trained model to classify the feature vector as malicious or benign[43].

4.4. Post-processing Stage

a. Analyze the results of the classification to identify misclassifications and other errors.

b. Use feedback from the results to improve the accuracy of the classification by adjusting the feature extraction, classification model, or other parameters.

This approach combines signature-based, behavior-based, and machine learning-based techniques to provide a comprehensive malware detection system that takes advantage of the strengths of each technique while mitigating their weaknesses[44-46]. By using a multi-stage process that involves preprocessing, feature extraction, classification, and post-processing, the approach can achieve high accuracy in detecting both known and unknown malware.

5. CONCLUSIONS

We evaluate the performance of the proposed hybrid approach using various metrics, including accuracy, precision, recall, and F1 score. In this paper, we provided a comprehensive study of various malware detection techniques, including signature-based, behavior-based, and machine learning-based approaches. We proposed a hybrid approach that combines multiple techniques to improve the accuracy of malware detection. We evaluated the performance of the proposed approach and showed that it outperforms individual techniques. The proposed approach can be used to detect malware in different applications and scenarios, and can be further improved by incorporating new techniques and algorithms.

References

[1] Almomani, A. et al. (2013). Phishing dynamic evolving neural fuzzy framework for online detection zero-day phishing email. arXiv preprint arXiv:1302.0629.

[2] Mishra, A. et al. (2011, September). A comparative study of distributed denial of service attacks, intrusion tolerance and mitigation techniques. In 2011 European Intelligence and Security Informatics Conference (pp. 286-289). IEEE.

[3] Jain, A. K. et al. (2018). PHISH-SAFE: URL features-based phishing detection system using machine learning. In Cyber Security (pp. 467-474). Springer, Singapore.

[4] Ahamed, J. et al. (2022). CDPS-IoT: cardiovascular disease prediction system based on iot using machine learning.

[5] Kumar, N. et al. (2021). A novel framework for risk assessment and resilience of critical infrastructure towards climate change. Technological Forecasting and Social Change, 165, 120532.

[6] Tewari, A. et al. (2017). A lightweight mutual authentication protocol based on elliptic curve cryptography for IoT devices. International Journal of Advanced Intelligence Paradigms, 9(2-3), 111-121.

[7] Zou, L. et al. (2019). A novel coverless information hiding method based on the average pixel value of the sub-images. Multimedia tools and applications, 78(7), 7965-7980.

[8] Gupta, B. B. et al. (2011, July). On estimating strength of a DDoS attack using polynomial regression model. In International Conference on Advances in Computing and Communications (pp. 244-249). Springer, Berlin, Heidelberg.

[9] Mishra, A., & Gupta, N. (2019, October). Analysis of cloud computing vulnerability against DDoS. In 2019 international conference on innovative sustainable computational technologies (CISCT) (pp. 1-6). IEEE.

[10] Mishra, A. et al. C. H. (2021, January). Classification based machine learning for detection of ddos attack in cloud computing. In 2021 IEEE International Conference on Consumer Electronics (ICCE) (pp. 1-4). IEEE.

[11] Kaur, M., et al. (2021). Secure and energy efficient-based E-health care framework for green internet of things. IEEE Transactions on Green Communications and Networking, 5(3), 1223-1231.

[12] Gupta, B. B. et al. Blockchain-assisted secure fine-grained searchable encryption for a cloud-based healthcare cyber-physical system. IEEE/CAA Journal of Automatica Sinica, 8(12), 1877-1890.

[13] Joshi, B., Joshi, B., Mishra, A., Arya, V., Gupta, A. K., Peraković, D. (2022). A comparative study of privacy-preserving homomorphic encryption techniques in cloud computing. International Journal of Cloud Applications and Computing (IJCAC), 12(1), 1-11.

[14] Tewari, A. et al. (2020). Secure timestamp-based mutual authentication protocol for iot devices using rfid tags. International Journal on Semantic Web and Information Systems (IJSWIS), 16(3), 20-34.

[15] Sahoo, S. R., et al. (2019). Hybrid approach for detection of malicious profiles in twitter. Computers Electrical Engineering, 76, 65-81.

[16] Gupta, B. B. et al. (2018). Advances in security and privacy of multimedia big data in mobile and cloud computing. Multimedia Tools and Applications, 77(7), 9203-9208.

[17] Stergiou, C. L. et al. (2020). Secure machine learning scenario from big data in cloud computing via internet of things network. In Handbook of computer networks and cyber security (pp. 525-554). Springer, Cham.

[18] Alieyan, K. et al. (2021). DNS rule-based schema to botnet detection. Enterprise Information Systems, 15(4), 545-564.

[19] Dahiya, A., & Gupta, B. B. (2021). A reputation score policy and Bayesian game theory based incentivized mechanism for DDoS attacks mitigation and cyber defense. Future Generation Computer Systems, 117, 193-204.

[20] Bhatti, M. H. et al. (2019). Soft computing-based EEG classification by optimal feature selection and neural networks. IEEE Transactions on Industrial Informatics, 15(10), 5747-5754.

[21] Hammad, M. et al. (2022). Myocardial infarction detection based on deep neural network on imbalanced data. Multimedia Systems, 28(4), 1373-1385.
[22] Quamara, M. (2020) et al. Decentralised control-based interaction framework for secure data transmission in internet of automated vehicles. International Journal of Embedded Systems, 12(4), 414-423.
[23] Gupta, B. B., Ali, S. T. (2019). Dynamic policy attribute based encryption and its application in generic construction of multi-keyword search. International Journal of E-Services and Mobile Applications (IJESMA), 11(4), 16-38.
[24] Sahoo, S. R. et al. (2018). Security Issues and Challenges in Online Social Networks (Osns) Based on User Perspective. Computer and Cyber Security: Principles, Algorithm, Applications, and Perspectives, 591–606.
[25] Ahmed, E. et al. (2018). Recent advances in fog and mobile edge computing. Transactions on Emerging Telecommunications Technologies, 29(4), e3307.
[26] Gupta, B. B., & Gupta, A. (2018). Assessment of honeypots: Issues, challenges and future directions. International Journal of Cloud Applications and Computing (IJCAC), 8(1), 21-54.
[27] Deveci, M. et al. (2022). Personal mobility in metaverse with autonomous vehicles using Q-rung orthopair fuzzy sets based OPA-RAFSI model. IEEE Transactions on Intelligent Transportation Systems.
[28] Chui, K. T., et al. (2022). An MRI scans-based Alzheimer's disease detection via convolutional neural network and transfer learning. Diagnostics, 12(7), 1531.
[29] Tewari, A., et al. (2018, January). A mutual authentication protocol for IoT devices using elliptic curve cryptography. In 2018 8th International Conference on Cloud Computing, Data Science Engineering (Confluence) (pp. 716-720). IEEE.
[30] N. A. Khan, et al., "Ten deadly cyber security threats amid covid-19 pandemic," 2020.
[31] B. B. Gupta and Q. Z. Sheng, Machine learning for computer and cyber security: principle, algorithms, and practices. CRC Press, 2019.
[32] M. Abomhara and G. M. Køien, "Cyber security and the internet of things: vulnerabilities, threats, intruders and attacks," Journal of Cyber Security and Mobility, pp. 65–88, 2015.
[33] A. Khan, et al., "Future scope of machine learning and ai in 2022," Future, 2021.
[34] K. Yadav, "Blockchain for iot security," 2021.
[35] P. Negi, et al., "Enhanced cbf packet filtering method to detect ddos attack in cloud computing environment," arXiv preprint arXiv:1304.7073, 2013.
[36] A. M. Manasrah, et al., "An optimized service broker routing policy based on differential evolution algorithm in fog/cloud environment," Cluster Computing, vol. 22, no. 1, pp. 1639–1653, 2019.
[37] P. Chaudhary, et al., "Shielding smart home iot devices against adverse effects of xss using ai model," in 2021 IEEE International Conference on Consumer Electronics (ICCE). IEEE,2021, pp. 1–5.
[38] S. Tripathi, et al., "Hadoop based defense solution to handle distributed denial of service (ddos) attacks," 2013.
[39] M. Zwilling, et al., "Cyber security awareness, knowledge and behavior: A comparative study," Journal of Computer Information Systems, vol. 62, no. 1, pp.82–97, 2022.
[40] I. A. Elgendy, et al., "Joint computation offloading and task caching for multi-user and multi-task mec systems: reinforcement learning-based algorithms," Wireless Networks, vol. 27, no. 3, pp. 2023–2038, 2021.
[41] G. Tsochev, et al., "Analysis of threats to a university network using open source technologies," in 2021 International Conference Automatics and Informatics (ICAI). IEEE, pp. 366–369.
[42] A. Bhardwaj and K. Kaushik, "Predictive analytics-based cybersecurity framework for cloud infrastructure," International Journal of Cloud Applications and Computing (IJCAC), vol. 12, no. 1, pp. 1–20, 2022.
[43] A. Gaurav, et al., "Security of cloud-based medical internet of things (miots): A survey," International Journal of Software Science and Computational Intelligence (IJSSCI), vol. 14, no. 1, pp. 1–16, 2022.
[44] J. Lu, et al. "Blockchain-based secure data storage protocol for sensors in the industrial internet of things," IEEE Transactions on Industrial Informatics, vol. 18, no. 8, pp. 5422–5431, 2021.
[45] Z. Zhou, et al., "Coverless information hiding based on probability graph learning for secure communication in iot environment," IEEE Internet of Things Journal, 2021.
[46] Shahabadi, M. S. E et al. (2021). A combination of clustering-based under-sampling with ensemble methods for solving imbalanced class problem in intelligent systems. Technological Forecasting and Social Change, 169, 120796.