How to Cite:
Reddy, K. N., Hasan, N., Abdullah, B., Mustaqeem, M., & Ara, T. (2022). An intelligent robust malware detection by implementing deep learning. *International Journal of Health Sciences*, *6*(S6), 2835–2843. https://doi.org/10.53730/ijhs.v6nS6.9818

An intelligent robust malware detection by implementing deep learning

Dr. K. Nagi Reddy
Professor, Department of Information Technology, Lords Institute of Engineering & Technology, Jawaharlal Nehru Technological University, Hyderabad (JNTUH)

Neha Hasan
Assistant Professor, Department of Information Technology, Lords Institute of Engineering & Technology, Jawaharlal Nehru Technological University, Hyderabad (JNTUH)

Bashar Abdullah
B.Tech Student, Department of Information technology, Lords Institute of Engineering &Technology, Jawaharlal Nehru Technological University, Hyderabad (JNTUH)

Mohd Mustaqeem
B.Tech Student, Department of Information technology, Lords Institute of Engineering &Technology, Jawaharlal Nehru Technological University, Hyderabad (JNTUH)

Tabassum Ara
B.Tech Student, Department of Information technology, Lords Institute of Engineering &Technology, Jawaharlal Nehru Technological University, Hyderabad (JNTUH)

***Abstract*---**Malicious software (ransom ware) cyber-attacks in frequency and severity, posing an increasingly serious threat to computer systems everywhere. Malware detection is a hot study area as several multiple computers, organisations, and governments have been affected by an exponential rise in malware attacks. Dynamic and static assessment of malicious characteristics and behaviour patterns is time expensive and useless in real-time malware detection, according to current technologies. It is becoming increasingly common for malicious apps to use polymorphic and adaptive techniques to rapidly modify their behaviour and develop a number of new malicious apps. In order to undertake an effective malware analysis, machine learning techniques (MLAs) are increasingly being used to create new malware varieties. This approach is time-consuming since it requires considerable feature engineering, learning and representation of features. Moreover the feature extraction process could be effectively

eliminated by using advanced MLAs like deep learning. These methods have been shown to perform better with a biased training dataset, which restricts their practical application in real-time scenarios. A new improved approach for successful zero-day malware detection must be developed in order to eliminate biases and analyze these approaches autonomously. In order to close the knowledge gap, this research compares and contrasts several commercial and government datasets using traditional MLAs and deep learning structures for detecting attacks, categorization, and segmentation. In the second step, we remove all dataset bias from the experimental evaluation by employing alternative splits of the public and private databases to training and validate the proposed at different timescales. A unique step in image analysis with optimal settings for MLAs and deep learning architectures is the third key contribution we have made. Deep learning architectures presented by our model outperform classical MLAs in a comprehensive comparison evaluation. First-of-its-kind in a big data environment, our innovation in the combination of visualisation and deep learning structures for static and dynamic hybrid approaches and needs to achieve hybrid approaches results in robust intelligence zero-day malware detection. With this paper's findings in mind, researchers can now move on with developing a deep learning approach that can identify malware in real time using a visual representation of the code.

***Keywords*---**Malware detection (static/dynamic), artificial intelligence.

Introduction

The rapid growth of technology has changed the daily activities of enterprises and individuals alike in this digitized field of Manufacturing 4.0. Apps and IoT had paved the way for modern conception of an informational economy being developed. In addition, cyber thieves target individual PCs and networks to steal personal information on the financial profit and disrupt systems, posing a serious threat to the advantages of this industrialization. Attackers like this one utilise computer viruses or virus to put information at risk [1]. Operation system-inflicting computer software is known as infection (OS). Based on the malware's aim and behaviour, it is given several titles, including malware (e.g. "adware,""spyware,""virus," and "worm"). Intrusion detection system and remediation is an ever-evolving issue in the world of computer security. The ability of malware authors to elude detection improves when new strategies are developed by researchers.

Research Background

When the Malware initially appeared as a malware attack in 1988-89, security software packages are developed to examine a matching with the virus signature databases that was updated regularly. A heuristics searching could be used in conjunction with biometrics detecting attacks to discover malware's activity. As a result, such traditional methodologies would be unable to identify lowest infection

since new strains of malware evade detection by employing approaches such as obfuscation techniques. [2]. In order to reverse engineer malware utilising Dynamic and Static research and assign a signature to it, a cryptography threat detection system need deep domain expertise. The decryption of viruses in a cryptography system takes longer, and an attacker could sneak into the system during that period. Additionally, signature-based systems are unable to detect emerging forms of virus. Scientists in the field of cyber security have discovered that hackers frequently employ polymorphism and metamorphic rocks to evade signature-based identification. Manual unpacking and analysis of API calls are carried out using software tools as a solution to this issue because this is a time-consuming procedure, [3] devised a four-step methodology for extracting API calls and identifying harmful traits. The malware gets unloaded in the first phase. Secondly, the binary application is broken down and analysed. Extracting API calls is the third step. Statistical feature analysis and API call mapping are part of Step 4 of the process. Improved [4] by employing a 5-step technique that includes machine learning algorithms (MLAs) like SVM with n-gram features acquired from huge samples of both normal and malicious executable code with 10-fold cross validation. [5] Later, a comparison of various machine learning algorithms for detecting attacks was carried out, and a framework for zero-day malware detection was presented. Methods of similarity-based analysis and machine learning are used to tackle harmful code variants depending on API call sequencing and periodicity [7]. A unified approach for extracting the features from virus binary was proposed during the extensive experimental study on a very large data set. When presenting novel classification algorithm techniques for improved detecting attacks, [8] made use of API calls and an SVM/MRMRF heuristics hybrid. This allowed for unique feature extraction techniques for increased malware identification to be presented. Several researchers have been working on enhancing malware detection MLAs in response to the recent rise in attacks using unknown virus [6] and other obfuscation virus. An important part of this study's inspiration comes from such fact.

Need for the Study

The approaches of feature engineering, feature selection, and feature representation are all essential to MLAs. A model is trained using a set of features and a corresponding class to build a dividing line in between benign and the malicious. This dividing plane aids in the detection and classification of a certain type of malware. Knowledge of a specific topic is required for both feature development and characteristic selection techniques. Dynamic and Static analysis can be used to acquire the various features. Without executing the binary programme, static analysis gathers data from it. Malware activity in an isolated manner can be monitored in real time using a technique called dynamic analysis. [10] Goes into great length about the challenges and complexities of Dynamic analysis. For a long-term solution to malware detection, dynamic analysis may be the best option. For real-time malware detection, the Dynamic model cannot be used since it takes too long to examine its behaviour, which means that dangerous payloads can be delivered.

When opposed to statically obtained data, malware detection systems based on evolutionary analytics are more resistant to obfuscation. Static and Dynamic

analysis methodologies are the most prevalent approaches used by commercial anti-malware solutions on the market. Machine learning-based malware detection systems rely on feature engineering, feature learning, and feature representation approaches that require substantial domain-level knowledge [11–13] to work. As a result, the malware detectors could be easily avoided if an attacker learns the features. In order to be effective, MLAs require a wide range of virus patterns in their data. The amount of publicly available baseline data for infection analyses is quite limited due to security and privacy considerations. There are only a handful datasets, and every one of them has its own set of [14] scathing comments because many of them are out-of date. Many of the reported results of computational intelligence malware analysis have been based on datasets provided by the researchers themselves. For research purposes, it's difficult to compile a comprehensive malware collection from publicly available sources.

A general machine learning-based malware analysis technique that could be implemented in real time has a number of limitations. More crucially, [15] examined in depth the compelling controversies affecting the application of data science approaches. A new type of neural networks known as deep learning, and that is an upgraded version of MLAs, has recently surpassed conventional MLAs on many tasks in NLP, machine vision, voice recognition as well as many other areas [14]. Training involves capturing relatively high characteristics in deep buried layers and then learning from them by making a mistake. New systems are discovered and linkages are made with previously discovered categories in machine learning. This allows for better executive functions, as MLA outputs decrease as the amount of data increases. Deep learning structures for malware classification have been the subject of only a few publications [13], [11], [12]. Industry 4.0 has led to an increase in the amount of malwares in recent years, though. For a real-time collection of malware, conventional systems are not extensible, requiring large amounts of store space and time to make effective decisions. In order to address the lack of scalability as well as decentralized infrastructures in malware classification, the recent research investigated the techniques and developed ScaleMalNet, a scalable structure.

Related Work

Big data for cyber security: Trends in vulnerability disclosure and dependency

Modern organisations' complex Big Data platforms are increasingly becoming attack targets for existing and developing threat agents. Elaborate and specialised assaults will become more common in exploiting weaknesses and flaws. With the growing trend of cybercriminals and occurrences caused by these vulnerabilities, proper vulnerabilities planning is fundamental for modern large organizations. Organizations, on the other hand, struggle to handle the enormous volume of vulnerability detected on their networks. Furthermore, vulnerability management is more reactive in actuality. Thorough predictive methods predicting the expected volume and dependency of vulnerabilities disclosure would surely provide valuable insights towards corporations and assist them in being more proactively in the monitoring of cyber threats. Our proposed fresh and rigorous paradigm enables this new potential by exploiting the extensive yet working in multiple

vulnerability data. Using this solid framework, we launched characteristics of the study on not only dealing with persistent volatility in data but also uncovering multivariate dependent structure among distinct vulnerability concerns. In contrast to previous research on dependent variable from several independent, we look at the broader multivariate situation, attempting to capture their fascinating linkages. We have demonstrated that a composite model may successfully identify and preserve long-term reliance between multiple vulnerabilities and exploit disclosures through comprehensive empirical investigations utilising real-world vulnerability data. Furthermore, the report sets the path for future research on the stochastic aspect of vulnerability proliferation in order to develop more accurate measurements for overall improved cyber risk management.

The issue of obfuscated malware in cybercrime

Malware is a major security concern to computers that has been around since the dawn of the computer age and has grown exponentially in recent years. Such malware risks are poorly understood, and there is a paucity of information on how to avoid and detect them. Obfuscated malware, which is becoming more common and more sophisticated, is the main focus of this article, which investigates the various strategies used by obfuscated malware. In particular, our research illustrates how file system weaknesses are exploited by fraudsters to infiltrate the system with concealed malware. Additionally, the paper discusses recent Zeus botnet developments and the critical role anomaly detection plays in combating the latest Zeus generation of malicious software.

Methodology

Static and dynamic analysis of transaction information is now being used to detect cyber-attacks in the modern day. Static research focuses on a signature to determine if a packet is normal or if it includes an attacker could exploit this vulnerability by comparing it to an existing attack signatures. In applied to measure malware/attacks, dynamic examination utilized the continuous execution flow[9]. However, simulation tool is night before going to bed. For this challenge, the author is using machine learning techniques, such as SVM Algorithm, Random Forest, Decision Tree, Naive Bayes and Logistic Regression, K-Nearest Neighbours and Deep Learning such as Convolution Neural Networks (CNN) and LSTM, to increase detection capability with old and new malware and viruses (Long Short Term Memory) [11]. In all algorithms, the performance of CNN and LSTM is superior.

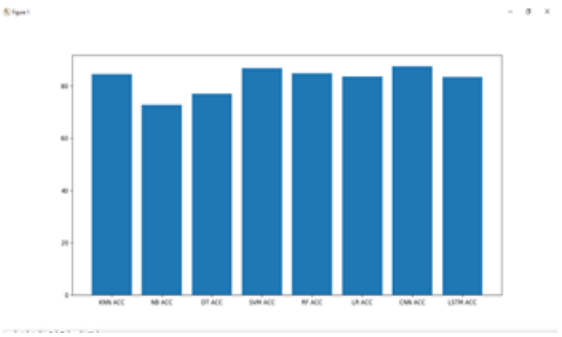
The author of this article is utilising a binaries infection database named 'MALIMG' to analyse the effectiveness of the training methods in this research. Using this binary information, the system would construct development / certification concepts for machine learning techniques by converting it into grey photos. To distinguish them from other techniques like EMBER and MalConv LSTM, these methods convert binary input into pictures before creating models. 80% of the dataset was utilised to train a model, while the remaining 20% was used for testing. In order to predict malware class, we must send new test malicious binaries information to the system whenever testing information is recorded.

Result and Discussion

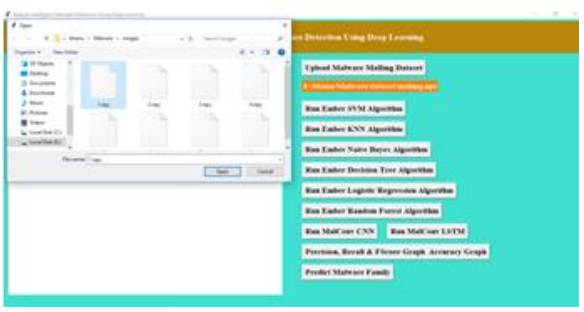
Run the project to get below screen

In above result click on 'Upload Malware MalImg dataset' button to upload dataset

In order to read malware dataset and create train and test model then apply SVM method to calculate its prediction performance, FSCORE, Precision and Recall, click on Run everything above algorithm' tab now. The accuracy, precision, and recall values would be closer to 100 if the algorithm performs well.



Click the "Anticipate Malware Family" tab and download a binary code to acquire or predict the type of malware you're dealing with now



A binary file named 1.npy is uploaded in the graph above. Its malware risk is forecasted to be as follows

As can be seen in the above result, the submitted test document includes a malware attack known as Dialer-a-Dialer, you can also add other files and use them to make predictions about your class.

Conclusion

In order to detect, classify, and categorise zero-day virus, this article examined classical machine learning techniques (MLAs) and computational intelligence architectures depending on simulation process, computational fluid dynamics, and image processing approaches. It also built ScaleMalNet, a scalability

infrastructure. The ransom ware gathered from end-user hosts is analysed in two stages using this framework, which employs deep convolutional neural network. First, spyware were classified using a combination of finite element modelling. Using image processing algorithms, browser hijackers were categorised into their respective groups in the second stage. This study found that deep continuing to learn techniques beat classical MLAs on both currently accessible benchmark problems and individually obtained datasets using a variety of scientific analysis. By adding a several more layers to performance to ensure, the developed system is capable of judging huge numbers of malicious in genuine and can be expanded out to analyse even bigger numbers of malware. Future research work is aimed at these variances with software additions that could be updated with new dataset. It is possible to sum up the findings of this study as follows:

Malimg's collection has a large number of malware categories that are disproportionately represented. A cost-conscious method can be used to address the problem of binary classification ransomware families. This makes it easier to incorporate the cost components into deep learning infrastructures' based on potential learning method. As a primary cost item, classifications relevance gives lower value for classes with a big number of tests and higher value for classes with a smaller number of examples. In an adverse context, deep learning architectures are susceptible. The generative adversarial network (GAN) approach can readily deceive recurrent neural networks throughout development or deployment. The resilience of the deep learning approaches is not addressed in the proposed method. Considering malware dispersion is a solution is prepared in a protection system, this is an essential path for future work. Misclassification can have a significant impact on an organization's bottom line.

References

1. R. Anderson et al., “Measuring the cost of cybercrime,” in The Economics of Information Security and Privacy. Berlin, Germany: Springer, 2013, pp. 265–300.
2. B. Li, K. Roundy, C. Gates, and Y. Vorobeychik, “Large-scale identification of malicious singleton files,” in Proc. 7th ACM Conf. Data Appl. Secur. Privacy. New York, NY, USA: ACM, Mar. 2017, pp. 227–238.
3. M. Alazab, S. Venkataraman, and P. Watters, “Towards understanding malware behaviour by the extraction of API calls,” in Proc. 2nd Cybercrime Trustworthy Comput. Workshop, Jul. 2010, pp. 52–59.
4. M. Tang, M. Alazab, and Y. Luo, “Big data for cybersecurity: Vulnerability disclosure trends and dependencies,” IEEE Trans. Big Data, to be published.
5. M. Alazab, S. Venkatraman, P. Watters, and M. Alazab, “Zero-day malware detection based on supervised learning algorithms of API call signatures,” in Proc. 9th Australas. Data Mining Conf., vol. 121. Ballarat, Australia: Australian Computer Society, Dec. 2011, pp. 171–182.
6. M. Alazab, S. Venkatraman, P. Watters, M. Alazab, and A. Alazab, “Cybercrime: The case of obfuscated malware,” in Global Security, Safety and Sustainability & e-Democracy (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering), vol. 99, C. K. Georgiadis, H. Jahankhani, E. Pimenidis, R. Bashroush, and A. Al-Nemrat, Eds. Berlin, Germany: Springer, 2012.

7. M. Alazab, “Profiling and classifying the behavior of malicious codes,” J. Syst. Softw., vol. 100, pp. 91–102, Feb. 2015.
8. S. Huda, J. Abawajy, M. Alazab, M. Abdollalihian, R. Islam, and J. Yearwood, “Hybrids of support vector machine wrapper and filter based framework for malware detection,” Future Gener. Comput. Syst., vol. 55, pp. 376–390, Feb. 2016.
9. [9] E. Raff, J. Sylvester, and C. Nicholas, “Learning the PE header, malware detection with minimal domain knowledge,” in Proc. 10th ACM Workshop Artif. Intell. Secur. New York, NY, USA: ACM, Nov. 2017, pp. 121–132.
10. C. Rossow, et al., “Prudent practices for designing malware experiments: Status quo and outlook,” in Proc. IEEE Symp. Secur. Privacy (SP), Mar. 2012, pp. 65–79.
11. E. Raff, J. Barker, J. Sylvester, R. Brandon, B. Catanzaro, and C. Nicholas. (2017). “Malware detection by eating a whole exe.” [Online]. Available: https://arxiv.org/abs/1710.09435
12. M. Krcál, O. Švec, M. Bálek, and O. Jašek. (2018). Deep Convolutional Malware Classifiers Can Learn from Raw Executables and Labels Only. [Online]. Available:https://openreview.net/forum?id=HkHrmM1PM
13. M. Rhode, P. Burnap, and K. Jones, “Early-stage malware prediction using recurrent neural networks,” Comput. Secur., vol. 77, pp. 578–594, Aug. 2018.
14. Thammana Ajay, K Nagi Reddy, Dasari Anantha Reddy, Pattem Sampath Kumar, K Saikumar, “Analysis on SAR Signal Processing for High-Performance Flexible System Design using Signal Processing” in proceedings of 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA) held during 02-04 December 2021, pp.30-34.
15. V Jothsna, Ibrahim Patel, K Raghu, P Jahnavi, K Nagi Reddy, K Saikumar, “A Fuzzy Expert System for The Drowsiness Detection from Blink Characteristics”, in proceedings of 7th International Conference on Advanced Computing and Communication Systems (ICACCS) held during 19-20 March 2021, pp.1976-1981.
16. Gustiani, R., Hakimi, M., & Suryaningsih, E. K. (2022). The implementation of referral system in postpartum hemorrhage cases by midwife. International Journal of Health & Medical Sciences, 5(3), 211-220. https://doi.org/10.21744/ijhms.v5n3.1917
17. Suryasa, I. W., Rodríguez-Gámez, M., & Koldoris, T. (2021). Get vaccinated when it is your turn and follow the local guidelines. *International Journal of Health Sciences*, *5*(3), x-xv. https://doi.org/10.53730/ijhs.v5n3.2938
18. Suryasa, I. W., Rodríguez-Gámez, M., & Koldoris, T. (2021). Health and treatment of diabetes mellitus. *International Journal of Health Sciences*, *5*(1), i-v. https://doi.org/10.53730/ijhs.v5n1.2864