International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0334
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2307-2311

Digital Fortress - Web Application Malware Detection

P. V. Kishore Kumar[1], K. vamsi[2], J. manasa[3], S. D V Swaroop[4], V. gayatri[5]
[1,2,3,4,5]Centre for Advanced Studies, RCE, Eluru, Andhra Pradesh, India.
Emails: *kishore.rce@gmail.com[1], vamsichowdarykuchipudi@gmail.com[2], jmanasa663@gmail.com[3], Sunkara777.ss56@gmail.com[4], gayatrividela@gmail.com[5]*

Abstract
Currently, the risk of network information insecurity is increasing rapidly in number and level of danger. The method mostly used by hackers today is to attack end to end technology and exploit human vulnerabilities. These techniques include social engineering, phishing, pharming, etc. one of the steps in conducting these attacks is to deceive users with malicious Uniform Resource Locators (URLs). As results, malicious URL detection is of great interest nowadays. there have been several scientific studies showing a number of methods to detect malicious URLs based on machine learning and deep learning techniques. In this paper, we propose a malicious URL detection method using machine learning techniques based on our proposed URL behavior and attributes. moreover, bigdata technology is also exploited to improve the capability of detection malicious URLs based on abnormal behaviour. In short, the proposed detection system consists of a new set of URLs features and behavior, a machine learning algorithm, and a bigdata technology. the experimental results show that the proposed URL attributes and behaviour can help improve the ability to detect malicious URL significantly. This is suggested that the proposed system may be considered as an optimized and friendly used solution for malicious URL detection.

Keywords: *ML Algorithms; Database; Python; Graphical User Interface; User browser.*

1. Introduction
Uniform Resource Locator (URL) is used to refer to resources on the Internet. In presented about the characteristics and two basic components of the URL as: protocol identifier, which indicates what protocol to use, and resource name, which specifies the IP address or the domain name where the resource is located. It can be seen that each URL has a specific structure and format. Attackers often try to change one or more components of the URL's structure to deceive users for spreading their malicious URL. Malicious URLs are known as links that adversely affect users. These URLs will redirect users to resources or pages on which attackers can execute codes on users' computers, redirect users to unwanted sites, malicious website, or other phishing site, or malware download Malicious URLs can also be hidden in download links that are deemed safe and can spread quickly through file and message sharing in shared network. Some attack techniques that use malicious URLs include Drive-by Download, Phishing and Social Engineering, and Spam. According to statistics presented in, in 2019, the attacks using spreading malicious URL technique are ranked first among the 10 most common attack techniques. Especially, according to this statistic, the three main URL spreading techniques, which are malicious URLs, botnet URLs, and phishing URLs, increase in number of attacks as well as danger level. From the statistics of the increase in the number of malicious URL distributions over the consecutive years, it is clear that there is a need to study and apply techniques or methods to detect and prevent these malicious URLs [1-3]. The paper also includes a new URL attribute extraction method. In our research, machine learning algorithms are used to classify URLs based on the features and behaviors of URLs. The features are extracted from static and dynamic behaviors of URLs and are new to the literature. Those newly proposed features are the main contribution of the research. Machine learning algorithms are a part of the whole malicious URL detection system. Two supervised machine learning algorithms are used, Support vector machine (SVM) and Random-forest (RF). 64 features that are non-

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0334
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2307-2311

binary contain both numerical and discrete, ordinal values. Classification accuracy has to be achieved by relying mainly on the power of the methods used. There is no involvement of domain experts to comment on feature importance nor correlations. For the entire dataset, roughly 33% of all URLs are classified as malicious [4-9].

1.1. Scope and context

The scope of this project encompasses the development and implementation of a comprehensive malicious URL detection system leveraging machine learning techniques and big data technology. Specifically, the project will focus on defining and extracting a refined set of URL features and behaviors that are indicative of malicious intent. These features will serve as inputs to the machine learning algorithm, which will be trained to accurately classify URLs as either malicious or benign. Additionally, the project will explore the integration of big data technology to enhance detection capabilities by identifying and analyzing abnormal URL behaviors at scale. The system will be designed to be adaptable and scalable, capable of handling large volumes of URL data in real-time or near-real-time scenarios. Moreover, the project will involve rigorous testing and evaluation to assess the effectiveness and efficiency of the proposed system in detecting malicious URLs across different environments and scenarios. Overall, the scope of the project encompasses the development of a robust and reliable solution for detecting malicious URLs, with the potential for broader applications in cybersecurity and threat detection.

1.2. Importance

Malicious web pages that launch client- side attacks on web browsers have become an increasing problem in recent years. High interaction client honeypots are security devices that can detect these malicious web pages on a network. However, high interaction client honeypots are both resource- intensive and known to miss attacks. This paper presents a novel classification method for detecting malicious web pages that involves inspecting the underlying static attributes of the initial HTTP response and HTML code. Identification of malicious web pages with static heuristics involves analyzing various attributes of a web page without executing its code or interacting with its content dynamically. This approach is useful for quickly determining whether a web page is potentially harmful based on its structure, code patterns, or other static characteristics.

2. Method

The literature review for URL detection that detects phishing, malware, safe, and defacement using the XG_Boost algorithm with 21 test cases is not explicitly mentioned in the provided sources. However, the sources do discuss various techniques and algorithms for detecting malicious URLs, including phishing, malware, and defacement, but they do not specifically focus on the XGBoost algorithm with 21 features. Uniform Resource Locator (URL) is created to address web pages. The Figure 1 & Figure 2 below shows relevant parts in the structure of a typical URL. A phisher has full control over the subdomain portions and can set any value to it. The URL may also have a path and file components which, too, can be changed by the phisher at will. The subdomain name and path are fully controllable by the phisher. We use the term FreeURL to refer to those parts of the URL in the rest of the article. URL can be set only once [10-13]. The phisher can change FreeURL at any time to create a new URL. The reason security defenders struggle to detect phishing domains is because of the unique part of the website domain (the FreeURL). When a domain detected as a fraudulent, it is easy to prevent this domain before an user access to it.

2.1. Dataset

In this dataset, there are 6 lakhs URLs, including 1500 malicious URL data and 1500 benign URL data. Phishing URLs were collected from a service called Phish Tank, an open-source service. Phish Tank provides collaborative data on phishing on the Internet through a database of phishing information, with the service providing multiple formats of data such as csv, json, and many more, which are updated hourly. My research led me to find a data set that contains benign, spam, phishing, malware & defacement URLs. My source is the University of New Brunswick and the number of legitimate URLs in this collection is 35,300. Malicious URLs and benign URLs are combined in this data set. Once the

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0334
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2307-2311

models are trained, they are used to make predictions or classifications on new or unseen.

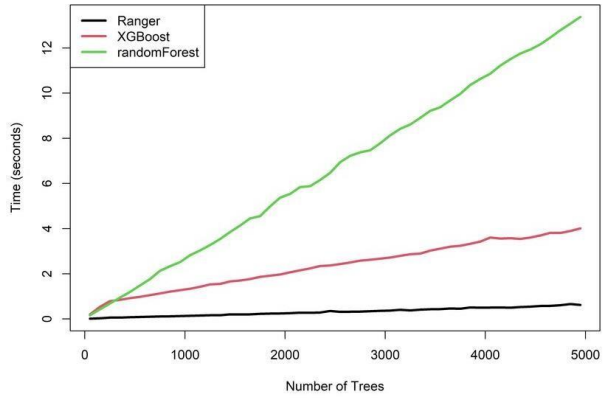

Figure 1 XG-Boost Algorithm

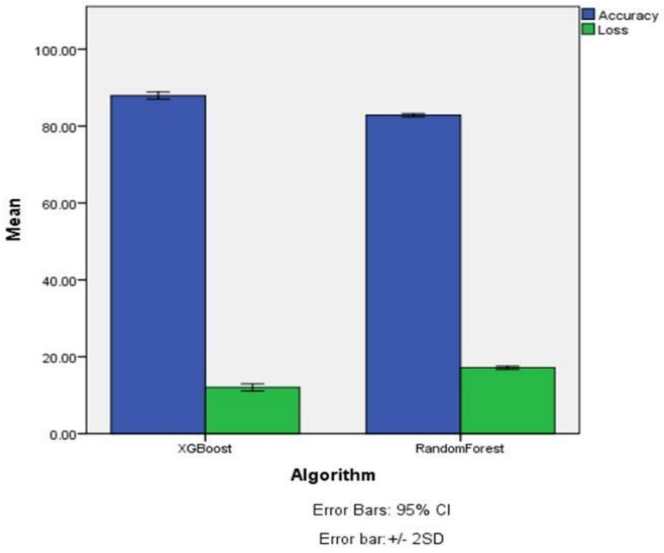

Figure 2 Comparison Between Two Algorithms

The activity diagram provides a visual representation of the workflow involved in handling modules, importing data, processing data, managing features and labels, training machine learning algorithms, and making predictions within a system. It helps in understanding the sequence of activities and decision points involved in the process, facilitating analysis, design, and communication of system behavior and the workflow involving modules, importing data, data processing, handling features and labels, training machine learning algorithms, and making predictions. In our dataset, we find that legitimate websites have been redirected one-time max. On the other hand, phishing websites containing this feature have been redirected at least 4 times.

3. Results and Discussion

3.1. Results

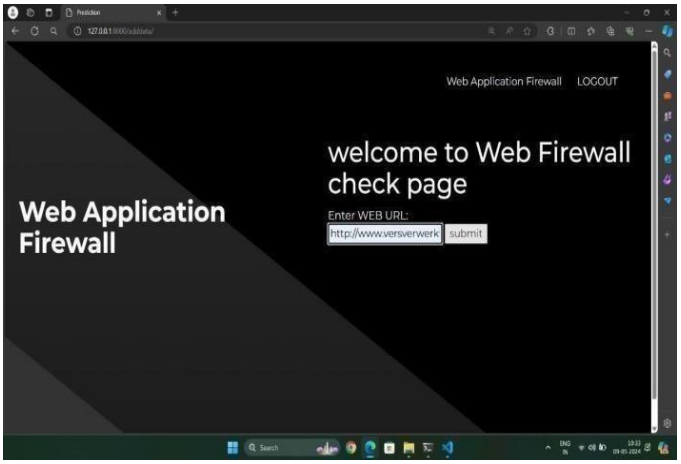

Figure 3 Web Application Firewall

3.2. Discussion

By analyzing test results Figure 3, stakeholders can make informed decisions regarding the quality, reliability, and readiness of the software for deployment or subsequent development phases. Test results serve as a critical mechanism for identifying defects, prioritizing corrective actions, and validating that the software meets specified requirements and quality standards [14]. Continued advancements in AI and machine learning algorithms will enable more accurate and efficient detection of malicious URLs. Deep learning techniques, such as neural networks and reinforcement learning, can be further optimized to analyze vast amounts of data and identify patterns indicative of malicious behavior. Enhancements in DPI technologies will allow for more granular inspection of network traffic to detect and block malicious URLs in real-time. This includes the development of faster and more efficient DPI hardware and software solutions capable of handling high-speed networks. Future systems will increasingly rely on behavioral analysis and anomaly detection techniques to identify suspicious URL activity. By establishing baselines of normal behavior, these systems can detect deviations that may indicate the presence of malicious URLs or cyber threats.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0334
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2307-2311

Conclusion

In this project, a method for malicious URL detection using machine learning is presented. The empirical results have shown the effectiveness of the proposed extracted attributes. In this study, we do not use special attributes, nor do we seek to create huge datasets to improve the accuracy of the system as many other traditional publications. Here, the combination between easy-to- calculate attributes and big data processing technologies to ensure the balance of the two factors is the processing time and accuracy of the system. The results of this research can be applied and implemented in information security technologies in information security systems. The results of this have been used to build a free tool to detect malicious URLs on web browsers. The future of malicious URL detection will be characterized by the integration of advanced technologies, collaboration among stakeholders, and a commitment to privacy and regulatory compliance. These enhancements will enable organizations to better protect against the growing threat of malicious URLs and cyberattacks.

Acknowledgements

We wish to take this opportunity to express our deep gratitude to all the people who have extended their cooperation in various ways during our project work. It is our pleasure and responsibility to acknowledge the help of all those individuals. We sincerely thank our guide Mr. P. V Kishore Kumar, Assistant Professor, the Department of CSE (Cyber Security) for helping us in successful completion of our project under his supervision. We express our deepest gratitude to The Management of Ramachandra College of Engineering, Eluru for their support and encouragement in completing our project work and providing us necessary facilities. We sincerely thank all the Faculty Members and Staff of the Department of CSE (Cyber Security) for their valuable advices, suggestions and constant encouragement which played a vital role in carrying out this project work. Finally, we thank one and all who directly or indirectly helped us to complete our project work successfully. We would like to express our sincere gratitude to their financial support and resources, which made this research possible. This work was supported by my team from my college. We are deeply grateful to our advisor, whose guidance, expertise, and invaluable feedback were crucial to the successful completion of this project.Special thanks to my professors for their insightful suggestions and constructive criticism, which significantly enhanced the quality of this research. We would like to thank the IT support team for their assistance in setting up the necessary infrastructure and for their continuous technical support. Our gratitude goes to my department for providing the essential tools and platforms used in our experiments. This project benefited greatly from the collaborative efforts of our students at Ramachandra College of Engineering, particularly, whose collaboration and shared knowledge were indispensable. And this project aims and stated that the machine learning is presented. The empirical results have shown the effectiveness of the proposed extracted attributes. In this study, we do not use special attributes, nor-do-we-seek-to-create-huge.

Future Enhancement

The future of malicious URL detection will likely involve a combination of advanced technologies and innovative approaches to address increasingly sophisticated cyber threats. Here are some potential enhancements: AI and Machine Learning Advancements: Continued advancements in AI and machine learning algorithms will enable more accurate and efficient detection of malicious URLs. Deep learning techniques, such as neural networks and reinforcement learning, can be further optimized to analyze vast amounts of data and identify patterns indicative of malicious behavior. Deep Packet Inspection (DPI) Improvements: Enhancements in DPI technologies will allow for more granular inspection of network traffic to detect and block malicious URLs in real-time. This includes the development of faster and more efficient DPI hardware and software solutions capable of handling high-speed networks. Behavioral Analysis and Anomaly Detection: Future systems will increasingly rely on behavioral analysis and anomaly detection techniques to identify suspicious URL activity. By establishing baselines of normal behavior, these systems can detect deviations that may indicate the

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0334
e ISSN: 2584-2854
Volume: 02
Issue: 07 July 2024
Page No: 2307-2311

presence of malicious URLs or cyber threats. Blockchain Technology: The use of blockchain technology can enhance the trustworthiness and integrity of URL data. By maintaining a decentralized and immutable ledger of URLs, blockchain-based solutions can provide a secure and transparent way to verify the legitimacy of URLs and prevent tampering or manipulation.The future of malicious URL detection will be characterized by the integration of advanced technologies, collaboration among stakeholders, and a commitment to privacy and regulatory compliance. These enhancements will enable organizations to better protect against the growing threat of malicious URLs and cyberattacks.

References

[1]. D. Sahoo, C. Liu, S.C.H. Hoi, “Malicious URL Detection using Machine Learning: A Survey”. CoRR, abs/1701.07179, 2017.

[2]. M. Khonji, Y. Iraqi, and A. Jones, “Phishing detection: a literature survey,” IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013.

[3]. M. Cova, C. Kruegel, and G. Vigna, “Detection and analysis of drive by download attacks and malicious JavaScript code,” in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp.281– 290.

[4]. R. Heartfield and G. Loukas, “A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks,” ACM Computing Surveys (CSUR), vol. 48, no. 3, p. 37, 2015.

[5]. S. Sheng, B. Wardman, G. Warner, L. F. Cranor, J. Hong, and C. Zhang, “An empirical analysis of phishing blacklists,” in Proceedings of Sixth Conference on Email and Anti-Spam (CEAS), 2009.

[6]. C. Seifert, I. Welch, and P. Komisarczuk, “Identification of malicious web pages with static heuristics,”in Telecommunication Networks and Applications Conference, 2008. ATNAC 2008. Australasian. IEEE, 2008, pp. 91–96. [7] S. Sinha, M. Bailey, and F. Jahanian, “Shades of grey: On the effectiveness of reputation- based “blacklists”,” in Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conferenceon. IEEE, 2008, pp. 57–64.

[7]. Y. Yu, J. Chen, Y. Xie, and H. Han, "A Dynamic URL Blacklist Based on Pattern Recognition," Proceedings of the 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, China, 2018.

[8]. M. Shar, R. Tan, and Y. Javed, "Whitelist-based Phishing URL Detection," Journal of Information Security and Applications, vol. 53, pp. 102499, 2020.

[9]. Z. Le, A. Markopoulou, and M. Faloutsos, "PhishDef: URL Names Say It All," Proceedings of the 2020 IEEE International Conference on Computer Communications (INFOCOM), Toronto, Canada, 2020.

[10]. Y. Li, T. Li, and C. Tian, "Malicious URL Detection Based on Machine Learning," Security and Communication Networks, vol. 2018, Article ID 8021340, 2018.

[11]. A. A. A. El-Mohandes, H. E. Badr, and M. E. Khalifa, "A Real-Time URL Phishing Detection System," Proceedings of the 2018 International Conference on Computer and Applications (ICCA), Beirut, Lebanon, 2018.

[12]. P. Marchal, P. Francois, R. Engle, and M. Matthieu, "Securing Web Servers Using Machine Learning," Proceedings of the 2019 IEEE Conference on Communications and Network Security (CNS), Washington, DC, USA, 2019

[13]. Birari, H. P., lohar, G. V., & Joshi, S. L. (2023). Advancements in Machine Vision for Automated Inspection of Assembly Parts: A Comprehensive Review. International Research Journal on Advanced Science Hub, 5(10), 365-371. doi: 10.47392/IRJASH.2023.065.

[14]. Rajan, P., Devi, A., B, A., Dusthackeer, A., & Iyer, P. (2023). A Green perspective on the ability of nanomedicine to inhibit tuberculosis and lung cancer. International Research Journal on Advanced Science Hub, 5(11), 389-396. doi: 10.47392/IRJASH.2023.071.