Predictive Modeling for Web Server Vulnerability Detection Using Machine Learning

THESIS

Submitted as fulfilment of the requirements for the completion of

Master of Computer Science Program

CHRISTANTO PAULUS RUMAPEA

20210130033

MASTER OF COMPUTER SCIENCE PROGRAM

SCHOOL OF COMPUTER SCIENCE

2024

STATEMENT OF AUTENTICITY

The undersigned below:

Name	:	Christanto Paulus Rumapea
ID of student	:	20210130033
Faculty	:	Computer Science
The Title of Thesis	:	Predictive Modeling for Web Server Vulnerability Detection Using Machine Learning

Stating truthfully that this thesis has nothing in common with other thesis. Thus this statement is made without coercion from any party. If this statement is not true, it will be sanctioned by the faculty leader.

Sukabumi, August 2024

Christanto Paulus Rumapea
NIM. 20210130033

APPROVAL OF THESIS

Title	:	Predictive Modeling for Web Server Vulnerability Detection Using Machine Learning
Name	:	Christanto Paulus Rumapea
ID of student	:	20210130033

The thesis has been reviewed and approved

Sukabumi, August 2024

Head of Study Program,	Supervisor
Prof. Ir. Teddy Mantoro, M.Sc., PhD.	Jelita Asian, M.Sc., PhD.
NIDN. 0323096491	NIDN. 0406097702

THESIS APPROVAL

Title	:	Predictive Modeling for Web Server Vulnerability Detection Using Machine Learning
Name	:	Christanto Paulus Rumapea
ID of student	:	20210130033

This Thesis has been tested and defended in front of the Board of Examiners in Thesis session on 9th August 2024. In our review, this Thesis adequate in terms of quality for the purpose of awarding the Master of Computer Degree.

Sukabumi, August 2024

Supervisor 1,	Examiner 1
Prof. Ir. Teddy Mantoro, M.Sc., PhD.	Prof. Ir. Media Anugerah Ayu, M.Sc., PhD
NIDN. 0323096491	NIDN. 0315046903
Supervisor 2,	Examiner 2,
Jelita Asian, M.Sc., PhD.	Haris Al Qodri Maarif, M.Sc., PhD.
NIDN. 0406097702	NIDN. 0418068505

PUBLICATION APPROVAL

As a member of the academic community of Nusa Putra University, i undersigned:

Name	: Christanto Paulus Rumapea
ID of Student	20210130033
Study Program	: Computer Science
Type of Work	: Thesis

For the sake of scientific development, agree to grant to the University of Nusa Putra the Non-Exclusive Royality-Free Right for my scientific work entitled: **Predictive Modeling for Web Server Vulnerability Detection Using Machine Learning.**

Along with existing devices (if needed). With this non-exclusive royalty-free right, Nusa Putra University has the right to store, transfer media/formats, process in the form of a database, maintain and publish my thesis as long as I keep my name as the author/creator and as the copyright owner.

This statement I made in truth.

Made in : Sukabumi

At the Date of : August 2024

That States

(Christanto Paulus Rumapea)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FOREWORD

Praise and gratitude to Allah SWT Almighty, because only with His blessings and grace can the writer complete this thesis. Writing this thesis is one of the requirements to achieve a Master's degree in Computers at Nusa Putra University. I realize that, without the help and guidance of various parties, from the lecture period to the preparation of this thesis, it is very difficult for the writer to complete this thesis. Therefore, I would like to thank:

1. Dr. Kurniawan, ST., M.Sc., MM. as Chancellor of Nusa Putra University;
2. Anggy Pradiftha Junfitrana, MT. as Vice Chancellor 1 for Academic Affairs;
3. Prof. Ir. Teddy Mantoro, M.Sc., PhD as Head of School Computer Science Nusa Putra University and Supervisor 1;
4. Jelita Asian, M.Sc., PhD. as Supervisor 2;
5. All Masters of Computer Science Lecturers who have provided very useful knowledge during lectures;
6. Mr. Lutfil Khakim, S.Kom., M.Si. for the support and encouragement to writer for completing study in Nusa Putra University;
7. Fellow comrades in Master of Computer Science batch 2021 who always give encouragement and support for the writer to complete this thesis;
8. All parties who have helped the writer in writing this thesis; For further improvement, suggestions and constructive criticism will be gladly accepted.

Sukabumi, August 2024

Writer

ABSTRACT

In 2023, the Indonesian government's websites faced significant challenges with web defacement attacks, totaling 189 incidents, with the highest number, 31, occurring in January. Web defacement, which exploits vulnerabilities in web servers to alter or delete web page content, poses serious risks to data integrity and user privacy. This study addresses these challenges by proposing a comprehensive framework for identifying and evaluating security vulnerabilities in website technologies using machine learning techniques. The framework integrates data collection, preprocessing, training and modeling, and analysis into a seamless process. Data is collected using a Website Technology Crawler to identify technology stacks and Common Vulnerabilities and Exposures (CVE) databases to gather information on known vulnerabilities. The research highlights the Random Forest Classifier as the most effective model, achieving an impressive accuracy of 98%, with precision and recall scores of approximately 0.98. These metrics underscore the model's capability to accurately identify and distinguish between safe and exploitable vulnerabilities. Key features such as `cve_number`, `version`, and `product_name` were critical indicators of exploitability, significantly enhancing predictive accuracy. Comparative analysis showed that the K-Nearest Neighbors (KNN) classifier also performed well, with an accuracy of 97%, while the Support Vector Classifier (SVC) had a slightly lower accuracy of 88%. The Random Forest model's ROC curve, with an AUC score of 0.99, highlighted its exceptional ability to discriminate between positive and negative classes. The deployment of the Random Forest model in a real-time prediction platform demonstrated its practical applicability, offering an efficient approach to vulnerability management. This research contributes to cybersecurity by providing a systematic and reliable approach to vulnerability detection, significantly improving the proactive identification and mitigation of exploitable vulnerabilities in Website Technologies.

Keyword: Website Vulnerabilities, Machine Learning, Random Forest, Support Vector Machine, K-Nearest Neighbors, Cyber Security

CHAPTER I
INTRODUCTION

1.1 Research Background

In the modern digital landscape, cybersecurity has emerged as a paramount concern for organizations, governments, and individuals alike. The reliance on digital systems for everyday operations has increased the potential for cyber threats, making robust cybersecurity measures indispensable. The evolution of cyberattacks in terms of frequency and sophistication has prompted the need for advanced security strategies to protect sensitive information and ensure the integrity of digital infrastructures. As technology progresses, so do the tactics and methods employed by malicious actors to exploit vulnerabilities in these systems. This growing threat landscape necessitates the continuous improvement of cybersecurity defenses to keep pace with emerging risks (Chio & Freeman, 2018).

In 2023, the Indonesian government's websites experienced significant issues with web defacement attacks. A total of 189 cases of web defacement were recorded, with the highest number of incidents occurring in January, where 31 cases were reported. Web defacement involves exploiting vulnerabilities in web servers to modify or delete content on the affected web pages, often causing reputational damage and loss of trust (BSSN, 2023).

The sector most impacted by these attacks was the Administrative Government sector, which accounted for 167 of the total cases. Other affected sectors included Health with 7 cases, Other sectors with 12 cases, and Defense with 3 cases. The majority of the web defacement incidents (93.1%) targeted hidden pages within websites, which are less likely to be immediately noticed by users. Only 6.9% of the cases affected the homepage of the websites, which typically results in more noticeable and immediate disruption.

Machine Learning-Based Website Website Vulnerability Identification presents several benefits that address the challenges faced by traditional methods, such as Machine learning models, particularly those like Random

Forest, can analyze complex datasets to identify patterns and correlations that might indicate potential vulnerabilities. These models can handle large volumes of data, allowing for a more comprehensive assessment compared to traditional methods. The ability to learn from data and improve over time enhances the precision of detecting vulnerabilities, reducing false positives and negatives. One of the primary advantages of machine learning in vulnerability detection is the automation of the identification process. This reduces the need for manual intervention, which can be time-consuming and error-prone. Additionally, machine learning models can be scaled to handle numerous websites and applications simultaneously, making them suitable for large-scale deployments. Traditional security measures often rely on predefined signatures or rules, making them less effective against new or evolving threats. Machine learning models, however, can adapt to new types of vulnerabilities by learning from new data. This dynamic adaptability is crucial in the ever-changing landscape of cybersecurity, where new vulnerabilities are constantly emerging. Website vulnerability detection often involves analyzing a mix of structured and unstructured data, including code, metadata, and behavioral patterns. Machine learning algorithms, especially ensemble methods like Random Forests, are capable of handling this complexity. They can manage missing data, non-linear relationships, and high-dimensional data, providing robust insights into potential security risks. Machine learning models can identify potential vulnerabilities before they are exploited by attackers. By analyzing patterns and anomalies, these models can detect unusual activity that may indicate an impending threat, allowing organizations to take preventive measures in advance. This proactive approach is crucial for maintaining the security and integrity of Website Technologiess. With access to ongoing data inputs, machine learning models can continuously learn and refine their predictions. This iterative learning process means that the models become more accurate and reliable over time, providing better protection against emerging vulnerabilities. Machine learning can integrate multiple data sources, such as CVE databases, Website Technologies logs, and network traffic, to provide a holistic view of potential vulnerabilities. This comprehensive analysis enables

a deeper understanding of the security posture and helps prioritize the most critical issues.

Machine learning (ML) has revolutionized various fields by enabling systems to learn from data and make intelligent decisions. In the realm of cybersecurity, ML algorithms can analyze vast amounts of data to detect patterns and predict potential threats. The application of ML in cybersecurity has shown promise in enhancing the effectiveness of security measures and providing more proactive defense strategies (Chio & Freeman, 2018). Machine learning has been applied to various aspects of cybersecurity, offering innovative solutions to complex problems. One notable application is anomaly detection, where ML algorithms are used to identify deviations from normal behavior, indicating potential security incidents. For instance, ML models can analyze network traffic patterns and detect anomalies that may signify an ongoing attack or unauthorized access. Another application is malware detection, where ML models are trained to classify and detect malware based on patterns in the data. These models can identify new and previously unknown malware variants by recognizing characteristics similar to known malicious software. Additionally, ML can enhance intrusion detection systems (IDS) by identifying suspicious activities in network traffic and triggering alerts for potential threats. The ability of ML to process and analyze large volumes of data in real-time makes it a valuable tool for improving the accuracy and efficiency of security measures (Sommer & Paxson, 2010).

While the integration of ML with website vulnerability identification testing shows promise, there are several challenges that need to be addressed to fully realize its potential. One of the primary challenges is data quality, as the effectiveness of ML models heavily depends on the quality and diversity of the training data. Ensuring that the data used to train the models accurately represents a wide range of attack scenarios and vulnerabilities is crucial for developing robust ML models. Another challenge is model interpretability, as understanding the decision-making process of ML models is important for validating their results and ensuring their reliability. Developing methods to interpret and explain the predictions made by ML models will be essential for

gaining trust in these systems. Additionally, the security of ML models themselves is a critical concern, as adversarial attacks can be used to manipulate the models and compromise their effectiveness. Research into techniques to secure ML models against such attacks will be necessary to maintain their integrity and reliability (Goodfellow et al., 2016).

This study aims to develop a comprehensive machine learning-based framework for identifying vulnerabilities in website technologies. The framework integrates several key processes: data collection, pre-processing, training and modeling, and analysis and evaluation. By systematically addressing each stage, the proposed solution seeks to provide a holistic and effective approach to web security. The framework utilizes a Website Technology Crawler to identify the technology stack of target websites and employs Common Vulnerabilities and Exposures (CVE) databases to gather relevant information on known vulnerabilities. This data is then subjected to rigorous cleansing and feature extraction to create a robust dataset for training the machine learning model. The core of the framework is the Random Forest algorithm, which is used to train a model capable of predicting exploitable vulnerabilities with high accuracy.

The significance of this research lies in its potential to enhance the security of Website Technologiess by providing a reliable and efficient method for vulnerability detection. The proposed framework not only aids in identifying existing vulnerabilities but also contributes to the proactive prevention of future attacks. By leveraging advanced machine learning techniques, this study aims to bridge the gap between traditional security practices and the needs of modern web environments. The findings of this research are expected to have practical implications for cybersecurity professionals, web developers, and organizations seeking to protect their digital assets.

In summary, this research addresses a critical need in the field of cybersecurity by proposing a novel machine learning-based approach to web vulnerability detection. The introduction sets the stage for a detailed exploration of the methods and techniques used in the study, providing a

foundation for understanding the complexities and challenges involved in securing Website Technologiess. The subsequent sections will delve deeper into the specifics of the proposed framework, including its design, implementation, and validation. Through this comprehensive analysis, the study aims to contribute valuable insights and tools to the ongoing efforts to safeguard digital infrastructures and protect user data.

1.2 Problem Statement

The rapid advancement of web technologies and the increasing complexity of Website Technologiess have heightened the potential for cyber threats. Traditional methods for identifying and mitigating security vulnerabilities, such as manual penetration testing, are often time-consuming, expensive, and reliant on expert knowledge. These methods struggle to keep pace with the evolving nature of cyberattacks and the sheer volume of data that must be analyzed to detect vulnerabilities. There is a critical need for more efficient and scalable solutions that can provide timely and accurate detection of security vulnerabilities in website technologies. This research aims to address this gap by exploring the application of machine learning, to develop a framework for website vulnerability identification.

1.3 Research Objectives

a. Evaluate and compare various machine learning algorithms to identify the most effective model for detecting vulnerabilities in Website Technologiess used by the Indonesian government. The results will be analyzed to identifying the best model for vulnerability detection of Indonesian government Website Technologiess.
b. Develop and implement a machine learning-based framework that uses the best-performing algorithm for real-time vulnerability detection in Indonesian government Website Technologies.

1.4 Significance of Research

This research contributes to the field of cybersecurity by offering a novel approach to vulnerability detection using machine learning. By addressing the limitations of traditional methods, the proposed framework can provide a more efficient, scalable, and accurate means of identifying vulnerabilities. The study aims to improve the security of Website Technologiess, protect sensitive data, and enhance the overall cybersecurity posture of organizations. The findings will be valuable for cybersecurity professionals, web developers, and researchers, offering practical tools and insights for safeguarding digital infrastructures.

1.5 Scope and Limitations

a. The study focuses on the development and evaluation of a machine learning-based framework for detecting vulnerabilities in Website Technologiess. It will cover various stages, including data collection using website technology crawlers and CVE databases, data preprocessing, feature extraction, model training, and evaluation.
b. The framework's effectiveness depends on the quality and diversity of the training data. Additionally, while machine learning models can significantly improve detection capabilities, they are not infallible and may still produce false positives or negatives. The study will primarily focus on vulnerabilities identifiable through publicly available data and may not cover all potential security issues. Furthermore, the models' adaptability to new threats relies on continuous learning and updating with new data, which may require additional resources and infrastructure.

1.6 Thesis Structure

The rest of thesis is organized as follows:

1. Chapter I describes the background of problem that will be discussed in the thesis.
2. Chapter II describes the literature review of thesis.

3. Chapter III describes the methodology of thesis.
4. Chapter IV presents the expereiment result and discussion.
5. Chapter V presents the conclusion the thesis and future work.

Nusa Putra
UNIVERSITY
2002
AMOR DEUS, PARENTIUM, CONCERVIS

CHAPTER V
CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusion

The research undertaken in this study focused on developing a robust machine learning model capable of accurately classifying vulnerabilities in Website Technologiess as either "Safe" or "Exploitable," despite the inherent class imbalance in the dataset. Utilizing advanced machine learning techniques, particularly the Random Forest Classifier, this study has demonstrated the potential of these models to predict security vulnerabilities with remarkable accuracy.

The Random Forest Classifier emerged as the most effective model, achieving an impressive accuracy of 98%, highlighting its ability to handle complex data interactions and provide reliable predictions. This high accuracy is largely attributed to the model's use of ensemble learning methods, which combine multiple decision trees to improve predictive performance. The model's precision and recall for both classes were approximately 0.98, indicating its strong capability to accurately identify and distinguish between safe and exploitable vulnerabilities.

The analysis of feature importance revealed that features such as cve_number, version, and product_name were critical in predicting vulnerabilities. These features were instrumental in the model's predictive accuracy, suggesting that specific product versions and their associated CVEs are key indicators of exploitability. These insights underscore the importance of carefully selecting features that capture essential characteristics of the dataset, enhancing the model's ability to make accurate predictions.

In a comparative analysis, the K-Nearest Neighbors (KNN) classifier also performed well, with an accuracy of 97%, indicating that instance-based learning methods, which rely on proximity to neighbors, are also effective for this classification task. However, the Support Vector Classifier (SVC) lagged behind with an accuracy of 88%, particularly showing lower performance in precision and recall for class 0. These results emphasize the necessity of

selecting the appropriate model for the dataset, as different algorithms interpret and process data in unique ways.

The model evaluation metrics, including confusion matrices and ROC curves, provided a comprehensive understanding of each model's performance. The Random Forest model's ROC curve, with an Area Under the Curve (AUC) score of 0.99, highlighted its exceptional ability to discriminate between positive and negative classes, further validating its effectiveness as a predictive model.

Additionally, the deployment of the Random Forest model within a real-time prediction platform demonstrated its practical applicability in assessing security vulnerabilities. The integration with nmap allowed for automated scanning and prediction, offering a streamlined approach to vulnerability management in Website Technologiess, thus proving the model's utility in real-world scenarios.

5.2 Recommendations

Based on the findings, several recommendations for future work and practical applications include exploring additional features and feature engineering techniques to further enhance model accuracy. Incorporating domain-specific features or leveraging additional data sources might provide deeper insights into vulnerability characteristics. Additionally, further efforts in hyperparameter tuning and exploring other ensemble methods, such as Gradient Boosting or XGBoost, could yield even better performance.

Ensuring scalability and optimizing the system for real-time predictions will be crucial as the platform is deployed on a larger scale. Improving the user interface to provide more intuitive insights and actionable recommendations based on predictions will enhance user engagement and decision-making processes.

In conclusion, this research highlights the significant potential of machine learning models in enhancing cybersecurity through accurate vulnerability prediction and management. By refining model techniques and expanding the platform's capabilities, organizations can significantly improve their ability to

preemptively identify and mitigate exploitable vulnerabilities in their Website Technologiess, thereby strengthening their overall security posture.

REFERENCES

Abdulghaffar, K., Elmrabit, N., & Yousefi, M. (2023). Enhancing Web Application Security through Automated Penetration Testing with Multiple Vulnerability Scanners. *Computers*, *12*(11), 235. https://doi.org/10.3390/computers12110235

Akram, J., Qi, L., & Luo, P. (2019). VCIPR: Vulnerable Code is Identifiable When a Patch is Released (Hacker's Perspective). *2019 12th IEEE Conference on Software Testing, Validation and Verification (ICST)*, 402–413. https://doi.org/10.1109/ICST.2019.00049

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

BSSN (2023). *Lanskap Keamanan Siber Indonesia 2023*. BSSN. Indonesia's Cyber Security Landscape 2023

Chio, C., & Freeman, D. (2018). *Machine learning and security: Protecting systems with data and algorithms* (First edition). O'Reilly Media.

Cox, D. R. (1959). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *21*(1), 238–238. https://doi.org/10.1111/j.2517-6161.1959.tb00334.x

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Hilario, E., Azam, S., Sundaram, J., Imran Mohammed, K., & Shanmugam, B. (2024). Generative AI for pentesting: The good, the bad, the ugly. *International Journal of Information Security*, *23*(3), 2075–2097. https://doi.org/10.1007/s10207-024-00835-x

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Third edition). Wiley.

J. Santhosh Kumar, B., & Pujitha, K. (2018). Web Application Vulnerability Detection Using Hybrid String Matching Algorithm. *International Journal of Engineering & Technology*, *7*(3.6), 106. https://doi.org/10.14419/ijet.v7i3.6.14950

Krasniqi, G., & Bejtullahu, V. (2018, October 27). Vulnerability Assessment and Penetration Testing: Case study on web application security. *2018 UBT International Conference*. University for Business and Technology International Conference, Pristina, Kosovo. https://doi.org/10.33107/ubt-ic.2018.213

Kumar, S., Mahajan, M., & Batra, S. (2023). A Recent Study of Machine Learning Based Techniques for the Detection of Cyber-Attacks on Web Applications. *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, 153–158. https://doi.org/10.1109/IC3I59117.2023.10397832

Liao, Y., & Vemuri, V. R. (2002). Use of K-Nearest Neighbor classifier for intrusion detection. *Computers & Security*, *21*(5), 439–448. https://doi.org/10.1016/S0167-4048(02)00514-X

Locasto, M. E., Wang, K., Keromytis, A. D., & Stolfo, S. J. (2006). FLIPS: Hybrid Adaptive Intrusion Prevention. In A. Valdes & D. Zamboni (Eds.), *Recent Advances in Intrusion Detection* (Vol. 3858, pp. 82–101). Springer Berlin Heidelberg. https://doi.org/10.1007/11663812_5

Nguyen, T. T. T., & Armitage, G. (2008). A survey of techniques for internet traffic classification using machine learning. *IEEE Communications Surveys & Tutorials*, *10*(4), 56–76. https://doi.org/10.1109/SURV.2008.080406

Sahoo, D., Liu, C., & Hoi, S. C. H. (2019). *Malicious URL Detection using Machine Learning: A Survey* (arXiv:1701.07179). arXiv. http://arxiv.org/abs/1701.07179

Sahu, D. R., & Tomar, D. S. (2017). Analysis of Web Application Code Vulnerabilities using Secure Coding Standards. *Arabian Journal for Science and Engineering*, *42*(2), 885–895. https://doi.org/10.1007/s13369-016-2362-5

Saini, J., & Bansal, A. (2024). *Automated Penetration Testing: Machine Learning Approach⋆*.

Seyyar, Y. E., Yavuz, A. G., & Unver, H. M. (2022). Detection of Web Attacks Using the BERT Model. *2022 30th Signal Processing and Communications Applications Conference (SIU)*, 1–4. https://doi.org/10.1109/SIU55565.2022.9864721

Shar, L. K., Briand, L. C., & Tan, H. B. K. (2015). Web Application Vulnerability Prediction Using Hybrid Program Analysis and Machine Learning. *IEEE Transactions on Dependable and Secure Computing*, *12*(6), 688–707. https://doi.org/10.1109/TDSC.2014.2373377

Sharma, C., & Jain, S. C. (2014). Analysis and classification of SQL injection vulnerabilities and attacks on web applications. *2014 International*

Conference on Advances in Engineering & Technology Research (ICAETR - 2014), 1–6. https://doi.org/10.1109/ICAETR.2014.7012815

Sharma, S., Zavarsky, P., & Butakov, S. (2020). Machine Learning based Intrusion Detection System for Web-Based Attacks. *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, 227–230. https://doi.org/10.1109/BigDataSecurity-HPSC-IDS49724.2020.00048

Sommer, R., & Paxson, V. (2010). Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. *2010 IEEE Symposium on Security and Privacy*, 305–316. https://doi.org/10.1109/SP.2010.25

Stolfo, S. J., Wei Fan, Wenke Lee, Prodromidis, A., & Chan, P. K. (1999). Cost-based modeling for fraud and intrusion detection: Results from the JAM project. *Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00*, *2*, 130–144. https://doi.org/10.1109/DISCEX.2000.821515

What is a CVE? (2021, November 21). Accessed in July 25th 2024 https://www.redhat.com/en/topics/security/what-is-cve

Ye, Y., Li, T., Adjeroh, D., & Iyengar, S. S. (2018). A Survey on Malware Detection Using Data Mining Techniques. *ACM Computing Surveys*, *50*(3), 1–40. https://doi.org/10.1145/3073559