JOURNAL OF INNOVATION AND TECHNOLOGY
eISSN:2805-5179 Vol. 2024, No.30

Phishing Website Detection using Machine Learning

Padmini Y[1], Usha Sree[1]

[1]Dayananda Sagar Academy of Technology and Management, Karnataka, India

Email: padhmini.yelkur@gmail.com; ushashree-mca@dsatm.edu.my

Abstract

Phishing attacks, a prevalent and significant form of cybercrime, involve attackers masquerading as reputable entities to deceive individuals into revealing sensitive details such as usernames, passwords,and credit card information. Deceptive websites are commonly used in these attacks, appearing legitimate and underscoring the need for individuals and organizations to heighten their awarenessand implement stronger and more advanced detection techniques. By luring sensitive information through deceptive websites, phishing attacks represent a serious cybersecurity threat. In this research, the effectiveness of machine learning algorithms, specifically the Gradient Boosting Classifier, in identifying phishing websites to enhance accuracy and response time is being assessed.

Keywords

Phishing attacks, Machine Learning, Cybersecurity, Gradient Boosting, Websites

Introduction

Phishing attacks represent a prevalent and serious type of cybercrime, in which malicious actors mimic reputable entities to trick individuals into disclosing confidential details such as usernames, passwords, and credit card information. These deceptive tactics frequently involve authentic-looking websites, underscoring the importance of remaining alert. It is imperative for both individuals and organizations to adopt more advanced and creative detection methods in order to efficiently combat these security risks. The goal of this project is to create a machine learning-based system for detecting phishing websites effectively. We will use machine learning techniques to analyze extensive data about both phishing and genuine websites, extracting important features like URL types, webpage content, and metadata to train different machine learning models. Our main emphasis will be on the Gradient Boosting Classifier, a strong ensemble learning technique recognized for its accuracy and resilience. After conducting thorough tests and analysis, our goal isto assess how well the Gradient Boosting Classifier can identify phishing websites. Our aims includeminimizing incorrect identifications, improving accuracy in detection, and speeding up the overall response time of the detection system. This project aims to combine sophisticated machine learning algorithms with phishing detection to

Submission: 13 October 2024; **Acceptance:** 1 November 2024

Copyright: © 2024. All the authors listed in this paper. The distribution, reproduction, and any other usage of the content of this paper is permitted, with credit given to all the author(s) and copyright owner(s) in accordance to common academic practice. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license, as stated in the website: https://creativecommons.org/licenses/by/4.0/

have a substantial impact on cybersecurity, offering users a dependable tool for defending against phishing attacks.

Problem Statement

Phishing detection methods currently face challenges related to their accuracy and ability to adapt to new phishing links. Our solution involves utilizing machine learning through the implementation of different classification algorithms to assess their efficacy on our dataset. We opted for the Gradient Boosting classifier due to its superior classification performance.

Literature Review

Ahmed and Abdullah created a software named Phish Checker, which identifies legitimate and phishing websites by examining URLs and domain names (Ahmed and Abdullah, 2016). The software attained a 96% precision rate by utilizing information from Yahoo and PhishTank catalogs. Nevertheless, its efficacy is constrained by heuristic approaches and the presence of particular discriminative traits. Furthermore, the assessment concentrates exclusively on URL authenticity to establish the genuineness of a website.

In another study, Aydin and Baykal investigated the detection of phishing websites through URL-based feature selection techniques (Aydin and Baykal, 2015). They conducted a comparison between Naive Bayes and Successive Negligible Optimization (SMO) classifiers using a dataset containing legitimate and phishing URLs sourced from Google and PhishTank. SMO demonstrated superior performance to Naive Bayes, achieving a 95.39% accuracy rate, while Naive Bayes attained 88.17% accuracy, despite both classifiers utilizing the same number of features.

Gupta involve the integration of a URL and DNS matching module with awhitelist of trusted websites, which is automatically updated based on users' browsing history, to detect phishing attempts (Ankit Kumar and Gupta, 2016). The accuracy of this method, evaluated using data from sources such as Stuffgate, Alexa, and PhishTank, reached 89.38%.

Jain focused on using only URL features for phishing detection, utilizing datasets from Kaggle and PhishTank (Jain, 2022). They employed a hybrid approach combiningPrincipal Component Analysis (PCA) with Support Vector Machine (SVM) and Random Forest algorithms, reducing the dataset's dimensionality while retaining crucial information. This method achieved an accuracy rate of 96.8%, surpassing other techniques evaluated in their study.

Karabatak and his team evaluates the effectiveness of classification algorithms on a compressed dataset of phishing websites from the UCI Machine Learning Repository (Karabatak and Mustafa, 2018). The results indicate varying effectiveness, with classifiers like KStar, LMT, ID3, and R.F. Classifier performing well, while others such as Lazy, BayesNet, and SGD Multilayer Perceptron were less effective. Notably, the Lazy classifier achieved a high accuracy rate of 97.58% on a compressed 27-feature dataset.

Smadi formulated a technique for differentiating genuine emails from phishing emails by assessing characteristics derived from the email headers and contents (Smadi et al., 2015). They applied ten distinct data mining methods and discovered that the RF, J48, and PART algorithms demonstrated impressive precision rates of 98.87%, 98.11%, and 90.10% correspondingly.

Tang and Yang gathered a dataset of phishing websites from the UCI repository and tested various machine learning techniques, including decision trees, AdaBoost, SVM, and random forests, analyzing features like web traffic, port numbers, URL length, IP address, and URL anchor (Tang et al., 2021; Yang, R et al., 2021). An advanced fusion classifier incorporating these algorithms achieved a 97% accuracy rate, outperforming previous phishing detection methods.

Methodology

The research included gathering a dataset from Kaggle, then performing data cleaning and preprocessing to ready it for machine learning. Essential characteristics, such as the length of the URL, presence of special characters, and domain age, were recognized and extracted. Multiple machine learning models were constructed, focusing on the Gradient Boosting Classifier, which was educated and authenticated using the prepared data. The parameters of the classifier were adjusted toenhance performance, and its outcomes were compared with other models such as XGBoost, Random Forest, and SVM. An intuitive interface for real-time phishing detection was developed, and the system underwent comprehensive testing to pinpoint areas for improvement based on false positives and negatives.

In the proposed system, Gradient Boosting Classifier, known for its high accuracy and ability to handle complex patterns,is the proposed strategy for detecting phishing sites. A website is currently in development to recognize legitimate websites using this machine learning algorithm, which outperforms other algorithms in terms of accuracy. The Gradient Boosting Classifier provides several benefits, such as exceptional predictive accuracy, resistance to overfitting compared to decision trees, and versatility in handling diverse data types and properties, making it suitable for various phishing detection scenarios.

Data Flow Diagram

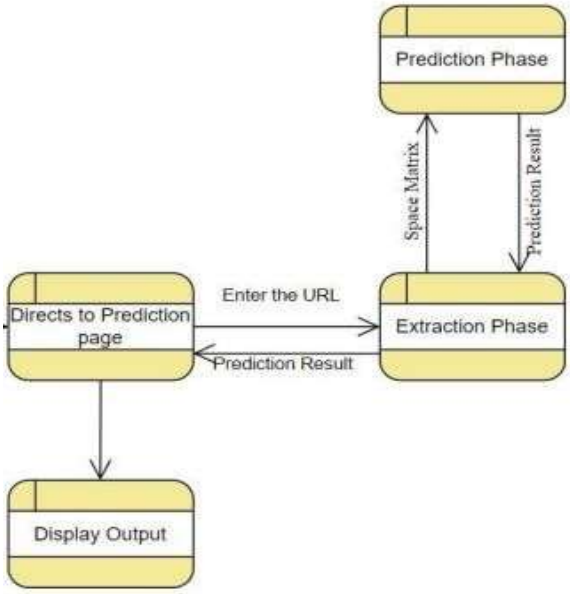

Figure 1: Data Flow Diagram

A Data Flow Diagram (DFD) visually illustrates how information moves within a system, demonstrating the entry, exit, processing, and storage of data. It acts as a tool for communication between system analysts and stakeholders, delineating the system's boundaries and constraints. In the realm of phishing detection, the DFD delineates the procedure for user login, URL entry, and feature extraction for analysis by the system. Following this, the machine learning model makes predictions on whether the website is involved in phishing activities, and subsequently notifies the user accordingly. If phishing is identified, a warning is generated; otherwise, the site is labeled as safe.

Gradient Boosting Classifier

Gradient boosting classifiers comprise a group of machine learning algorithms that aggregate multipleweak learning models in order to create a robust predictive model. When employing gradient boosting, decision trees are commonly used. Boosting algorithms are crucial in managing the trade- off between bias and variance. Unlike bagging algorithms, which primarily address high variance in amodel, boosting addresses both bias and variance, making it generally more effective.

Comparison of Models

Table 1: Comparison of Models

SL.NO	ML Model	Accuracy	f1_score	Recall	Precision
1	Gradient Boosting Classifier	0.974	0.977	0.994	0.9860.967
2	XGBoost Classifier	0.969	0.973	0.993	0.984
3	Random Forest	0.967	0.970	0.993	0.991
4	Support Vector Machine	0.964	0.968	0.980	0.965
5	Decision Tree	0.961	0.965	0.991	0.993

The performance of different machine learning models for detecting phishing is compared in thetable, with evaluation based on metrics such as accuracy, F1 score, recall, and precision. The GradientBoosting Classifier stands out as the top performer, achieving an accuracy of 0.974, an F1 score of 0.977, a recall of 0.994, and a precision of 0.986, demonstrating strong predictive capabilities.Following closely is the XGBoost Classifier, which boasts an accuracy of 0.969 along with a similarly high F1 score and recall, positioning it as a robust alternative. Additionally, the Random Forest model delivers good performance, with an accuracy of 0.967 and a noteworthy precision of 0.991. The Support Vector Machine and Decision Tree models, although slightly less accurate, still show good performance, with the Decision Tree achieving a precision of 0.993. This analysis underscores the Gradient Boosting Classifier's higher accuracy and well-rounded performance in various metrics, establishing it as the preferred option for this use case.

Conclusion

Phishing, a cybercrime method that involves social engineering and technical deception to steal sensitive personal information, is crucial to detect and prevent data breaches. While current classifiers can reasonably predict phishing attacks, a combined approach could potentially improve detection rates. Our study suggests a new method for detecting phishing using URL-based attributes and multiple machine learning classifiers. Through testing, it was found that the Gradient Boosting classifier achieved a high classification accuracy of 97.4% on the phishing dataset. Further work willinvolve applying this model to larger datasets and evaluating the performance of these algorithms based on classification accuracy.

Acknowledgements

The authors would like to express their heartfelt gratitude to Dayananda Sagar Academy of Technology and Management (DSATM) for providing the necessary resources and facilities to conduct this research project on "Guardian of Cybersecurity: Unveiling the Tactics in Phishing Website Detection". The institution's support and encouragement have been crucial to the successful completion of this endeavour. Furthermore, we extend our deepest thanks to our families, especially our mothers, for their unwavering love, support, and understanding throughout this journey. Their encouragement and belief in our abilities have been a constant source of motivation, and their financial support has enabled us to pursue this research project with dedication and commitment. We are deeply grateful to all the individuals and institutions mentioned above for their support and contributions, which have been pivotal in shaping this research paper on "Guardian of Cybersecurity: Unveiling the Tactics in Phishing Website Detection".

References

Ahmed, A. A., & Abdullah, N. A. (2016). Real-time detection of phishing websites. 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 1-6. IEEE. https://doi.org/10.1109/IEMCON.2016.7746247

Aydin, M., & Baykal, N. (2015). Feature extraction and classification of phishing websites based on URL. 2015 IEEE Conference on Communications and Network Security (CNS), 769-770. IEEE. http://dx.doi.org/10.1109/CNS.2015.7346927

Jain, A. K., & Gupta, B. B. (2016). A novel approach to protect against phishing attacks at client side using auto-updated white-list. EURASIP Journal on Information Security, 2016(1), 1-11. https://doi.org/10.1186/s13635-016-0034-3

Jain, S. (2022). Phishing websites detection using machine learning. Available at SSRN 4121102.

Karabatak, M., & Mustafa, T. (2018). Performance comparison of classifiers on reduced phishing website dataset. 2018 6th International Symposium on Digital Forensic and Security (ISDFS), 1-5. IEEE. https://doi.org/10.1109/ISDFS.2018.8355357

Smadi, S., Aslam, N., Zhang, L., Alasem, R., & Hossain, M. A. (2015). Detection of phishing emails using data mining algorithms. 2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), 1-8. IEEE. https://doi.org/10.1109/SKIMA.2015.7399985

Tang, L., & Mahmoud, Q. H. (2021). A deep learning-based framework for phishing website detection. IEEE Access, 10, 1509-1521. https://doi.org/10.1109/ACCESS.2021.3137636

Yang, R., Zheng, K., Wu, B., Wu, C., & Wang, X. (2021). Phishing website detection based on deep convolutional neural network and random forest ensemble learning. *Sensors*, *21*(24), 8281. https://doi.org/10.3390/s21248281

Yang, P., Zhao, G., & Zeng, P. (2019). Phishing website detection based on multidimensional features driven by deep learning. *IEEE access*, *7*, 15196-15209. https://doi.org/10.1109/ACCESS.2019.2892066

Tang, L., & Mahmoud, Q. H. (2021). A survey of machine learning-based solutions for phishing website detection. *Machine Learning and Knowledge Extraction*, *3*(3), 672-694. http://dx.doi.org/10.3390/make3030034