STM JOURNALS

Journal of
Web Engineering & Technology

ISSN: 2455-1880
Volume 11, Issue 3, 2024
September–December
DOI (Journal): 10.37591/JoWET

https://journals.stmjournals.com/jowet

Review JoWET

Detection of Phished URLs Using Machine Learning

Shilpa Mundodagi[1], Nyamatulla Patel[2], Ziaullah Choudhari[3,*]

Abstract

Phishing attacks remain a significant cybersecurity challenge, requiring innovative detection strategies. This study investigates the use of machine learning to detect phishing URLs, to improve the accuracy and reliability of detection systems. Utilizing a diverse dataset of legitimate and phishing URLs we extracted the features such as lexical properties, domain-specific details, and HTML content to train various machine learning models. Algorithms including Random Forest, support vector machine (SVM), and gradient boosting were evaluated for their effectiveness. The results demonstrate that these models can reliably identify phishing URLs with high accuracy and low false-positive rates. This work underscores the potential of machine learning to improve phishing detection mechanisms, contributing to stronger cybersecurity measures. Future work will involve refining these models, investigating deep learning approaches, and integrating real-time detection capabilities.

Keywords: Phishing detection, machine learning, URL analysis, classification algorithms, cybersecurity

INTRODUCTION

The internet has become an important part of our lives in the digital era, allowing us to connect, communicate, and engage in several online activities. However, this convenience also exposes us to various cybersecurity risks such as phishing. Phishing is a tricky plan used by hackers to cheat users into disclosing data such as usernames, passwords, and financial information by posing them as trustworthy organizations. Each year, phishing scams cause "billions of dollars" in losses for consumers. This relates to the methods used by thieves to steal private data from unaware internet users. False emails and phishing software are used by scammers to acquire private data, including personal and financial accounts. ML techniques are used to analyze different elements of legal and phishing URLs to evaluate techniques for identifying phishing websites. The selected machine learning technique for differentiating between fake and safe websites is made easier by fine-tuning parameters. Before attempting to steal sensitive information, criminals frequently work with financial institutions and businesses that handle financial data to create counterfeit versions of real websites and email accounts. The email features genuine company phrases and logos [1].

The ease with which trademarks, brand names, and other business identities can be misused, which customers rely on as methods of verification, and the distinct factors with the internet's explosive rise as a medium for communication. Many people receive "spoof" emails in an effort to enlist their participation in criminal fraud. Whenever these emails are opened or a link on the email is clicked, customers are paid on a counterfeit website that appears to be from a legitimate business. To allow computers to acquire knowledge from facts and make predictions or judgments without explicit programming, machine learning uses statistical

***Author for Correspondence**
Ziaullah Choudhari
E-mail: mdziaullah_ec@secab.org

[1]Student, Department of Master of Computer Application, SECAB Institute of Engineering and Technology, Vijayapur, Karnataka, India
[2]Associate Professor, Department of Master of Computer Application, SECAB Institute of Engineering and Technology, Vijayapur, Karnataka, India
[3]Associate Professor, Department of Electronics and Communication Engineering, SECAB Institute of Engineering and Technology, Vijayapur, Karnataka, India

Received Date: September 12, 2024
Accepted Date: September 23, 2024
Published Date: October 29, 2024

Citation: Shilpa Mundodagi, Nyamatulla Patel, Ziaullah Choudhari. Detection of Phished URLs Using Machine Learning. Journal of Web Engineering & Technology. 2024; 11(3): 1–7p.

models and algorithms. Using "machine learning" models that contain characteristics that distinguish between legitimate websites and phishing websites that are trained on a huge quantity of identifiable data.

LITERATURE SURVEY

Jones A [2]—Investigated to identify phishing attacks, focusing on exploiting human-related vulnerabilities within systems. Users are often the most susceptible point of failure in the safety chain, as cyber-attacks frequently capitalize on the weaknesses found in end users. With this pervasive issue of phishing, a group of strategies is employed to counteract specific attacks with the goal of exploring recently proposed approaches aimed at reducing vulnerabilities. It is important to delineate the appropriate application of phishing detection techniques for various mitigation strategies, including detection, active defense, rectification, and prevention.

Saranga Komanduri [3]—They stated the challenges that users face when making security and privacy decisions in the context of technological advancements. This work explores individual decision-making processes concerning the trade-offs between privacy and information security, as well as the factors influencing these decisions and strategies to overcome decisional constraints.

Overink FJ, Junger M, and Montoya L [4]—They examined the effectiveness of two interventions: using triggering cues to raise awareness about sociotechnical cyber-attack risks and providing cautions against sharing personal information. Their research aimed to protect these interventions and offer users protection against social engineering attacks. Their findings highlight individuals' susceptibility to social engineering attacks and underscore the importance of effective intervention strategies to mitigate such risks.

El-Alfy and El-Sayed [5]—Highlights the increasing frequency and severe consequences of phishing attacks, emphasizing the importance of exploring preventive solutions in information security. Phishing poses various security threats such as information leaks, identity theft, financial losses, and damage to reputations.

OBJECTIVES

Identification of phished URLs using ML aims to enhance internet security by identifying and blocking fraudulent websites designed to steal sensitive information, such as usernames, passwords, and credit card details [6–10]. Here are the primary objectives of this approach:

- *Accuracy:* Develop a model to obtain assured results for legitimate and phishing websites by analyzing various features and patterns.
- *Real-time detection:* Implement the model in a way that allows for real-time detection and prevention of phished-link attempts, thereby providing immediate protection to users.
- *Feature extraction:* Identify and utilize relevant features from websites (e.g., URL structure, HTML content, domain information) that can effectively differentiate phishing sites from legitimate ones.

PROBLEM STATEMENT

Consumers lose billions of dollars annually owing to phishing schemes, which involve deceitful tactics used to obtain private information through deceptive means. Internet fraudsters engage in criminal deception to acquire personal and financial account information from users [11]. Consumers are directed to fake websites that mimic a real company when they open emails or click on a link within them. Cybersecurity comprises a range of technologies and practices designed to protect computers, networks, projects, and information from attacks and unauthorized access, alteration, or destruction [12].

PROPOSED SYSTEM

This idea was created using a website similar to a platform for all users. This flexible and interactive website was designed to determine whether a site is legitimate or a phishing attempt. It was developed

using various web design languages, including HTML, CSS, JavaScript, and Python Flask framework [13]. Websites were created using the basic HTML framework. These effects may be introduced to a website to improve its aesthetic appeal and usability. It is important to keep in mind that because the website is designed for all users, it must be user-friendly. Noting that the dataset lacks website URLs, the proposed system was trained using a dataset with various attributes [14].

When assessing whether a website's URL is genuine or fraudulent, the dataset includes various factors that need to be considered. A Gradient Boosting Classifier was used to obtain the results suggested by the system. The classifier identifies the supplied URL depending on the preparatory information once the system is trained with the dataset. If the site is phishing, it alerts the user that the website is phishing. It notifies users that the website is legitimate. Using a Gradient Boosting Classifier, phishing websites can be identified with a 97% accuracy rate [15].

MOTIVATION

The motivation behind the detection of phishing websites is to prevent financial losses and to protect users' personal information. Phishing attacks cost internet users billions of dollars per year, and the COVID-19 pandemic has increased the use of technology, providing more opportunities for phishers to carry out attacks. Detecting phishing websites is crucial to prevent these attacks, which can have financial, psychological, and professional impacts.

BLOCK DIAGRAM

The block diagram shown in Figure 1 includes the working flow of the model, which includes the dataset and analysis of the features for the best results. In the third stage, the algorithm divides or classifies the website features. Then, it will determine whether a URL is safe [16].

METHODOLOGY

Data Collection

- *Source:* Gather datasets containing both legitimate and phished URLs. These can be obtained from sources such as the Phish Tank, Open Phish, and Alexa Top Sites [17].
- *Preprocessing:* The data are cleaned by removing duplicates, normalizing URLs, and removing irrelevant information.

Feature Extraction

- *URL-based features:* Length of URL, number of dots, presence of IP addresses, use of hyphens, special characters, etc.
- *Domain-based features:* Age of the domain, WHOIS information, domain registration length, etc.
- *Content-based features:* Presence of certain keywords, number of external links, presence of forms, etc.
- *Behavior-based features:* Click-through rates, traffic patterns, user interaction data, etc.

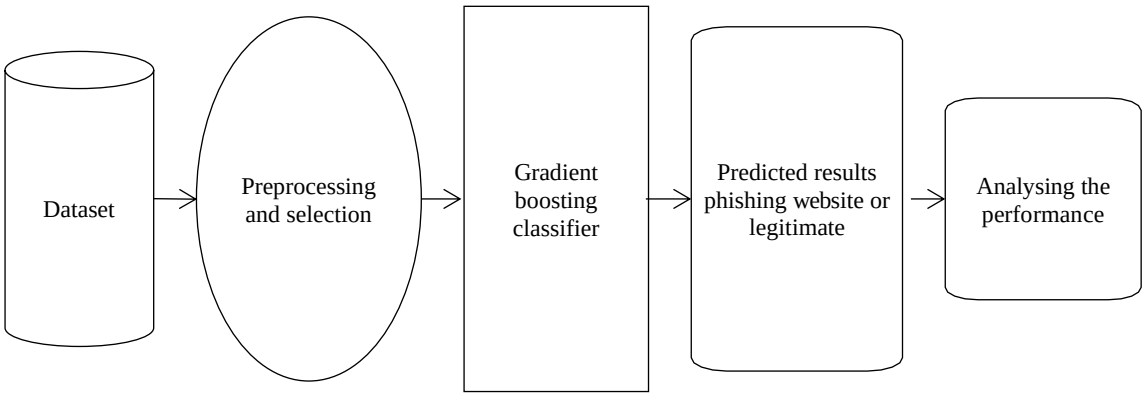

Figure 1. Detailed view of system architecture.

Model Selection

- *Algorithms:* Select appropriate machine learning algorithms, such as logistic regression, decision trees, random forests, support vector machines, and neural networks.
- *Training and testing:* Divide the dataset into training and testing sets (e.g., an 80–20 split) and apply cross-validation techniques to ensure robustness [18, 19].

Model Training

The chosen models were trained using the training dataset. Feature scaling and normalization were applied if needed. Hyperparameters are tuned using methods such as Grid Search or Random Search.

Model Evaluation

The model's performance was assessed using metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC. We compared various models and chose the model with the best performance according to these metrics.

Deployment

Implement the trained model in a real-world setting using frameworks such as Flask and Django. Set up a pipeline for real-time URL analysis and detection.

RESULTS AND ANALYSIS

Figures 2–5 illustrate key stages of the project designed to detect phishing URLs. Figure 2 depicts the home page or login page of the system, serving as the entry point for users. Figure 3 shows the page where users can upload the dataset, which is essential for training the model. Figure 4 presents the results page displaying the safety status of URLs, indicating whether they are safe or not. Similarly, Figure 4 also displays the results of the URL checks, confirming their safety or potential risks.

Figure 2. Home page.

Figure 3. Upload page.

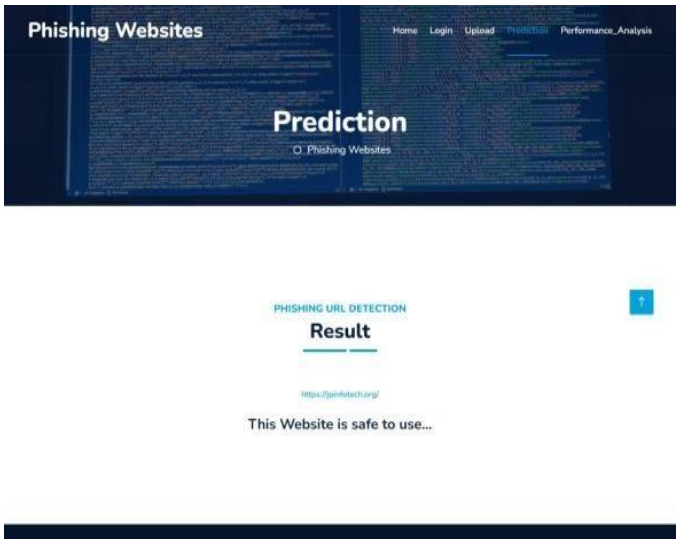

Figure 4. Result page.

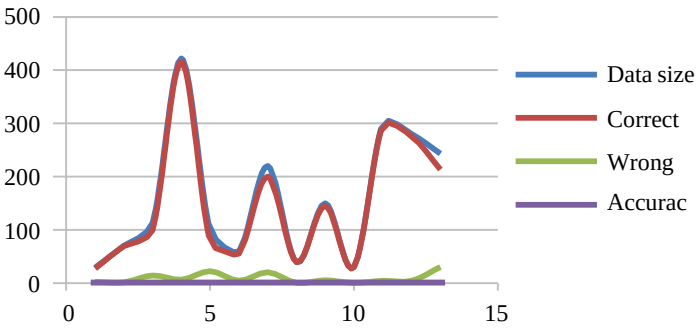

Figure 5. Accuracy graph.

The current method only determines the precision of phishing websites using different types of machine learning approaches. The proposed method generates a useful model for phishing website detection, saves and deploys the model, and then uses the URL to determine whether the website is legitimate or used for identity theft. The authors claim that the presentation of this procedure provides improved accuracy in phishing website detection compared to the previous method. The entire comparison is shown, and the precision of the gradient boosting model is displayed in Figure 5, where a complete comparison is presented.

Extreme Gradient Booster (Also Known as XGBoost)

On the other hand, XGBoost refers to the technical objective of leveraging the computational power of the boosted tree algorithm. A software library called XGBoost is accessible through interfaces after it is downloaded, installed, and used.

CONCLUSION AND FUTURE WORK

A good anti-phishing system should be able to recognize phishing attacks in real time. To improve phishing site detection, it is also crucial to know when powerful anti-phishing technologies should be deployed. The Gradient Boosting Classifier is the only system that is currently used to identify phishing websites. Using the Gradient Boosting Classifier, we achieved a low false-positive rate and 97% detection accuracy.

While relying solely on the URL components proved to be more effective, phishers discovered how challenging it was to forecast where a URL would go. As a result, they carefully changed the URL to prevent detection, such as hosting, nearly all efficient methods.

To swiftly identify novel phishing assault patterns and enhance the accuracy of our models through feature extraction, we want to extend the phishing detection system to be a customizable web service coupled with online learning in future improvements.

In the future, machine learning could play a major role in enhancing the identification of phishing websites. The following are some potential advancements that have been made:

- *Feature engineering:* Under the conditions of phishing website detection, future enhancements could involve the growth of more sophisticated features that capture the subtle characteristics of phishing attempts. These features may include analyzing the HTML structure, JavaScript behavior, SSL certificates, URL patterns, and text content.
- *Deep learning architectures:* Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have demonstrated outstanding performance in various fields. Future research could investigate how these architectures can enhance the accuracy of phishing website detections. Deep learning models can grasp complex patterns and relationships within data, enabling them to identify subtle indicators of phishing attempts.
- *Large-scale datasets:* The availability of large-scale, diverse, and up-to-date datasets is crucial for training effective machine learning models. Future enhancements should focus on creating comprehensive datasets that encompass various types of phishing attacks and incorporate emerging attack vectors. Incorporating real-time data from reliable sources, such as security companies and community reports, provides information for training and updating models.
- *Transfer learning and domain adaptation:* Transfer learning techniques can enhance phishing website detection by leveraging pre-trained models from similar tasks. By transferring knowledge from general cybersecurity domains or related tasks, such as junk detection or malware analysis, models can learn from existing expertise and modify the specific characteristics of phishing attacks.
- *Real-time detection:* Promptness is essential for the detection and addressing of phishing attacks. Future enhancements should focus on developing real-time detection systems that can analyze websites within milliseconds and provide immediate feedback to users. This can involve the use of lightweight models or model compression techniques to ensure efficient inferences on various platforms and devices.
- *Collaborative filtering and crowdsourcing:* Building a collective intelligence system that leverages the wisdom of a crowd can enhance the accuracy of phishing website detection. Future enhancements may involve incorporating collaborative filtering techniques to aggregate feedback and insights from users, security experts, and other trusted sources. The identification procedure can be aided using this common knowledge to generate reputation ratings or labels for websites.
- *Adversarial robustness:* Adversarial attacks, where malicious actors attempt to manipulate or evade detection systems, pose a significant challenge at cyber security sites. Future enhancements should focus on the development of ML models. Techniques, such as adversarial training, model assembly, and anomaly detection, can also help improve the resilience of phishing detection systems.

REFERENCES

1. Sheng S, Wardman B, Warner G, Cranor LF, Hong J, Zhang C. Empirical analysis of phishing blacklists. In: 6th International Conference on Email and Anti Spam; 2009 Jul 6; Mountain View, California. p. 16–17.
2. Khonji M, Iraqi Y, Jones A. Phishing detection: A literature survey. IEEE Commun Surv Tutor. 2013;15:2091–2121. doi:10.1109/SURV.2013.032213.00009.
3. Acquisti A, Adjerid I, Balebako R, Brandimarte L, Cranor LF, Komanduri S, Leon PG, Sadeh N, Schaub F, Sleeper M, Wang Y, Wilson S. Nudges for privacy and security: understanding and assisting users' choices online. ACM Comput Surv. 2018;50:1–41. doi:10.1145/3054926.
4. Junger M, Montoya L, Overink FJ. Priming and warnings are not effective to prevent social engineering attacks. Comput Hum Behav. 2017;66:75–87. doi:10.1016/j.chb.2016.09.012.
5. El-Alfy ESM. Detection of phishing websites based on probabilistic neural networks and K-medoids clustering. Comput J. 2017;60:1745–1759. doi:10.1093/comjnl/bxx035.

6. Moreno-Fernández MM, Blanco F, Garaizar P, Matute H. Fishing for phishers: improving internet users' sensitivity to visual deception cues to prevent electronic fraud. Comput Hum Behav. 2017;69:421–436. doi:10.1016/j.chb.2016.12.044.
7. Kamalam GK, Suresh P, Nivash R, Ramya A, Raviprasath G. Detection of phishing websites using machine learning. In: 2022 International Conference on Computer Communication and Informatics (ICCCI); 2022. p. 1–4. doi:10.1109/ICCCI54379.2022.9740763.
8. Beknazarova SS, Latipova NM, Maxmudova MJ, Alekseeva VS, Turakulova AS. Machine learning algorithms are used to detect and track objects on video images. In: Third International Conference on Optics, Computer Applications, and Materials Science (CMSD-III 2023); 2024. Vol. 13065. SPIE. doi:10.1117/12.3025015.
9. Deshpande A, Pedamkar O, Chaudhary N, Borde S. Detection of phishing websites using machine learning. Int J Eng Res Technol (IJERT). 2021;10.
10. Wu L, Du X, Wu J. Effective defense schemes for phishing attacks on mobile computing platforms. IEEE Trans Veh Technol. 2015;65:6678–6691. doi:10.1109/TVT.2015.2472993.
11. Gowtham R, Krishnamurthi I. A comprehensive and efficacious architecture for detecting phishing webpages. Comput Secur. 2014;40:23–37. doi:10.1016/j.cose.2013.10.004.
12. Xiang G, Hong J, Rose CP, Cranor L. CANTINA+: A feature-rich machine learning framework for detecting phishing web sites. ACM Trans Inf Syst Secur. 2011;14:1–28. doi:10.1145/2019599.2019606.
13. Zhu E, Chen Y, Ye C, Li X, Liu F. OFS-NN: An effective phishing websites detection model based on optimal feature selection and neural network. IEEE Access. 2019;7:73271–73284. doi:10.1109/ACCESS.2019.2920655.
14. Dou Z, Khalil I, Khreishah A, Al-Fuqaha A, Guizani M. Systematization of knowledge (SoK): A systematic review of software-based web phishing detection. IEEE Commun Surv Tutor. 2017;19:2797–2819. doi:10.1109/COMST.2017.2752087.
15. Cova M, Kruegel C, Vigna G. Detection and analysis of drive-by-download attacks and malicious JavaScript code. In: Proceedings of the 19th International Conference on World Wide Web; 2010 Apr 26. p. 281–290. doi:10.1145/1772690.1772720.
16. Tan CL, Chiew KL, Sze SN. Phishing website detection using URL-assisted brand name weighting system. In: Proceedings of the 2014 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS); 2014. p. 54–59. doi:10.1109/ISPACS.2014.7024424.
17. Basnet RB, Sung AH. Mining web to detect phishing URLs. In: 11th International Conference on Machine Learning and Applications; 2012 Dec 12; Vol. 1. IEEE Publications. p. 568–573. doi:10.1109/ICMLA.2012.104.
18. Kurniabudi K, Purnama B, Sharipuddin S, Darmawijoyo D, Stiawan D, Samsuryadi S, Heryanto A, Budiarto R. Network anomaly detection research: A survey. Indones J Electr Eng Inform. 2019;7:37–50. doi:10.52549/ijeei.v7i1.773.
19. Nguyen LA, To BL, Nguyen HK, Nguyen MH. A novel approach for phishing detection using URL-based heuristic. In: 2014 International Conference on Computing, Management and Telecommunications (ComManTel); 2014. p. 298–303.

