Viable Detection of URL Phishing using Machine Learning Approach

Machikuri Santoshi Kumari[1*], *Chiguru* Keerthi Priya[1], *Gondhi* Bhavya[1], *Haridas* Neha[2], *Monisha* Awasthi[2], *Surendra* Tripathi[3]

[1]Department of CSE (DS), GRIET, Hyderabad, Telangana State, India
[2]Uttaranchal Institute of Technology, Uttaranchal University, Dehradun, 248007, India
[3]KG Reddy College of Engineering & Technology, Hyderabad, India

Abstract. The objective of paper is to detect phishing URLs using machine learning algorithms. Phishing is a fraudulent activity that involves tricking users into giving away sensitive information, such as passwords and credit card numbers, by impersonating legitimate websites. The main objective of this work is to build a model that can accurately detect viable phishing URLs and classify them as either legitimate or fraudulent. This will help to prevent users from falling victim to phishing attacks and protect their personal information. The model will be trained on a large dataset of annotated URLs and will be optimised for high accuracy and low false positive rates. The paper consists of two datasets in which one of the dataset consists of phishing URLs and other datasets consist of features of URLs. The performance of the phishing detection model will be evaluated using various metrics, such as precision, recall, and F1 score. We will also conduct an in-depth analysis of the results and discuss the effectiveness of the approach. This work aims to build a robust model for phishing URL detection using machine learning algorithms. Future enhancements to this work could include incorporating more advanced feature extraction techniques, exploring the use of deep learning models, and expanding the dataset to include more diverse types of URLs.

1 Introduction

1.1 Introduction

Viable Phishing URL detection is crucial in cybersecurity, aiming to identify and block deceptive URLs used for phishing attacks. These attacks employ social engineering to trick users into divulging sensitive data. Detection involves methods like URL structure analysis, content inspection, domain examination, SSL certificate checking, and machine learning for pattern identification in large datasets. Phishing URL identification is crucial due to the persistent threat of attacks targeting individuals and organizations. Machine learning methods are employed in this research to discern legitimate and fraudulent URLs, aiming to safeguard personal information and prevent phishing scams. The paper employs feature

*Corresponding author: machikuri.santoshikumari@gmail.com

extraction and ML algorithms on a large dataset, aiming for high accuracy and minimal false positives. Future enhancements could involve advanced feature techniques and deep learning. Once a phishing URL is detected, it can be blocked or flagged to prevent users from accessing the malicious website. Phishing URL detection is an essential component of cybersecurity, as phishing attacks continue to be a prevalent threat to businesses and individuals alike. As phishing attacks become increasingly sophisticated, detection systems must continually evolve to keep up with the evolving threat landscape.

1.2 History

Phishing URL detection has evolved over time as a response to the growing threat of online phishing attacks. In the early days of the internet, phishing attacks were relatively rare. The concept of phishing emerged in the mid-1990s, where attackers would send fraudulent emails or messages to trick users into revealing sensitive information. Initially, it was difficult to detect phishing URLs because attackers often used simple HTML tricks to create convincing but malicious websites. As phishing attacks became more prevalent, security experts started developing techniques to combat them. One of the early approaches was to maintain blacklists of known phishing URLs. Websites and email providers would compare the URLs in incoming messages against these lists and block or warn users about potentially malicious links.

2 Existing methods

2.1 Literature survey

The study surveys pertinent aspects and approaches to evaluate the use of machine learning in identifying and detecting phishing websites. Attackers like phishing because it makes it easier to trick people into clicking on links that look genuine but are actually dangerous than to directly get past computer defenses. The website not only empowers users to identify fraudulent websites but also raises awareness about the ongoing malpractices. By utilizing the website, individuals can protect their personal information, such as email addresses, passwords, debit card numbers, credit card details, CVV codes, bank account numbers, and more, from exploitation. The study emphasizes that the task of real-time phishing website detection and identification is a complex and dynamic one, involving numerous variables and requirements. Fuzzy logic approaches are crucial for recognising and assessing phishing websites due to the detection's inherent ambiguities.

In order to increase effectiveness and accuracy in recognising phishing assaults, the paper provides a novel model for detecting phishing attempts using machine learning techniques. The websites from the PhishTank dataset are parsed to apply this model using a separate programme that gathers and compiles features.

The document offers a thorough analysis of the available research on the many techniques for identifying phishing websites. It admits the shortcomings of current anti-phishing programmes and provides a comparative analysis of them. The research paper describes a model that combines the URL detection method and the Random Forest algorithm to identify phishing websites. A dataset from PhishTank was acquired to carry out the complete method. The initial phase involved parsing, which seeks to analyse the feature set given the volume of input. A two-part Attribute Subset Selector is an element of the parsing process.

The paper presents a simple framework for phishing website detection. The method involves matching URLs using the quick Minhash signature. The authors of a different

work, suggest a method that makes use of several classifiers (ensemble learning) to forecast the class probabilities of URLs. The classifiers' judgements are then filtered using a threshold.

The article proposes a phishing detection method that uses only nine lexical features to produce reliable findings. The ISCX URL-2016 dataset, which included 11,964 instances of legal and phishing URLs, was used in the study's experiment. The study suggests a novel PhishZoo phishing detection method. This technique makes use of profiles of legitimate websites' visual looks to spot phishing websites. The method yields a high accuracy of 96%, comparable to blacklisting techniques, but has the extra benefit of being able to identify targeted attacks on smaller websites, such as corporate intranets, and zero-day phishing attacks.Author [11] devised efficient methods to detect and avoid wormhole attacks in the BSR protocol.Author [12] discussed possible attacks on BSR protocol in network. These attacks are difficult to be detected. To avoid the attacks on BSR protocol, proposed two schemes namely Reverse Routing Scheme (RRS) and Authentication of Nodes Scheme (ANS).Authors [13] presented the key issues of data security in the cloud computing technology, and are described in four areas: storage security, network security, data security and virtualization.Authors [14] has discussed the importance of data sharing and the necessity to ensure security and privacy.

Table 1. Summary of literature survey.

Ref. No.	Methodology	Results	Drawbacks
[1]	Random Forest And decision tree classifier	Random forest algorithm with 97.31% and Decision tree with 95%	Random forest and decision trees has the potential for limited generalization
[2]	Used fuzzy logic and machine learning algorithms	Accuracy of 95.6%	Require significant computational resources, potentially resulting in slower processing speeds for real-time phishing detection
[3]	Logistic regression, XGBoost algorithm	Accuracy of 92%	Too specific to the training data and struggle to generalize to new, unseen phishing instances
[4]	Random Forest	They used 3 datasets and achieve an accuracy of 96.92%, 99.77% and 89.73%	The increased complexity and computational overhead introduced by combining multiple algorithms
[5]	Random Forest	Accuracy of 90%	Few features are missed in the dataset
[6]	Random Forest SVM, KNN	98% with Random forest, 97% with KNN, 96% with SVM	-
[7]	Naive Bayes Classifier	The Naive Bayes classifier model with 97% Accuracy	The data set used in this paper is little old and it needs regular updates
[8]	Ensemble Learning	NA	The complexity of ensemble learning, which may hinder real-time performance and scalability of the system.
[9]	Random Forest	Accuracy of 99.57%	Lexical-based approaches may struggle to capture more nuanced features and patterns in the URLs.
[10]	Computer Vision	Accuracy of 96%	Malicious actors can manipulate the visual elements of the phishing websites to evade detection

3 Problem statement and objectives

3.1 Problem statement

The paper aims to establish a robust phishing URL detection system by integrating a dependable blacklist and the XGBoost algorithm. The process involves collecting an updated blacklist from diverse sources, containing URLs associated with phishing. This blacklist is then matched against input URLs for potential phishing identification. XGBoost, a proficient machine learning algorithm, is employed for classification. A dataset with labeled URLs is utilized to train the model, considering features like domain traits, URL length, and keywords. The trained model predicts the legitimacy and viability of new URLs based on extracted features, enhancing overall detection accuracy.

3.2 Objectives

Phishing URL detection is pivotal for shielding brand reputation and data security. It prevents fraudsters from executing successful phishing attacks and stealing sensitive information, including usernames and passwords. Such attacks can also spread malware, jeopardizing users' systems. Various methods like blacklisting, machine learning (e.g., XGBoost), and content analysis identify and block phishing URLs. Real-time protection, continuous updates, user-friendly interfaces, and minimizing false positives are vital considerations. Overall, the goal is to provide efficient and comprehensive protection against evolving phishing threats, safeguarding individuals and organizations from financial and security risks.

4 Proposed method

4.1 Description

Detecting phishing URLs is vital for cybersecurity. Combining blacklist methods with XGBoost, a robust machine learning classifier, enhances phishing URL identification. Blacklists contain known malicious URLs, while XGBoost handles complex data effectively. This hybrid approach boosts accuracy and safeguards users against phishing scams.

4.2 Architecture diagram and its detailed explanation

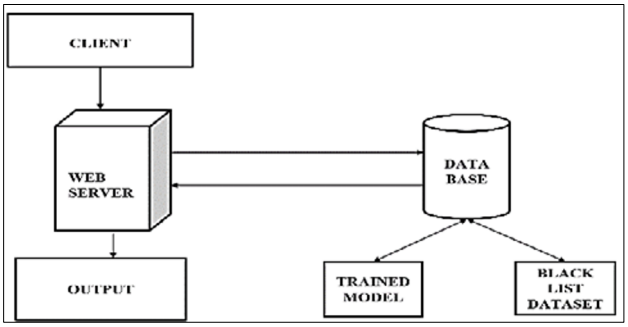

Fig. 1. User interface of phishing URL detection.

In phishing URL detection using XGBoost and blacklisting, the client connects to the server. The server queries the database; if data is in the blacklist, results are displayed; if not, the XGBoost model is trained and results shown. Collecting URL data and extracting features differentiate genuine and phishing URLs. The XGBoost model learns from labeled URLs and the blacklist identifies harmful ones. Real-time detection compares URLs to blacklist; if safe, features are input to the XGBoost model for analysis. Regular blacklist updates ensure dynamic protection against evolving phishing attacks.

4.3 Modules and its description

Depending on the precise method and implementation employed, different phishing URL detection modules may be available. However, the descriptions of a few typical modules used in phishing URL detection systems are provided below.

- Data collection Module This module is in charge of gathering information from a variety of sources, including phishing reports, social media posts, and email correspondence. The collected data may include URLs, website content, and metadata.
- URL Pre-processing Module In this module, the collected URLs are cleaned pre-processed to remove any unwanted characters or symbols. This is done to ensure that the URLs are in a standardized format for further processing.
- Feature Extraction Module This module collects a number of characteristics from URLs, including the length of the URL, the number of subdomains, the presence of special characters, and the age of the domain. Following that, the URLs are categorised as being legitimate or phishing using these features.
- Blacklist Module: The Blacklist module maintains a list of known phishing URLs obtained from various sources such as public databases, security companies, and security researchers. The URLs are checked against this list, and if a match is found, the URL is labelled as a Phishing URL.
- Machine Learning Module: The Machine Learning module categorises the URLs as authentic or phishing using a variety of supervised learning methods, including logistic regression, decision trees, and support vector machines (SVMs).
- Front-End Web Interface Module: This module deals with the development of the user interface (UI) components. It includes HTML, CSS, and JavaScript files responsible for rendering the frontend web page where users can interact with the system. It includes taking URLs from user and display output whether phishing URL or not.
- Backend Integration Module This module handles the integration of the front-end web interface with the backend components. It includes components responsible for routing and handling HTTP requests from the front-end UI to the appropriate backend modules.
- Integration Module This module integrates the Flask framework to connect the front-end UI, backend modules, and data processing components. It includes components responsible for handling HTTP requests, routing, and coordinating the flow of data between different modules.

5 Results and discussions

5.1 Description about dataset

The paper employs two datasets: one contains phishing URLs, while the other holds URL features. The blacklist phishing URL dataset comprises known phishing URLs sourced

from initiatives like Phishtank and Open Phishtank. Serving as a valuable training resource, it covers diverse phishing attacks like spoofed sites and malicious downloads. Collated from reputable sources, duplicate and broken URLs were removed, and normalization applied for consistency. The model is trained on the second dataset with URL features extracted to capture patterns associated with phishing URLs, enhancing detection accuracy.

- Domain-related Features:
- Host and IP-based Features Hosting Provider:
- URL Structure
- Content-related Features
- SSL Certificate Features

5.2 Detailed explanation about experimental results

Xgboost, Randomforest, Decision Tree, Auto Encoder, and Multilayer Perceptron were among the machine learning techniques that were taken into consideration for this paper. The most accurate of these was Xgboost.

	ML Model	Train Accuracy	Test Accuracy
3	XGBoost	1.000	0.967
1	Random Forest	0.936	0.925
0	Decision Tree	0.925	0.905
2	Multilayer Perceptrons	0.866	0.880
4	AutoEncoder	0.000	0.000

Fig. 2. ML model accuracy.

The below figure depicts the some of the important features of the URLs like HTTPS, Anchor URL, Website Traffic, subDomains etc.

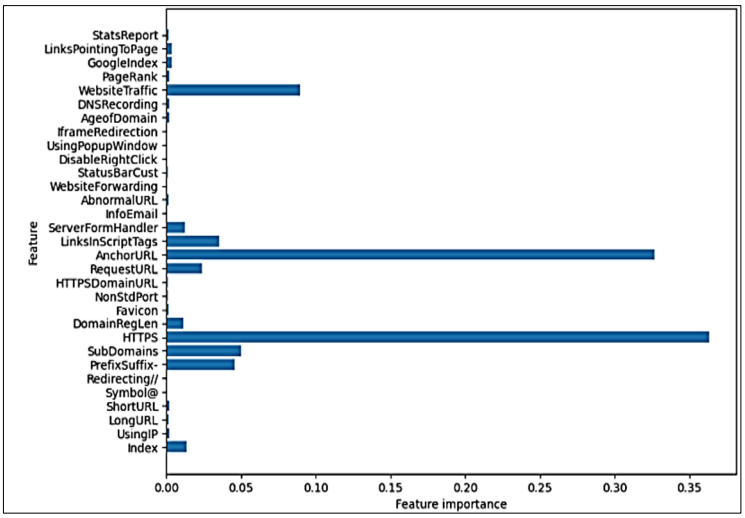

Fig. 3. Important features of URL's.

The figure shows the user interface of phishing URL Detection Where the user should enter the URLs to know whether it is legitimate or fraudulent URL.

S

Fig. 4. User interface of phishing URL detection.

In the below figure a URL was entered to depict whether it is Phishing or not.

Fig. 5. Entry of the URL in phishing URL detection.

Figure below shows the result of the entered URL and here it shows that the entered URL is safe.

Fig. 6. Result of the entered URL.

In the below figure the user entered the URL in the Phishing URL Detection webpage.

Fig. 7. Entered URL.

The URL which was entered in the above figure identified as unsafe because it was found in blacklist phishing URL detection.

Fig. 8. Result of entered URL.

The below figure shows the URL which is entered to know whether it is safe or not.

Fig. 9. URL entered.

The below Figure shows that URL which is entered in the above figure is identified as phishing URL by the model based on the feature of the URL.

Fig. 10. Result of the entered URL.

5.3 Significance of proposed method with its advantages

Viable Phishing URL detection's importance lies in shielding users from online threats. Leveraging blacklists and the XGBoost algorithm enhances this process. Blacklists, curated by security experts, swiftly detect potential phishing URLs, offering proactive defense by comparing URLs and preventing access to harmful sites. Blacklists contain reported phishing URLs, providing protection against known threats and reducing susceptibility to attacks. Regular updates by experts maintain current detection capabilities. XGBoost, a potent machine learning algorithm, enhances phishing URL detection. It highlights significant features for classification and employs an ensemble of decision trees for accurate predictions, countering overfitting. Addressing class imbalance, XGBoost optimizes predictions on extensive datasets. Its efficacy in handling unbalanced data and optimal performance contributes to robust phishing defense.

6 Conclusion and future enhancement

Using the XGBoost algorithm for phishing URL detection brings benefits such as improved prediction accuracy, balanced handling of class imbalances, and real-time processing. Its ensemble of decision trees reduces overfitting, while weight-based adjustments address class imbalance. XGBoost's scalability and parallel processing enable quick training, vital for real-time detection. Hyperparameter customization and feature importance analysis enhance adaptability. The algorithm's ensemble learning and imbalanced data handling offer accuracy and fairness. Additionally, integration with blacklist datasets fortifies defense mechanisms against phishing attacks, minimizing false positives. This approach combines XGBoost's power with blacklist insights for robust, real-time protection, mitigating the risk of falling victim to phishing.

Combining blacklists and XGBoost algorithm enhances phishing URL detection. Blacklists promptly block known malicious URLs, while XGBoost's advanced feature analysis and ensemble learning improve accuracy and adapt to evolving threats. This multi-layered approach strengthens defense against phishing attacks, safeguarding user security by swiftly blocking malicious sites and accurately identifying new threats.

Phishing URL detection is crucial for thwarting cyberattacks. Future improvements encompass content scrutiny to identify plagiarism, detecting peculiar URL characters, and assessing URL structures. API-driven blacklist updates enhance efficiency, while user

behavior analysis and real-time website monitoring counter social engineering. By combining these strategies with evolving technology and security protocols, viable phishing URL recognition can be fortified, safeguarding users against deceptive assaults in the ever-evolving digital landscape.

References

1. S. Alrefaai, G. Özdemir, A. Mohamed, *Detecting Phishing Websites Using Machine Learning*, in Proceedings of the 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey (2022)
2. M. D. Bhagwat, P. H. Patil and T. S. Vishawanath, A MethodicalOverview on Detection, Identification and Proactive Prevention of PhishingWebsites, in Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India (2021)
3. P. Bhavani, Amba, Chalamala, Madhumitha, Likhitha, Sree Sai, C. P. Sai, Intl. J. Appl. Res. Tech **8**, 2511 (2022)
4. Chenguang Wang, Yuanyuan Chen, Knowl. Based Sys. **258**, 109955 (2022)
5. S. Parekh, D. Parikh, S. Kotak and S. Sankhe, *A New Method for Detection of PhishingWebsites: URL Detection*, in Proceedings of the 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India (2018)
6. A. Saleem Raja, R. Vinodini, A. Kavitha, J. Adv. Appl. Scien. Res **5**, 4 (2021)
7. J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran, B. S. Bindhumadhava, *PhishingWebsite Classification and DetectionUsing Machine Learning*, in Proceedings of the 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India (2020)
8. D. K. Mondal, B. C. Singh, H. Hu, S. Biswas, Z. Alom, M. A. Azim, J. Inform. Secu. Appl **62** (2021)
9. B. B. Gupta, K. Yadav, I. Razzak, K. Psannis, A. Castiglione, X. Chang, Comp. Communi. **175** (2021)
10. S. Afroz, R. Greenstadt,*PhishZoo: Detecting Phishing Websites by Looking at Them,* in Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing, Palo Alto, CA, USA (2011)
11. E.Poornima, C. Bindhu,Intl. J. Comp. Netwo. Secur **3**, 1(2010)
12. E. Poornima, C. S. Bindu and S. K. Munwar, *Detection and Prevention of Layer-3 Wormhole Attacks on Boundary State Routing in Ad Hoc Networks*, 2010 International Conference on Advances in Computer Engineering, Bangalore, India(2010)
13. E.Poornima, N.Kasiviswanath, C.S.Bindu, Ind. J. Sci. Tech **10**, 19 (2017)
14. E.Poornima, N.Kasiviswanath, &C.ShobaBindu*Secured Data Sharing in Groups using Attribute-Based Broadcast Encryption in Hybrid Cloud*,in Proceedings of the Emerging Trends in Expert Applications and Security. Advances in Intelligent Systems and Computing, Springer, Singapore, **841** (2019)