ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

URL PHISHING DETECTION USING MACHINE LEARNING

G Bhavya Teja
21N81A6261
Computer Science and Engineering
(Cyber Security)
Sphoorthy Engineering College,
Nadergul, Hyderabad,501510
bhavyateja.gntp@gmail.com

M Vishal
21N81A6284
Computer Science and Engineering
(Cyber Security)
Sphoorthy Engineering College,
Nadergul, Hyderabad,501510
vishalmavilla1947@gmail.com

V Gayathri Devi
21N81A6260
Computer Science and Engineering
(Cyber Security)
Sphoorthy Engineering College,
Nadergul, Hyderabad,501510
gayathridevi052003@gmial.com

K Amarnath
21N81A6272
Computer Science and Engineering
(Cyber Security)
Sphoorthy Engineering College,
Nadergul, Hyderabad,501510
amarnathchowan18@gmail.com

Mr.G.Mukesh
Assistant Professor
Computer Science and Engineering (Cyber Security)
Sphoorthy Engineering College,
Nadergul, Hyderabad,501510

ABSTRACT: Phishing is a type of cybersecurity attack that involves stealing personal information such as passwords, credit card numbers, etc. To avoid phishing scams, we have used Machine learning techniques to detect Phishing Websites. Therefore, in this paper, we are trying to find the total number of ways to find Machine Learning techniques and algorithms that will be used to detect these phishing websites. We are using different Machine Learning algorithms such as KNN, Naive Bayes, Gradient boosting, and Decision Tree to detect these malicious websites. The research is divided into the following parts. The introduction represents the focused zone, techniques, and tools used. The Preliminaries section has details of the preparation of the information that is required to move further. Later the paper emphasizes the detailed discussion of the sources of information. Keywords—Algorithms, Cybersecurity, Machine Learning, Phishing

1. INTRODUCTION

In contemporary times, it has become increasingly simple for cybercriminals to establish counterfeit websites that closely resemble legitimate ones. Consequently, phishing has emerged as a significant concern for security researchers. While experts possess the skills to identify fraudulent websites, not all users are equipped to recognize them, making them susceptible to phishing attacks. Although the internet has introduced unparalleled convenience for individuals managing their finances and investments, it simultaneously offers opportunities for large-scale fraud at minimal cost to the perpetrators. Fraudsters exploit victims by gathering information under the guise of legitimacy, rather than relying solely on hardware or software systems that feature advanced security measures. Phishing ranks among the most prevalent forms of internet fraud, primarily targeting the theft of sensitive personal information, including passwords and credit card details. Phishing attacks manifest in two primary forms: one involves deceiving victims into disclosing their sensitive information by impersonating trustworthy entities with a legitimate need for such data, while the other seeks to obtain secrets by installing malware on victims' devices. The specific malware utilized in phishing attacks is a subject of ongoing research within the virus and malware community and is not the focus of this thesis. This thesis will concentrate on phishing attacks that successfully deceive users, and the term 'phishing attack' will be employed to describe this particular type of assault. Phishing URLs may be disseminated to consumers via emails, instant messages, or text messages. In this study, we will implement the Gradient Boosting Classifier Algorithm to assess the safety percentage of websites. Our model encompasses both frontend and backend components. For the frontend, we are utilizing HTML and CSS, while the backend is developed in Python. We will extract 30 features, which will be analyzed by the machine to predict the safety percentage of the website.

2. LITERATURE REVIEW

[1] Mohammed Hazim Alkawaz and Stephanie Joanne Steven (2021) present a paper titled "A Comprehensive Survey on Identification and Analysis of Phishing Websites Based on Machine Learning Methods." In this study, the authors introduce a hybrid approach known as Phish-Alert, which integrates content similarity whitelists, style similarity, and heuristics. When compared to existing algorithms such as CANTINA and CANTINA+, Phish-Alert demonstrated superior performance on an experimental dataset comprising 500 phishing sites and 500 legitimate sites. However, the effectiveness of the Phish-Alert model diminishes as the dataset size increases. To enhance detection efficiency, the authors suggest incorporating additional functions into future iterations of the Phish-Alert algorithm. Furthermore, the authors propose a phishing detection method utilizing the Resource Description Framework (RDF) and Random Forests. This method consists of a two-level process: the first level is based on the RDD model of webpages, while the second level employs machine learning techniques. Both levels work in tandem to minimize false positives, thereby improving the system's accuracy and precision. The generation of RDF from Hypertext Markup Language occurs after the extraction of features from suspicious webpages. A total of 21 features were selected, ensuring that no similar webpages share the same element sent. Additional vocabularies, including Extensible Hypertext Markup Language (XHTML), HTTP, and Dublin Core, have also been incorporated. Among the various machine learning algorithms, Random Forest has exhibited the best performance in classification, demonstrating high accuracy even in the presence of less sensitive outliers and missing values in parameter selections.

[2]Almaha Abuzuraiq, Mouhammad Alkasassbeh, and Mohammad Ali (January 2020) present the study titled "Intelligent methods for accurately detecting phishing websites." In the development of phishing detection systems, two critical components are the algorithms used to construct the model and the dataset employed for its training and testing. The primary objective of this research is to create a phishing detection system utilizing a fuzzy logic algorithm. Consequently, the dataset's validity will be confirmed by testing it across various machine learning algorithms. Subsequently, different feature selection methods will be applied to this dataset to improve the model's performance. Following this, four distinct fuzzy logic algorithms will be implemented on the same dataset. Ultimately, the experimental outcomes from these approaches will be compared and analyzed. The dataset utilized in this study comprises 5000 phishing websites and 5000 legitimate websites. The development of this model follows four key steps: i) Feature selection: This paper employs two algorithms for the dataset, namely Info-gain and Relief-F, both of which are filter-type methods. The top 15 features identified by each algorithm have been taken into account. ii) Model evaluation: The model is assessed using the accuracy equation, as the dataset is binary and balanced. iii) Machine learning experimental results. iv) Application of fuzzy logic.

[3] J. James, Sandhya L. and C. Thomas, “Detection of phishing URLs using machine learning techniques,” International Conference on Control Communication and Computing (ICCC), December 2013: The proposed system employed a method that utilized lexical features, host properties, and characteristics related to the webpage for identifying phishing websites. To gain a comprehensive understanding of the URL patterns, various data mining algorithms were applied. The classification algorithms evaluated included Naïve Bayes, J48 Decision Tree, K-NN, and SVM for the identification of phishing websites. The Decision Tree demonstrated superior accuracy of 91.08% in comparison to the other algorithms. Therefore, Tree-based classifiers are most effective for the classification of phishing URLs.

3. METHODOLOGY or Existing methods

Ref. No	Methodology	Results	Drawbacks
1	Random Forest And decision tree classifier	The random forest algorithm achieved an accuracy of 97.31%, while the decision tree reached 95%.	Both random forests and decision trees exhibit limited generalization capabilities.
2	Utilized fuzzy logic alongside machine learning algorithms.	Achieved an accuracy of 95.6%.	This approach necessitates considerable computational resources, which may lead to slower processing speeds for real-time phishing detection.

3	Logistic regression and the XGBoost algorithm.	Achieved an accuracy of 92%.	These methods tend to be overly specific to the training data, making it difficult to generalize to new, unseen phishing instances.
4	Random Forest	Utilized three datasets ,achieving accuracies of 96.92%, 99.77%, and 89.73%.	The combination of multiple algorithms introduces increased complexity and computational overhead.
5	Random Forest, SVM, KNN.	Achieved 98% accuracy with Random Forest, 97% with KNN, and 96% with SVM.	-
6	Naive Bayes Classifier.	The Naive Bayes classifier model achieved an accuracy of 97%.	The dataset utilized in this study is somewhat outdated and requires regular updates.
7	Ensemble Learning.	Not applicable.	The complexity inherent in ensemble learning may impede real-time performance and the scalability of the system.
8	Computer Vision	Accuracy of 96%	Malicious actors can manipulate the visual elements of the phishing websites to evade detection

PROPOSED SYSTEM

In our project, the dataset comprises 30 lexical features. Analyzing these lexical features allows us to capture properties for classification purposes. We begin by distinguishing the two components of a URL: the host name and the path, from which we extract a bag of words. Our research has revealed that the primary differences between phishing

and legitimate URLs include the length of the URL, the number of tokens, a greater number of tokens in the domain path, and more levels (delimited by dots) in the case of phishing. Furthermore, the structure of URLs for phishing or malware websites is often designed to resemble benign ones by incorporating popular brand names as tokens, apart from those in the Second Level domain. In instances of phishing and malware websites, attackers may utilize an IP address to obscure the suspicious elements of the URL, which is not a practice seen in legitimate cases. Additionally, phishing URLs tend to include several suggestive word tokens (such as confirm, account, banking, secure, ebayisapi, login, signin); we assess the presence of these security-sensitive words and incorporate their binary values into our features. Intuitively, malicious sites are less popular than benign ones, making site popularity a significant feature. Traffic rank, one of the features obtained from Alexa.com, supports this observation. Since malicious websites are often registered with less reputable or recognizable hosting centers or regions, host-based features are derived from this insight. Our dataset is categorized as a regression problem. Therefore, the supervised machine learning model selected for training the dataset in our project is Gradient Boosting. We have developed a web page that assists users in determining the safety of opening a URL by providing a safety percentage. The code used to create our web page includes an anchor tag that takes the link provided by the user, displaying it on the web page based on the safety percentage. If the user chooses to open the URL by clicking the link, they are directed to the corresponding website. Consequently, the web page in our project can also serve as a source for opening links, similar to Chrome, Firefox, etc., with the added feature of safety assessment.

4. IMPLEMENTATION

I. Start the procedure by accepting a URL as input.

II. Generate rules for extracting specific features from the URL. These rules are designed to identify characteristics commonly associated with phishing, such as excessive use of special characters, presence of IP addresses, or abnormal URL lengths.

III. Extract URL-based features (UF1 to UF20): Analyze the structure and content of the input URL to generate 20 features. These may include:

a. Length of the URL
b. Use of HTTPS
c. Presence of suspicious symbols (e.g., “@” or “//”)
d. Number of subdomains
e. Use of IP address instead of domain name
f. Age of domain, and more

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

IV. Extract HTML source code features (UF21 to UF30): Download and parse the HTML content of the web page corresponding to the input URL. From the source code, extract 10 additional features such as:
   a. Number of <script> and <iframe> tags
   b. Presence of JavaScript-based redirects
   c. Frequency of external links
   d. Usage of form handlers pointing to different domains
   e. Obfuscation in scripts, and more

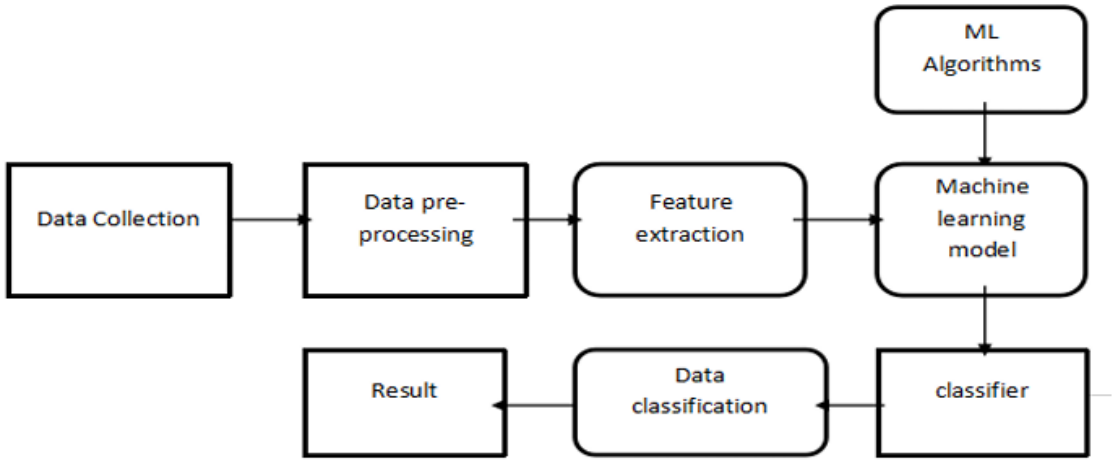
V. Combine URL and HTML features into a hybrid feature set. This results in a single set of 30 features that represent both the structure of the URL and the content of the web page.

VI. Apply the hybrid feature set to a pre-trained, high-performance machine learning classifier. In this case, a Gradient Boosting Classifier is used, which is well-suited for handling complex relationships between features and providing accurate predictions.

VII. Evaluate the classifier's prediction:
   a. If the classifier predicts the URL as phishing, set the prediction output to 1.
   b. Else, if the classifier predicts the URL as legitimate, set the prediction output to -1.

VIII. Return the final prediction result to the user.

SYSTEM ARCHITECTURE



INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT
ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

5. RESULT ANALYSIS

	Precision	Recall	F1 Score	Support
-1	0.95	0.96	0.97	976
1	0.97			1235
Accuracy			0.96	2211
Macro avg	0.96	0.96	0.96	2211
Weighted avg	0.96	0.96	0.96	2211

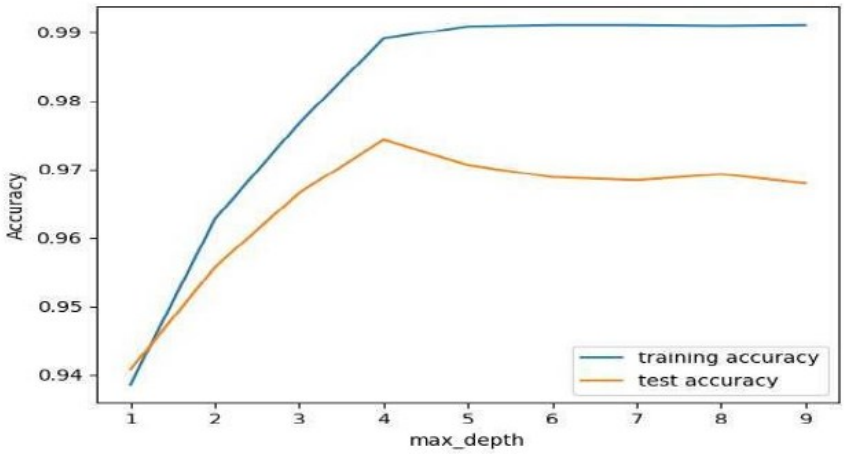


Fig 1: Plotting the training and test accuracy

6. CONCLUSION

Phishing has emerged as a significant threat to network security, making phishing prediction a critical challenge. In this project, we explored traditional phishing detection techniques such as blacklisting and heuristic-based methods, along with their limitations. To address these challenges, we developed a lexical analysis model in Python that analyses various URL features to classify websites as phishing or legitimate. We also evaluated the model's accuracy.

This work provided valuable insights into the distinguishing features between phishing and benign URLs, and how URL components can be extracted and transformed into machine-readable features. Through the process, I learned how to train and test machine learning models, and gained a deeper understanding of how models interpret datasets to generate predictions.

Our final analysis of the phishing dataset revealed that certain features—such as the use of HTTPS, anchor tags, and website traffic—play a more significant role in determining whether a URL is phishing or not.

ISSN 2454-9940
www.ijasem.org
Vol 19, Issue 2, 2025

7. FUTURE SCOPE

The project can also include other variants of phishing like smishing, vishing, etc. to complete the system. Looking even further out, the methodology needs to be evaluated on how it might handle collection growth. As phishing websites increases day by day, some features may be included or replaced with new ones to predict them**.**

8. REFERENCES

[1] Mohammed Hazim Alkawaz, Stephanie Joanne Steven (2021) "A Comprehensive Survey on Identification and Analysis of Phishing Website based on Machine Learning Methods".

[2] Almaha Abuzuraiq, Mouhammad Alkasassbeh, Mohammad Ali (Jan 2020) "Intelligent methods for accurately detecting phishing websites."

[3] Dhanushka Niroshan Atimorathanna, JayaniRukshila Perera (March 2020) "No-Fish; total anti- phishing protection system".

[4] Peng Yang, Guangzhen Zhao (2019) "Phishing Website Detection Based on Multidimensional features driven by Deep Learning".

[5] Malaika ERastogi, Anmol chethri, divyanshu kumar singh, gokul rajan (2021) "Survey on Detection and prevention of phishing websites using machine learning".