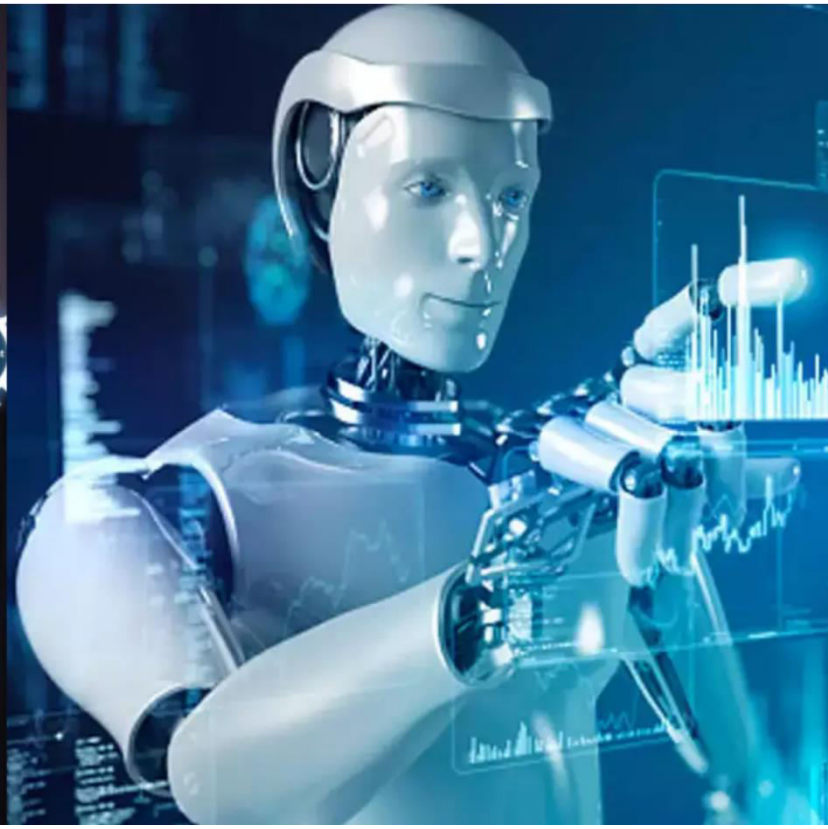
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404199|

Legit Link: Detecting Phishing Websites sing Machine Learning

Mrs.P.Shymala[1], Mohammad Sohail[2], Kashi Sreeya[3], Puli Nikhil[4], Thirukovela Srivat[5]

Department of Computer Science and Engineering, Malla Reddy University, Hyderabad, India[1-5]

ABSTRACT: Phishing is an internet scam in which an attacker sends out fake messages that look to come from a trusted source. A URL or file will be included in the mail, which when clicked will steal personal information or infect a computer with a virus. Traditionally, phishing attempts were carried out through wide-scale spam campaigns that targeted broad groups of people indiscriminately. The goal was to get as many people to click on a link or open an infected file as possible. There are various approaches to detect this type of attack. One of the approaches is machine learning. The URL's received by the user will be given input to the machine learning model then the algorithm will process the input and display the output whether it is phishing or legitimate. There are various ML algorithms like SVM, Neural Networks, Random Forest, Decision Tree, XG boost etc. that can be used to classify these URLs. By extracting and comparing different characteristics between legitimate and phishing URLs, the suggested method uses gradient boosting classifier to identify phishing URLs. The studies' findings demonstrate that the suggested approach successfully identifies legitimate websites from bogus ones in real time.

I. INTRODUCTION

Phishing attacks remain one of the most prevalent and successful forms of cybersecurity threats, with millions of users falling victim each year. These attacks typically involve creating deceptive websites that mimic legitimate ones to steal sensitive information such as login credentials, financial details, and personal data. The sophistication of these attacks has grown substantially, making traditional detection methods increasingly inadequate. Machine learning offers a promising approach to address this challenge by enabling automated, adaptive, and scalable detection systems. This research investigates the application of machine learning techniques to detect phishing websites effectively. By analyzing various website features and patterns associated with phishing attempts, machine learning models can be trained to distinguish between legitimate and malicious websites with high accuracy. The goal of this research is to develop a robust system, "Legit Link," that can provide real-time protection against phishing threats while minimizing false positives that might disrupt legitimate web browsing experiences.

The general method to detect phishing websitupdating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as "blacklist" method. To evade blacklists attackers uses creative techniques to fool users by modifying the URL to appear legitimate via obfuscation and many other simple techniques including: fast-flux, in which proxies are automatically generated to host the web-page; algorithmic generation of new URLs; etc. Major drawback of this method is that, it cannot detect zero- hour phishing attack. Heuristic based detection which includes characteristics that are found to exist in phishing attacks in reality and can detect zero-hour phishing attack, but the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high .To overcome the drawbacks of blacklist and heuristics based method, many security researchers now focused on machine learning techniques. Machine learning technology consists of a many algorithms which requires past data to make a decision or prediction on future data. Using this technique, algorithm will analyze various blacklisted and legitimate URLs and their features to accurately detect the phishing websites including zero- hour phishing websites.

II. OBJECTIVE

The objective of this project is to develop a machine learning-based solution for detecting phishing URLs in real-time. The focus is on using a gradient boosting classifier to accurately classify URLs as either legitimate or phishing. By extracting key features from the URLs, the model aims to distinguish between benign and malicious sites, thereby improving detection accuracy. The primary goal is to provide a reliable tool that can identify phishing attempts and help protect users from online scams and security threats. This project seeks toenhance internet security by offering an

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404199|

effective, automated method for identifying phishing URLs and preventing users from falling victim to these types of attacks.

III. LITERATURE SURVEY

Sahingoz et al. (2019) proposed a system that utilizes a self-constructed dataset, with phishing websites sourced from PhishTank and legitimate URLs obtained from the Yandex Search API. The study aimed to detect suspicious URLs based on characteristics such as similarity to brand names, the presence of random characters, and keyword identification. A range of classification algorithms, including Naive Bayes, Random Forest, k-Nearest Neighbors (k-NN), Adaboost, K-star, SMO, and Decision Trees, were used. Feature extraction methods, including Natural Language Processing (NLP) and Word Vectors, were applied, achieving high accuracy in phishing detection.

James et al. (2013) focused on detecting phishing websites by analyzing lexical features, host properties, and page-related characteristics of URLs. Their study compared several data mining algorithms, including Naïve Bayes, J48 Decision Tree, k-NN, and Support Vector Machines (SVM). Among these, the Decision Tree classifier achieved the highest accuracy (91.08%), suggesting that tree- based classifiers are well-suited for phishing URL classification.

Pradeepthi and Kannan (2014) examined a dataset consisting of 4500 URLs, including 2500 legitimate URLs from the DMOZ repository and 2000 phishing URLs from PhishTank. They evaluated various classification algorithms such as Naive Bayes, Random Forest, Random Tree, Multi-layer Perceptron, C-RT, J48 Tree, LMT, C4.5, ID3, and k-Nearest Neighbors. Their findings indicated that Random Forest provided the highest classification accuracy in detecting phishing URLs.

Kiruthiga and Akila (2019) reviewed 15 studies and proposed a method incorporating five classification algorithms: Decision Tree, Generalized Linear Model, Gradient Boosting, Generalized Additive Model, and Random Forest. Their results showed that the Random Forest algorithm achieved the highest performance, with an accuracy of 98.4%, a recall of 98.59%, and a precision of 97.70%. The dataset for this study was sourced from the UCI machine learning repository.

IV. FEATURE EXTRACTION

We have implemented python program to extract features from URL. Below are the features that we have extracted for detection of phishing URLs.

1) **Presence of IP address in URL**: If IP address present in URL then the feature is set to 1 else set to 0. Most of the benign sites do not use IP address as an URL to download a webpage. Use of IP address in URL indicates that attacker is trying to steal sensitive information.

2) **Presence of @ symbol in URL:** If @ symbol present in URL then the feature is set to 1 else set to 0. Phishers add special symbol @ in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol

3) **Number of dots in Hostname**: Phishing URLs have many dots in URL. For examplehttp://shop.fun.amazon.phishing.c om, in this URL phishing.com is an actual domain name, whereas use of "amazon" word is to trick users to click on it. Average number of dots in benign URLs is

4) **Prefix or Suffix separated by (-) to domain:** If domain name separated by dash (-) symbol then feature is set to 1 else to 0. The dash symbol is rarely used in legitimate URLs. Phishers add dash symbol (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example Actual site is http://www.onlineamazon.com but phisher can create another fake website like http://www.online-amazon.com to confuse the innocent users.

5) **URL redirection:** If "//" present in URL path then feature is set to 1 else to 0. The existence of "//" within the URL path means that the user will be redirected to another website

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404199|

6) **HTTPS token in URL:** If HTTPS token present inURL then the feature is set to 1 else to 0. Phishers may add the "HTTPS" token to the domain part of a URL in order to trick users. For example, http://https- www-paypal-it-mpp-home.soft-hair.com

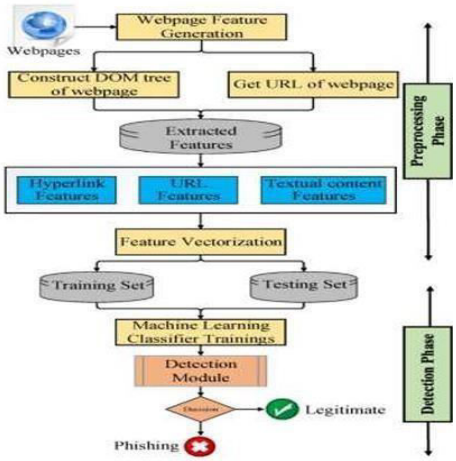
7) **Information submission to Email:** Phisher might use "mail()" or "mailto:" functions to redirect the user's information to his personal email[4]. If such functions are present in the URL then feature is set to 1 else to 0.

8) **Presence of sensitive words in URL:** Phishing sites use sensitive words in its URL so that users feel that they are dealing with a legitimate webpage. Below are the words that found in many phishing URLs
: 'confirm', 'account', 'banking', 'secure',
'ebyisapi', 'webscr', 'signin', 'mail','install', 'toolbar', 'backup', 'paypal', 'password', 'username', etc;

9) **Length of Host name:** Average length of the benign URLs is found to be a 25,

If URL's length is greater than 25 then the feature is set to 1 else to 0

V. METHODOLOGY



I. **Data Collection and URL Retrieval** : The system starts by retrieving URLs to be checked for phishing. These URLs can be collected from user input in the webpage created.

1. **URL Validation:**
The system first validates the user- provided URL to ensure it is well- formed and syntactically correct. This step helps prevent errors due to invalid input.

2. **Feature Extraction:**
Once the URLs are obtained, the system extracts relevant features from the web pages. These features are essential for training and evaluating the machine learning models. Various features were extracted from the URL database based on Domain, HTML and Address bar of the URLs

3. **Data Preprocessing** :
The extracted features often need to be preprocessed to ensure consistency and compatibility with the machine learning models.

4. **Machine Learning Model:**
A machine learning model, trained on a labeled dataset containing both legitimate and phishing URLs, is used to make predictions.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

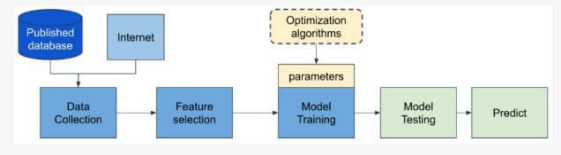
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404199|

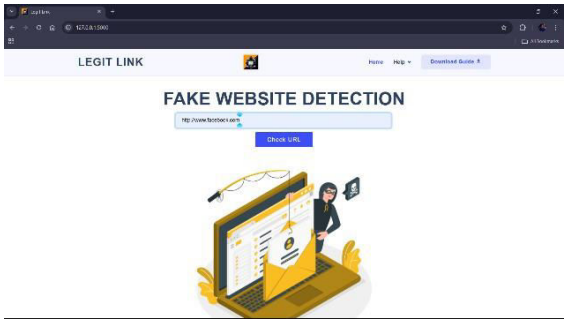
5. **Prediction and Scoring:**
   The machine learning model predicts whether the URL is a phishing site or not. It provides a probability score or a binary classification (phishing or not phishing) based on the trained model's decision boundary.

6. **Result Presentation:**
   The system categorize URLs into "phishing" or "legitimate" and the result is finally displayed on the webpage.

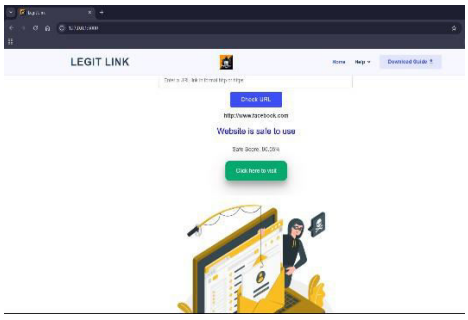


V. RESULT

(a)



(b)



(c)

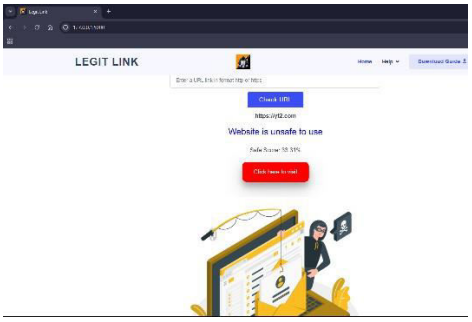
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404199|



(d)

(e)

VII. CONCULSION

It is found that phishing attacks is very crucial and it is important for us to get a mechanism to detect it. As very important and personal information of the user can be leaked through phishing websites, it becomes more critical to take care of this issue. This problem can be easily solved by using any ofthe machine learning algorithm with the classifier. We already have classifiers which gives good prediction rate of the phishing besides, but after our survey that it will be better to use a hybrid approach for the prediction and further improve the accuracyprediction rate of phishing websites. We have seen that existing system gives less accuracy so we proposed a new phishing method that employs URL based features and also, we generated classifiers through several machine learning. We have got the desired results of testing the site is phishing or not by using five different classifiers.

VIII. FUTURE ENHANCEMENT

Further work can be done to enhance the model by using ensembling models to get greater accuracy score. Ensemble methods is a ML technique that combines many base models to generate an optimal predictive model. Further reaching future work would be combining multiple classifiers, trained on different aspects of the same training set, into a single classifier that may provide a more robust prediction than any of the single classifiers on their own. 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. The collections will ideally grow incrementall.

It is found that phishing attacks is very crucial and it is important for us to get a mechanism to detect it. As very important and personal information of the user can be leaked through phishing websites, it becomes more critical to take care of this issue. This problem can be easily solved by using any of the machine learning algorithm with the classifier. We already have classifiers which gives good prediction rate of the phishing besides, but after our survey that it will be over time so there will need to be a way to apply a classifier incrementally to the new data, but also potentially have this classifier receive feedback that might modify it over time.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404199|

REFERENCES

[1] M. H. Alkawaz, S. J. Steven and A. I. Hajamydeen, "Detecting Phishing Website Using Machine Learning," 2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA), 2020, pp. 111-114

[2] V. Patil, P. Thakkar, C. Shah, T. Bhat and S. P. Godse, "Detection and Prevention of Phishing Websites Using Machine Learning Approach," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA),2018,pp.1-5.

[3] W. Bai, "Phishing Website Detection Based on Machine Learning Algorithm," 2020 International Conference on Computing and Data Science (CDS),2020,pp.293-298

[4] A. Razaque, M. B. H. Frej, D. Sabyrov,A. Shaikhyn F. Amsaad and A. Oun, "Detection of Phishing Websites using Machine Learning," 2020 IEEE Cloud Summit, 2020, pp. 103-107.

[5] A. Alswailem, B. Alabdullah, N. Alrumayh and A. Alsedrani, "Detecting Phishing Websites Using Machine Learning," 2019 2nd International Conference on Computer Applications & Information Security(ICCAIS),2019,pp.1-6.

[6] Yuan, H., Chen, X., Li, Y., Yang, Z., & Liu, "Detecting Phishing Websites and Targets Based on URLs and Webpage Links," 2018 24th International Conference on Pattern Recognition 2018, pp. 3669- 3674.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details