IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

Phishing Website Detection using Machine Learning

Indubu Nutana[1], J. Tagore Babu[2], J. Jahnavi[3], J. Yogendra[4], J. Siva Tanay Akash[5]
GMR Institute of Technology, Rajam, Andhra Pradesh, India
20341A0571@gmrit.edu.in[1], 20341A0572@gmrit.edu.in[2], 20341A0575@gmrit.edu.in[3],
20341A0574@gmrit.edu.in[4], 20341A0573@gmrit.edu.in[5]

Abstract: *Phishing websites today pose serious risks because of their almost invisible risk. In these attacks, fraudsters trick users into handing over their login information or other sensitive information through a login form that mimics the target website and submits the information to an evil server. In recent years, numerous anti-phishing strategies have been created, using various resources such as the URL and HTML code from trustworthy and fraudulent index websites. These methods have significant restrictions when predicting authentic login websites because they frequently lack login forms to specify the appropriate class that was used to train the suggested model. The URL, HTML, and web technology properties are used to detect phishing websites in actual situations. In this work the phishing website datasets will be used with machine learning algorithms like Random Forest, Logistic regression, KNN and SVM to test the best method, so the crawl trusted websites are taken to align with phishing perspective. The study involves comparison of the four algorithms and finding efficient algorithm in the basis of accuracy.*

Keywords: Phishing, Cybercrime, Detection, Phishing Dataset, Machine Learning

I. INTRODUCTION

Phishing is an attack in which an attacker tries to convince victims to provide critical and private information. This includes credit card numbers, passwords. Phishing Attacks are increasing yearly, due to more internet users. Distinguishing between legal and fraudulent web pages is difficult because they appear to be alike. Phishing attack detection is mostly characterized as a classification problem. Machine learning methods are therefore viewed as promising approaches. When using such techniques, the following three crucial factors must be taken into account: choosing effective classifiers, utilising distinguishing characteristics, and gathering samples of a representative dataset for training. The majority of phishing detection technologies fall into two categories: Content-based and URL-based systems. Among many algorithms, the analytical and comparative analysis of machine learning methods like Random Forest, Logistic regression, KNN and SVM are used. This paper is organized as: firstly, listing some widely used phishing techniques. Secondly, providing an overview of different machine learning methods. Thirdly, showing evaluation results of suggested machine learning methods and finally concluding which algorithm is efficient.

II. LITERATURE SURVEY

[1] Sánchez-Paniagua, M., Fidalgo, E., Alegre, E., & Alaiz-Rodríguez, R. (2022). Phishing websites detection using a novel multipurpose dataset and web technologies features. ***Expert Systems with Applications*****, *207*, 118010.**
This study presented an approach that makes use of URL, HTML, and web technology properties to identify phishing websites in actual situations. The author introduced PILWD,offline dataset that demonstrated a new phishing detection method that is differentiated by including legitimate samples primarily from login sites.In addition, the new set of web-based features improved the detection process and accuracy.According to experimental findings, phishing websites when tested on the PILWD dataset, was detected with 97.95% accuracy using a Light BGM classifier and the entire set of the 54 features used.

[2] Hannousse, A., & Yahiouche, S. (2021). Towards benchmark datasets for machine learning based website phishing detection: An experimental study. Engineering Applications of Artificial Intelligence, 104, 104347.
This paper presented a general approach for creating extendable datasets that may be used to detect online phishing. It shows Random Forest classifier based systems can be used in browsers to effectively predict phishing web pages. They

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

have comparatively higher accuracy than all the other classifiers for different categories of features like Extraction order, Chi-Square order,Pearson Correlation order,Information gain order and Relief order.Random Forest is feature order sensitive, filter methods can effectively be used to improve the performance of the model and reduce less important features.Random Forest accuracy in different categories extraction order 96.61%,chi-square order 96.83%, Pearson correlation 96.65%,information gain 96.76%,and relief order 96.66%.

[3] Shahrivari, V., Darabi, M. M., & Izadi, M. (2020). Phishing detection using machine learning techniques. arXiv preprint arXiv:2009.11116.

This study offered a comparative and analytical review of various machine learning techniques such as Logistic Regression, Decision Tree, Random Forest, Ada-Boost, Support Vector Machine, KNN, Artificial Neural Networks, Gradient Boosting, and XGBoost, to identify phishing websites. Author have implemented and evaluated twelve classifiers on the phishing website dataset that consists of 6157 legitimate websites and 4898 phishing websites. The result shows, a very good performance in ensembling classifiers namely, Random Forest, XGBoost both on computation duration and accuracy. The reason for this amazing result from Random Forest is because of the trees protect each other from individual errors. The most important factor behind the success of XGBoost is its scalability in all scenarios.

[4] Ramana, A. V., Rao, K. L., & Rao, R. S. (2021). Stop-Phish: an intelligent phishing detection method using feature selection ensemble. Social Network Analysis and Mining, 11(1), 1-9.

This study worked to provide an intelligent model that can effectively identify phishing websites. The author used various machine learning algorithms for identifying the best classifier and developed an ensemble model with Random forest, Decision tree and XGBoost algorithms and also used various feature selection ensembles like ANOVA, information gain for the classification of phishing websites. The proposed model with ensemble of Random forest, Decision tree and XGBoost achieved better performance with an accuracy of 97.38% than the existing classifiers.

[5] Rao, R. S., Vaishnavi, T., & Pais, A. R. (2020). CatchPhish: detection of phishing websites by inspecting URLs. Journal of Ambient Intelligence and Humanized Computing, 11(2), 813-825.

The author proposed a light-weight application, Catch Phish which predicts the URL legitimacy without visiting the website. The technique uses hostname, full URL, Term Frequency-Inverse Document Frequency (TF-IDF) features and phish hinted words from the suspicious URL for the classification using the Random forest classifier. It is independent of third party services and source code. This behavior makes the technique adaptable at client side due to its low response time. Random forest classifier is used for the classification of phishing and legitimate sites. The Catch Phish application achieved an accuracy of 94.26%precision on the dataset taken and and an accuracy of 98.25% on a benchmark dataset.

III. METHODOLOGY

Phishing is a form of social engineering assault that is frequently employed to obtain user information, such as login information and credit card information. It happens when an attacker deceives a target into open an email, text messenger, or sms by presenting themselves as a reliable source. Next, a dangerous link is deceived into being clicked by the recipient. This can cause malware to be installed on the recipient's computer, a ransomware assault to lock it down, or the disclosure of private data. An assault can have disastrous consequences. For people, this includes theft of money or identity, as well as illicit transactions.

The process of phishing website detection-

Step-1 : The data set is collected. The analysis of data set is done and organized.

Step-2: The data is filtered 3 times to increase dataset quality.

Filter 1: Repeated legitimate websites in phishing reports.

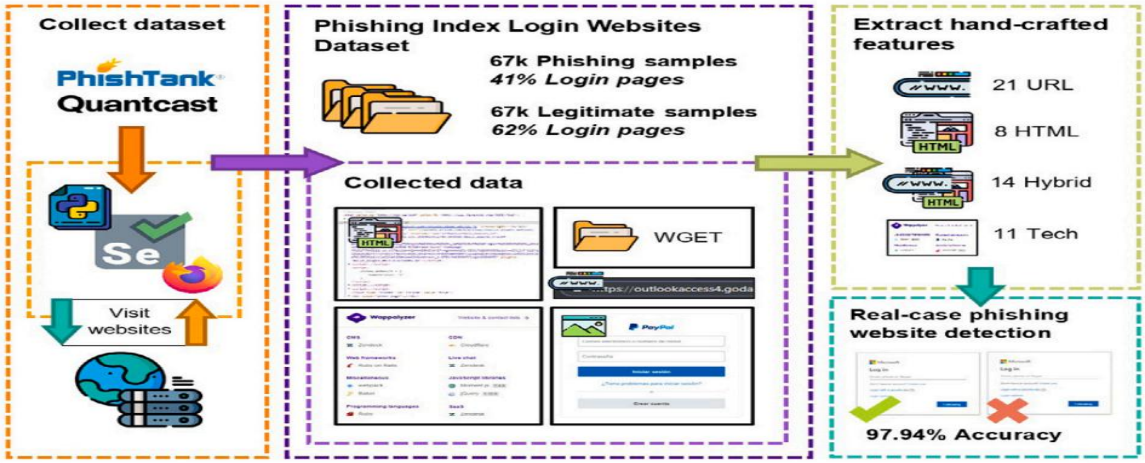
Filter 2: Empty and error samples.

Filter 3: Verified phishing samples.

Step-3: The features are extracted in the different sets .

The strategy depends on integrating four categories.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252



3.1 Features: URL, HTML, Hybrid and Technologies

A. URL Features

The URL is a unique identifier of the website and locates it on the Internet.
The final set of URL characteristics is described below:

- **Subdomain level:** Hiding the original domain name and using the initial section of the URL as the target to trick people.
- **IP address:** Attackers must purchase this service in order to use a domain name on the website. If not, users only would be able to access the website by entering its IP address with in menu.
- **Common Top-Level Domain (TLD):** The "com,""org,""net,""edu," and "gob," as well as all Country Code TLDs (ccTLD), are regarded as regular. This binary characteristic is set to 1 if the target URL TLD is not present in list.
- **Length:** When comparing to legal URLs,phish URLs were lengthier.
- **Digits:** Fake domain names typically have more numbers than real domain names.
- **Special characters:** Fake URLs frequently contain strange symbols.
- **Random words:** Phish websites continue to use random words.

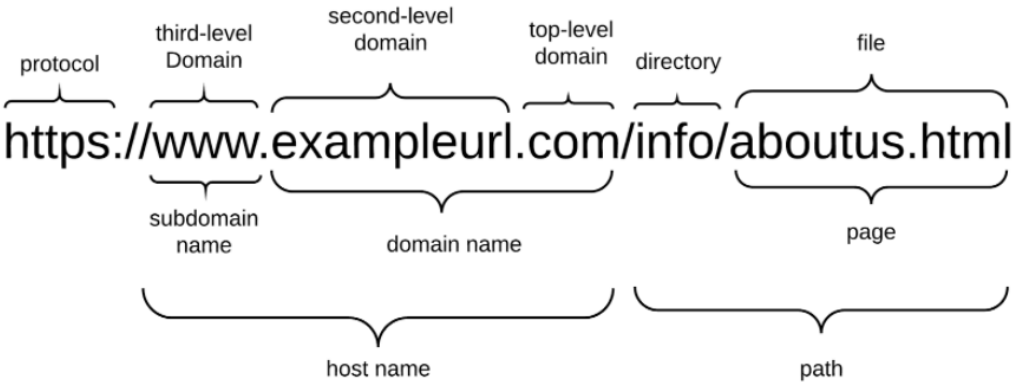


Figure 2. URL parts.

B. HTML Features

Free hosting companies frequently offer low-resource servers with minimal RAM, few CPU cores, and little disc space. These constraints prevent the attacker from hosting and storing sophisticated websites. Because of this, phish websites often only have a single HTML page with the login screen on it.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

Links: The majority of phishing sites use a distinct web domain to host themselves while pretending to be a reputable website. Attackers employ connections from the main website resources to improve their websites while their sitemap and internal link count are relatively little as a means of misleading users. The amount of external links will rise as more important reputable websites start to store their resources in the cloud.

Body length and tags: Used a method to compute the length of the websites by using the content of style, script, link, comments and form tags.

C. Hybrid Features

Copyright in the HTML: Phishers objective is to spoof the brands to increase users confidence. We detect whether a website has the copyright disclaimer, or not.

Domain in body: Only the HTML body is used by this functionality. In this method, metadata which can contain information about the domain is excluded.

some other features include-

Title-Domain; Title-Domain-Copyright etc.

Technology-based features: Some technologies are more frequently used on legitimate websites than on phishing ones. To select those technologies, we introduced the total of 134 detected technologies into SelectKBest library from scikit-learn and used the f_classification algorithm to create the technologies ranking.

IV. RESULTS AND DISCUSSION

4.1 Performance Metrics and Evaluation

The effectiveness of phishing models is evaluated using accuracy, precision, recall, and F-score. The real class is the negative class, and the phishing class is the positive one.

ACCURACY: The no. of correctly classified samples is taken main.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

F1-SCORE: The mean of precision and recall.

$$F1Score = 2 * \frac{Recall * Precision}{Recall + Precision}$$

RECALL: Recall refers to the fraction of correctly classified phishing samples among the total number of phishing instances.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

PRECISION: It evaluate the fraction of correctly classified instances or samples among the ones classified as positives.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

True Positive (TP) denotes how many phishing samples were accurately identified as such, whereas True Negative (TN) denotes how many genuine samples were correctly identified as such. False Positive (FP) reflects the number of valid samples that were incorrectly labeled as phishing. False Negative (FN) indicates how many phishing samples there were incorrectly categorized as legal.

The proposed subset, D1, was built to represent this scenario by using completely filtered and verified phishing samples along with legitimate login samples and legitimate homepage samples. The result is a dataset where both classes, Legitimate and Phishing, have a similar amount of login websites.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

D2 depicts the methodology used in most state-of-the-art papers, where authors collect only the homepage of the most visited domains and the phishing websites.

Subsets D3 and D4 were created to benchmark the behaviour of different proposals when facing a different set of legitimate websites. D3 contains only legitimate login samples, and D4 was built using legitimate homepage samples only.

Algorithm	Precision	Recall	F1 score	Accuracy
Random Forest	0.980	0.972	0.976	**97.63%**
Logistic regression	0.890	0.901	0.899	89.79%
KNN	0.931	0.948	0.940	93.91%
SVM	0.955	0.942	0.948	94.86%

V. CONCLUSION

The phishing detection is done with a dataset. Here firstly, a dataset is taken. Secondly, the new web technology features were added to improve accuracy. Thirdly, the novel technologies were introduced to give more protection. The new model combines two novel phishing detection ideas: the first is the use of a large number of valid login websites in the dataset; the second is the application of new hand-crafted features combined with website technology analysis.

Finally, the taken algorithms were used to test. The dataset collected includes screenshots, URL HTML and other data. At last the new features introduced and a new body length measure method performed well. Due to less samples and usage of other services there are few limitations.The Logistic Regression algorithm gave 89.79% accuracy, Random Forest gave 97.63%, KNN gave 93.1% accuracy, SVM gave 94.86% on the taken dataset. Thus, Random Forest was best among the other selected classifiers and it also had 97.63% accuracy.

REFERENCES

[1]. Sánchez-Paniagua, M., Fidalgo, E., Alegre, E., & Alaiz-Rodríguez, R. (2022). Phishing websites detection using a novel multipurpose dataset and web technologies features. Expert Systems with Applications, 207, 118010.

[2]. Hannousse, A., & Yahiouche, S. (2021). Towards benchmark datasets for machine learning based website phishing detection: An experimental study. Engineering Applications of Artificial Intelligence, 104, 104347.

[3]. Shahrivari, V., Darabi, M. M., & Izadi, M. (2020). Phishing detection using machine learning techniques. arXiv preprint arXiv:2009.11116.

[4]. Ramana, A. V., Rao, K. L., & Rao, R. S. (2021). Stop-Phish: an intelligent phishing detection method using feature selection ensemble. Social Network Analysis and Mining, 11(1), 1-9.

[5]. Rao, R. S., Vaishnavi, T., & Pais, A. R. (2020). CatchPhish: detection of phishing websites by inspecting URLs. Journal of Ambient Intelligence and Humanized Computing, 11(2), 813-825.

[6]. Fitzpatrick, B., Liang, X., & Straub, J. (2021). Fake news and phishing detection using a machine learning trained expert system. arXiv preprint arXiv:2108.08264.

[7]. Christou, O., Pitropakis, N., Papadopoulos, P., McKeown, S., & Buchanan, W. J. (2020). Phishing url detection through top-level domain analysis: A descriptive approach. arXiv preprint arXiv:2005.06599.

[8]. Tupsamudre, H., Jain, S., & Lodha, S. (2021). PhishMatch: A Layered Approach for Effective Detection of Phishing URLs. arXiv preprint arXiv:2112.02226.

[9]. Lakshmanarao, A., Rao, P. S. P., & Krishna, M. B. (2021, March). Phishing website detection using novel machine learning fusion approach. In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) (pp. 1164-1169). IEEE.

[10]. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. Expert Systems with Applications, 117, 345-357.

[11]. Abuzuraiq, A., Alkasassbeh, M., & Almseidin, M. (2020, April). Intelligent methods for accurately detecting phishing websites. In 2020 19th International Conference on Information and Communication Systems (ICICS) (pp. 085-090). IEEE.

[12]. He, S., Li, B., Peng, H., Xin, J., & Zhang, E. (2021). An effective cost-sensitive XGBoost method for

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, November 2022
Impact Factor: 6.252

malicious URLs detection in imbalanced dataset. IEEE Access, 9, 93089-93096.

[13]. Mao, J., Bian, J., Tian, W., Zhu, S., Wei, T., Li, A., & Liang, Z. (2019). Detecting phishing websites via aggregation analysis of page layouts. Procedia Computer Science, 129, 224-230.

[14]. Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q., & Wang, Y. (2022). An effective detection approach for phishing websites using URL and HTML features. Scientific Reports, 12(1), 1-19.

[15]. Wood, T., Basto-Fernandes, V., Boiten, E., & Yevseyeva, I. (2022). Systematic Literature Review: Anti-Phishing Defences and Their Application to Before-the-click Phishing Email Detection. arXiv preprint arXiv:2204.13054.