INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact Factor: 6.078
(Volume 10, Issue 6 - V10I6-1316)
Available online at: https://www.ijariit.com

Phishing Web Detection using Machine Learning Technique

Samiksha Sachin Karanjkar
samikshakaranjkar10102003@gmail.com
G.H. Raisoni College of Engineering and Management, Wagholi, Pune, Maharashtra

Shruti Jadhav
uniqueshrutij0507@gmail.com
G.H. Raisoni College of Engineering and Management, Wagholi, Pune, Maharashtra

Vrushali Pimpale
vrushalipimpale17@gmail.com
G.H. Raisoni College of Engineering and Management, Wagholi, Pune, Maharashtra

Rahul Navale
tyco1.48.samikshakaranjkar@gmail.com
G.H. Raisoni College of Engineering and Management, Wagholi, Pune, Maharashtra

ABSTRACT

Everyone is addicted to the internet these days. All of us have made reservations, recharged, shopped, and banked online. Phishing is a type of online threat to websites. Phishing, according to the original website, is an illegal attempt to collect information such a credit card number, login ID, and password. We presented a successful machine learning-based phishing detection technique in this research. Overall, the experimental findings demonstrated that the suggested method performs best when used in combination with support vector machine classifiers, detecting 95.66% of phishing attempts and matching websites with just 22.5% of novel functionality. When compared to many popular phishing datasets from UCI's repository, the suggested method yields encouraging results. For machine learning-based phishing detection, the suggested method is therefore the one that is favored and utilized.

Keywords: *Phishing, Web, Machine Learning, Principal Component Analysis, Support Vector Machine, Random Forest, Decision Tree, Naïve Bayes*

1.INTRODUCTION

Several criminal enterprises now use the internet as a tool to send spam, commit financial fraud, and distribute malware. All of them must guarantee that customers are not compelled to visit their website, even though the legitimate business justifications for this strategy may differ. This visit should be facilitated by email, web query items, or links from other site pages; however, the client must act quickly to obtain crucial information, such as by providing the correct URL (Uniform Resource Locator). In response, the security community developed a blacklist service that provides precise feedback along with alerts or warnings and is integrated into toolbars, gadgets, and search engines. Because they are too new, unclassified, or misclassified, many dangerous websites are not banned.

One kind of cyberattack called phishing uses websites to obtain important customer data, including credit card numbers, accounts, login credentials, and much more. In June 2018, 51,401 phishing sites were formally suggested by the "APWG (ANTI-PHISHING WORKING GROUP)" [1]. According to another RSA analysis, the estimated global cost of phishing attacks is close to $9 billion. Statistics from 2016 [2] demonstrate the failure of traditional anti-phishing tactics and measures.

The blacklist warning system, which is present in popular browsers including Chrome, Internet Explorer, and Mozilla Firefox, is the most widely used anti-phishing solution. Due to its central database of suspected phishing URLs, the Blacklisted Survey Device is unable to locate recently open phishing sites [3, 4].

Accuracy aspects are the foundation of the machine learning-based phishing detection device's effectiveness. While developing appropriate analysis and selection techniques is not a crucial strategy, the majority of anti-phishing researchers concentrate on new features or optimization of classification algorithms [5, 6].

The Website's 12 Features Are Phishing-Friendly and Actual, Achieving A 97% Effective Positive Rate and A 4% False Positive Rate, Considering [5]. Features are obtained via Tf-Idf, URLs, Hyperlinks, Web Page Content, Meta Coding, And Other Means. Therefore, the average accuracy is unaffected by external factors that may still raise the technique's cost (such as training time, storage, electricity, etc.). Therefore, A Powerful Machine Learning-Based Technique for Phishing Detection Is Required to Detect Truly Effective Compact Features.

Machine learning-based phishing detection is presented in this paper. Moreover, classifiers developed using machine learning algorithms are able to identify authentic phishing websites [19]. The Suggested Technique Used Svm with A 95.66% Accuracy and A Very Low False-Positive Rate. The proposed technique can detect new, transient phishing websites and lessen the damage caused by phishing attacks. It has been suggested that the machine learning-based approach outperforms earlier phishing detection technologies. Hybrids [14] Use Different Extreme Machine Learning (ELM) Technologies For Future Research To Extract

Features From Web Content, Text, And URLs. To use ELM to determine the label text's content, the first step in this process is to write the classifier's text content.

In [15], a technique for creating a stochastic neural network (PNN) was introduced. Outliers' Numbness, Generalization, And Fast Train Pnn Time Are the Best Features. However, PNNS can significantly increase data, requiring a significant amount of time and space. Therefore, the author's group used PNNS in K-Medoids to decrease the number of training cases. proposed a method to prevent phishing attacks on the Iranian electronic banking system in [16]. The Author Identifies 28 Characteristics Used By Hackers To Deceive Iranian Banking Websites. The accuracy of the Iranian banking system's detection was 88%. This technique, which can only distinguish legitimate websites from phishing ones, was developed specifically to locate Iranian bank websites.

Comparing a suspect site's operations with a pre-established set of operations is a common strategy used in machine learning techniques. Therefore, the quality of these functions and the accuracy with which the defendant chooses them determine the accuracy of the system. In a recent study, natural language processing (NLP) was used to detect phishing emails by examining the email text's semantic content to find any malicious intent [17]. In order to identify phishing websites, this work uses machine learning algorithms to detect web page URLs. When implementing a machine learning algorithm, the extraction of features from the dataset is a crucial step. As a result, Google uses pre- existing datasets to compile a sizable number of legitimate and fraudulent website URLs. Measuring according to the function that the term vector defines, the effectiveness of the suggested system is assessed.

2.RELATED WORK

i. **A Methodical Overview on Detection, Identification and Proactive Prevention of Phishing Websites:** Bhagwat M. D., Dr. Patil P. H (2020) He Represents the Methodical Overview on Detection, Identification and Proactive Prevention of Phishing Websites. Detecting and finding some phishing websites in real-time for a day now is really a dynamic and nuanced topic involving several variables and requirements.

ii. **A Machine Learning Approach for URL Based Web Phishing Using Fuzzy Logic as Classifier:** Happy Chapla, Riddhi Kotak (2021) stated that A Machine Learning Approach for URL Based Web Phishing Using Fuzzy Logic as Classifier. Phishing is the major problem of the internet era. In this era of internet, the security of our data in web is gaining an increasing importance.

iii. **Detection and Prevention of Phishing Websites Using Machine Learning Approach:** Vaibhav Patil, Pritesh Thakkar, Chirag Shah, Tushar Bhat, S. P. Godse (2020) They Suggests That Detection and Prevention of Phishing Websites Using Machine Learning Approach. Phishing costs Internet user's lots of dollars per year. It refers to exploiting weakness on the user side, which is vulnerable to such attacks.

iv. **Detection of Phishing Attacks with Machine Learning Techniques in Cognitive Security Architecture:** Chirag Pawar ,D. P. Bhat, Dr. Thakkar P. H.(2019) states that Detection of Phishing Attacks with Machine Learning Techniques in Cognitive Security Architecture. The number of phishing attacks has increased in Latin America, exceeding the operational skills of cybersecurity analysts

v. **On Effectiveness of Source Code and SSL Based Features for Phishing Website Detection**: Roopak.S, P Vijayaraghavan, Tony Thomas (2021) They Suggest that On Effectiveness of Source Code and SSL Based Features for Phishing Website Detection. Phishing is a social engineering method to steal user credentials through data entry forms from malicious websites. Currently available anti-malware software’s can only detect black listed phishing websites.

3.ARCHITECTURE OF PHISHING WEB DETECTION

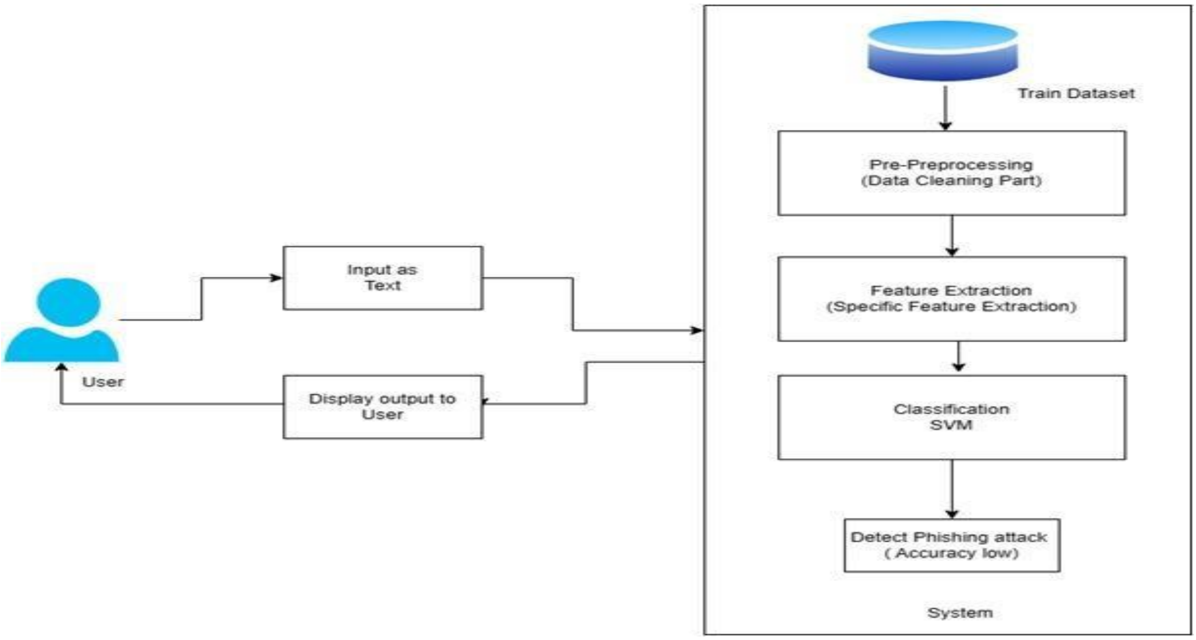


Fig a: Architecture of Phishing web detection

a. Data Collection:

Phish Tank, an open-source application, was used to collect the phishing URLs. This website offers a collection of hourly-updated phishing URLs in a number of formats, such as csv, Json, and others. Machine learning models are trained using this dataset and phishing URLs.

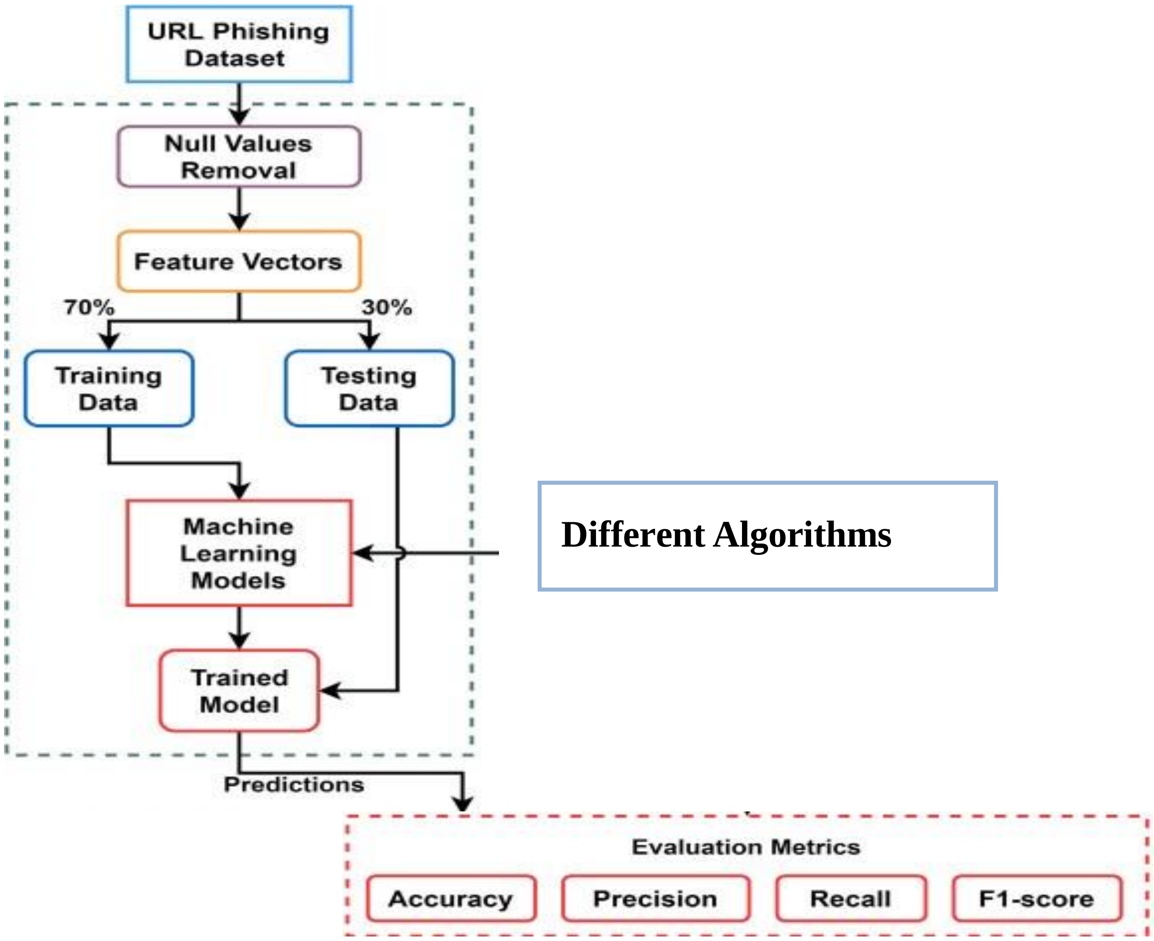
b. Data Pre-processing:

Data Purification To clean up the data, add missing values, smooth out shaky data, find and eliminate outliers, and fix anomalies. Preparing unstructured raw data into a clean, well-structured dataset that may be utilized for additional research is known as data preparation. Data preprocessing is a cleansing technique that turns raw, unstructured data into a clean, well-organized dataset that may be utilized for additional research.

c. Feature Extraction:

In Feature Extraction we extract the Required Features from the URL Dataset.

4.PROPOSED SYSTEM



i. Module:

Admin

The administrator must enter a working user name and password to access this module. He can perform some actions, like viewing all users and authorizing, after successfully logging in.

End User

A total of n users is present in this module. Before beginning any operations, the user should register. The user's information will be added to the database as soon as they register. Following a successful registration, he must use his password and permitted user name to log in.

ii. Support Vector Machine:

One of the most widely used supervised learning techniques for resolving regression and classification issues is the Support Vector Machine, or SVM. Nonetheless, its primary use is to address classification issues in machine learning. In order to make it simple to place fresh data in the appropriate point category later on, the SVM algorithm aims to provide the best line or decision boundary that can split the n-dimensional space into categories. We refer to this optimal decision boundary as the hyperplane. To aid in the creation of the hyperplane, the SVM chooses extreme points and vectors. The approach is referred described as a support vector machine since these extreme situations are known as support vectors. Examine the picture below, which uses a decision boundary or hyperplane to classify two distinct classes.

iii. Random Forest:

According to its name, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Rather than depending on a single decision tree, the random forest forecasts the final result by taking the predictions from each tree and calculating the majority vote of predictions. Accuracy increases and overfitting is avoided when there are more trees in the forest.

iv. Decision Tree:

Although it may be applied to both classification and regression problems, decision trees—a supervised learning technique—are primarily employed to solve classification difficulties. This classifier is tree-structured, with internal nodes standing in for a dataset's features, branches for the decision rules, and each leaf node for the result. The Decision Node and the Leaf Node are the two nodes that make up a decision tree. Leaf nodes are the results of those decisions and do not have any additional branches, while decision nodes are utilized to make any decision and have numerous branches. The features of the provided dataset are used to make choices or run tests. It provides a visual depiction of every potential answer to a issue or choice based on the circumstances. The reason it is named a decision tree is that, like a tree, it begins with the root node, which grows on additional branches to form a structure resembling a tree.

v. Naïve Bayes:

Based on the Bayes theorem, the Naïve Bayes algorithm is a supervised learning technique used to solve classification issues. Text classification with a high-dimensional training dataset is its primary use. One of the simplest and most efficient classification techniques for creating short machine learning models with rapid prediction capabilities is the Naïve Bayes Classifier. Being a probabilistic classifier, it makes predictions based on an object's likelihood.

5.RESULT ANALYSIS

The proposed machine learning-based method is used to compare it with the previous method. In our work, a similar classification method is used to train the split test. The proposed machine learning-based strategy is compared to the earlier methodology. In our experiment, we also train the split test using a similar categorization technique.

i. **Precision:** The accuracy of a machine learning model's positive prediction is one measure of the model's performance. The number of genuine positives divided by the total number of positive predictions—that is, the sum of the true positives and false positives—is known as precision. Precision measures the percentage of cases or samples that are accurately classified out of those that are labeled as positives.
Thus, the formula to calculate the precision is given by: **Precision = TP / (TP+ FP)**

ii. **F1 Score:** A metric called the F1 score balances precision and recall to assess a machine learning model's performance. Both binary and multiclass classification problems frequently employ it. The F1 score goes from 0 to 1, where 0 represents the lowest possible score and 1 represents an ideal outcome. While a low F1 score frequently denotes a trade-off between recall and precision, a high F1 score shows a well-balanced performance. The main purpose of the F1 score is to compare two classifiers' performances. For unbalanced data, it might not provide an accurate representation of the model's performance. This is so because precision and recall are given equal weight in the standard F1 score.
The formula for calculating the F1 score is: **F1 = 2 * (Precision + Recall) / (Precision * Recall)**

iii. **Accuracy:** A performance statistic called accuracy is a percentage of how many accurate predictions the model makes. It is a statistic that provides a general description of the model's performance in every class.
It is defined as, **Accuracy = Number of Correct Prediction / Total Number of predictions**

iv. **Recall:** The frequency with which a model accurately detects positive instances (true positives) out of all the real positive samples in the dataset is known as recall. The true positive rate (TPR) is another name for it. The number of true positives divided by the total number of positive cases plus the number of false negatives is the recall. the capacity of a model to locate every pertinent instance in a dataset. Mathematically, we define recall as the number of true positives divided by the number of true positives plus the number of false negatives.
Recall= TP / TP + FN

v. Mathematical Model

Let S be the Whole system S=I, P,
O I-input
P procedureO-
output Input(I)
I=Text
Dataset Where,
Dataset- Text dataset Classification Using SVM/RF/NB Algorithm Procedure(P),
P=I, Using I System Detect Phishing Web or Not.
Accuracy = Number of Correct Prediction / Total Number of predictions

6.FUTURE SCOPE

A better and more secure online environment can be achieved by progressively improving the efficacy and dependability of phishing website detection utilizing SVM and RF algorithms in the future. The time between detections can be greatly decreased by creating real-time phishing detection systems that can evaluate website data and make conclusions instantly. expanding phishing detection models to include a variety of internet channels, such as chat applications and social media.

7.CONCLUSION

This study introduces phishing detection based on machine learning. Additionally, classifiers developed using machine learning techniques are able to identify legitimate phishing websites. The proposed method, which employed SVM, had an accuracy of 95.66% and an extremely low false-positive rate. The proposed technique can detect new, temporary phishing websites and reduce the damage caused by phishing attempts. Previous phishing detection solutions are outperformed by the proposed machine learning-based method. It will be helpful to conduct future studies that investigate the impact of feature selection using various classification algorithms.

REFERENCES

[1] Higashino, M., et al. An Anti-phishing Training System for Security Awareness and Education Considering Prevention of Information Leakage. in 2019 5th International Conference on Information Management (ICIM). 2019.

[2] H. Bleau, Global Fraud and Cybercrime Forecast,. 2017.

[3] Michel Lange, V., et al., Planning and production of grammatical and lexical verbs in multi-word messages. PloS one, 2017. 12(11): p. e0186685- e0186685.

[4] Rahman, S.S.M.M., et al. Performance Assessment of Multiple Machine Learning Classifiers for Detecting the Phishing URLs. 2020. Singapore: Springer Singapore.

[5] He, M., et al., An efficient phishing webpage detector. Expert Systems with Applications, 2011. 38(10): p. 12018- 12027.

[6] Mohammad, R.M., F. Thabtah, and L. McCluskey. An assessment of features related to phishing websites using an automated technique. in 2012 International Conference for Internet Technology and Secured Transactions. 2012.

[7] Abdelhamid, N., A. Ayesh, and F. Thabtah, Phishing detection based Associative Classification data mining. Expert Systems with Applications, 2014. 41(13): p. 5948 5959.

[8] Toolan, F. and J. Carthy. Feature selection for Spam and Phishing detection. in 2010 eCrime Researchers Summit. 2010.

[9] Jain, A.K. and B.B. Gupta, Towards detection of phishing websites on client-side using machine learning based approach. Telecommunication Systems, 2018. 68(4): p. 687-700.

[10] 1PhishTank, Phishing dataset. 2018, Verified phishing URL.

[11] Openfish, Phishing dataset. 2018.

[12] 1Chiew, K.L., et al., Utilisation of website logo for phishing detection. Computers & Security, 2015. 54: p. 16 26.

[13] Benavides, E., et al. Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review. 2020. Singapore: Springer Singapore.

[14] Zhang, W., et al., Two-stage ELM for phishing Web pages detection using hybrid features. World Wide Web, 2017. 20(4): p. 797-813. 15. Wide Web, 2017. 20(4): p. 797-813.

[15] El-Alfy, E. S.M., Detection of phishing websites based on probabilistic neural networks and Kmedoids clustering. The Computer Journal, 2017. 60(12): p. 1745-1759.

[16] Montazer, G.A. and S. ArabYarmohammadi, Detection of phishing attacks in Iranian e-banking using a fuzzy–rough hybrid system. Applied Soft Computing, 2015. 35: p. 482-492.

[17] Dr. Nilesh B. Korade, Amol C. Jadhav, "Exploring NLP Techniques for Duplicate Question Detection to Maximizing Responses on Q&A Websites",International Journal of Intelligent system and applications in engineering,Vol. 12 No. 3 (2024), ISSN:2147-6799,

[18] Wang, Y.-G., G. Zhu, and Y.-Q. Shi, Transportation spherical watermarking. IEEE Transactions on Image Processing, 2018. 27(4): p. 2063-2077.

[19] Amol C. Jadhav, A. M. Pawar, “Enhancement in Phishing Detection Using Features Clustering”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 6, ISSN: 2277 128X, June 2016

[20] De Maio, C., et al., Social media marketing through timeΘaware collaborative filtering. Concurrency and Computation: Practice and Experience, 2018. 30(1): p. e409