JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Phishing Website Detection Using Machine Learning Algorithms

[1]Pulkit Garg, [2]Kartikay, [3]Ankush Kalsotra, [4]B. Mohit Manihara, [5]Padmashree T.

[1]pulkitgarg.is20@rvce.edu.in, [2]kartikay.is20@rvce.edu.in, [3]ankushkalsotra.is20@rvce.edu.in, [4]mohitmaniharab.is20@rvce.edu.in, [5]padmashreet@rvce.edu.in
[12345]Information science and engineering,
[12345]RV College of Engineering, Bengaluru, Karnataka

Abstract : **Phishing websites represent a substantial cybersecurity threat, resulting in financial losses, data breaches, and compromised user privacy. Traditional detection methods are struggling to keep up with the constantly evolving tactics of cybercriminals. Thus, employing machine learning emerges as the most effective approach for automatically identifying phishing websites. In this the dataset used is from Kaggle, consists of 31 features and 11054 rows describing them with both phishing and legitimate data, 80:20 split for training and testing. Six machine learning classifiers are incorporated to train the data that are logistic regression, gradient boost, naive bayes classifier, decision tree, random forest and support vector machine with the highest accuracy of 97.4 using the gradient boost classifier.**

IndexTerms **- .Phishing, machine learning, legitimate, cybersecurity, classifier.**

I. INTRODUCTION

Phishing attacks are widespread in the online realm and continue to pose a serious danger to cybersecurity, putting both organizations and individuals at serious risk. These malicious schemes, which often commence by fraudulent websites or emails, try to dupe individuals into revealing private information, such as bank account details and login credentials. Although security measures have advanced, traditional detection techniques often do not keep up with the ever-evolving intelligence of cybercriminal operations, leaving consumers open to exploitation. The incorporation of machine learning classifiers is now a crucial tactic for strengthening defenses against phishing attacks in response to this serious challenge. Machine learning techniques discover patterns that are suspicious of phishing activities independently by using datasets that include a variety of features gathered from URLs and website properties making it possible for immediate recognition of threats.

A paradigm shift in cybersecurity has occurred with the use of machine learning for phishing detection, which provide a proactive and flexible defense against ever-evolving threats. Machine learning models can differentiate between harmful and genuine URls with greater accuracy by analyzing data including domain attributes, URL characteristics, and website content. This allows for quicker response times. It examines the effectiveness of machine learning algorithms in identifying phishing websites, revealing the effectiveness of these algorithms at detecting and preventing phishing attacks and protecting the internet from malicious use.

II. LITERATURE SURVEY

Studies have explored the use of machine learning for phishing detection those are as follows.

[1]In the year 2022 Adarsh, Saikiran, Vishnu and R Kavitha, focus on extraction of feature techniques from URLs and website characteristics, including domain-based, HTML/JavaScript-based, and address bar-based features. It employ the algorithms such as Decision Trees and Random Forests, achieving high accuracies of 87% and 82.4%. The study emphasizes ensemble learning techniques and to enhance accuracy and address zero-hour phishing attacks.

[2]In the year 2022 Safa, Ghina and Afnan Mohamed emphasizes on applying different machine learning methods, such Support Vector Machine, Random Forest Tree, XGBoost, Naive Bayes, K-Nearest Neighbors, AdaBoost, and Gradient Boosting for classification tasks. By analysing datasets with features of URLs and website characteristics, it aims to enhance detecting performance and accuracy while decreasing the rate of false positives and negatives. Comparative analyses highlight the effectiveness of XGBoost in achieving high accuracy rates.

[3]Aniket, Namrata, Sammed, Twinkle and Sandeep Gorein in 2021 evaluates various algorithms, including k-Nearest Neighbors, Naive Bayes, Decision Trees, and Gradient Boosting, for phishing website detection. Through comparative analysis, decision trees emerge as the preferred approach, exhibiting superior efficacy in phishing detection.

[4]Nishitha, Revanth, Mourya Vardhan and Kumaran in year 2023 proposes the study with the help of machine learning techniques to reduce the phishing attacks, emphasizing the shortcomings of traditional detection methods like blacklisting and heuristic-based approaches. Through meticulous data preprocessing and evaluation of multiple learning algorithms such as logistic regression and convolutional neural networks, it achieves better accuracy rates, with logistic regression and CNN reaching 95% and 96%, respectively.
[5]In the year 2019 Nathezhtha, Sangeetha and Vaidehi, emphasizes the importance of employing a multiple approach, including software-based solutions, hardware-based approaches, machine learning techniques, heuristic analysis, and web crawling strategies, to effectively combat phishing attacks. Their study explores various methodologies aimed at identifying and reducing phishing attacks, highlighting the effectiveness of web crawling in identifying zero-day phishing attacks.
[6]AD. Kulkarni and Brown in the year 2019 provide a comprehensive examination of various methodologies and algorithms for phishing website detection. They discuss a range of techniques, including Support Vector Machines, Decision Trees, Naïve Bayes' classifier, and Neural Networks, addressing challenges such as feature extraction from URLs and classification based on different attributes.
[7]In the year 2023 Mahesh, Ananth, and Dheepthi examines a diverse dataset of over 651,191 URLs, employing a range of features including URL length, presence of symbols, and hostname length. Through rigorous experimentation, they identify the multi-layer perceptron (MLP) architecture as an effective model, achieving an impressive accuracy of 95.6%.
[8]Deshpande, Atharva, Omka, Nachike, and Dr. Swapna in 2021 focus on the evolving nature of phishing and the importance of robust detection mechanisms in the face of advancing technology. Gaps persist in understanding the effectiveness of the techniques across different contexts and integrating processing methods for enhanced detection accuracy. It aims to bridge these gaps by providing a comprehensive overview of existing approaches, delineating key features, methodologies, and challenges in the field, with the ultimate goal of informing the development of more effective anti-phishing solutions.

III. METHODOLOGY

The methodology consists of the data preprocessing and training the machine learning classifiers.

A. *Dataset Preprocessing*

The Data set contains 11054 URLs with 6157 as phished and rest legitimate. The data is collected from Kaggle. Out of 31, 30 features are extracted using packages like BeautifulSoup with python programming language.
The database is trained on the features UsingIP, LongURL, ShortURL, Symbol@, Redirecting//, PrefixSuffix-, SubDomains, HTTPS, DomainRegLen, Favicon, NonStdPort, HTTPSDomainURL, RequestURL, AnchorURL, LinksInScriptTags, ServerFormHandler, InfoEmail, AbnormalURL, WebsiteForwarding, StatusBarCust, DisableRightClick, UsingPopupWindow, IframeRedirection, AgeofDomain, DNSRecording, WebsiteTraffic, PageRank, GoogleIndex, LinksPointingToPage and StatsReport.
The data is split into 80:20 ratio for the training and testing respectively. The training data is used to train the models on the features mentioned above and the testing data is used to analyse the performance and accuracy of the model.

B. *Classifiers*

Logistic Regression It is used for solving problems related to classifications, which predicts the output as categorial dependent variable. Hence the outcome will be either categorial or discrete value.
LogisticRegression() creates an instance of the logistic regression classifier. It is used for binary classification tasks. It models the possibility that a given input belongs to a particular class.
Support Vector Machine It creates a decision boundary that segregate the data in n dimensional space into classes so that it can be placed in correct category.
Kernel - rbf, linear are used which will try two kernels: RBF (nonlinear) and linear to get the best space. It is supervised learning algorithm which used for both classification as well as regression problems.
Naïve Bayes algorithm It makes quick predictions. It is based on Bayes theorem and used for solving classification problems.
GaussianNB() creates an instance of the Gaussian Naive Bayes classifier. It assumes that features follow a Gaussian distribution.

	precision	recall	f1-score	support
-1	0.94	0.91	0.92	976
1	0.93	0.95	0.94	1235
accuracy			0.93	2211
macro avg	0.93	0.93	0.93	2211
weighted avg	0.93	0.93	0.93	2211

	precision	recall	f1-score	support
-1	0.97	0.94	0.96	976
1	0.96	0.98	0.97	1235
accuracy			0.96	2211
macro avg	0.97	0.96	0.96	2211
weighted avg	0.96	0.96	0.96	2211

	precision	recall	f1-score	support
-1	0.97	0.94	0.96	976
1	0.96	0.98	0.97	1235
accuracy			0.96	2211
macro avg	0.97	0.96	0.96	2211
weighted avg	0.96	0.96	0.96	2211

Fig. 1 Classification of logistic regression, support vector machine and naïve bayes respectively

Decision Trees It is mostly used for solving Classification problems. It is a tree-structured classifier, where dataset features are represented by the node, decision rules are represented by branches and each leaf node represents the outcome.
DecisionTreeClassifier(max_depth=30) creates an instance of the decision tree classifier. The maximum depth is set to 30, which means the tree will grow till the depth of 30 or until all leaves are pure.

Random Forest It is based on the concept of ensemble learning. It uses various decision trees on different subsets of a data and averages the results to increase the accuracy.
RandomForestClassifier(n_estimators=10) creates an instance of the random forest classifier that consists of 10 decision trees.

Gradient boosting classifiers It combines different weak learning models to create a strong model. Decision trees are used when doing gradient boosting. Boosting algorithms play a role in dealing with bias variance trade-off, boosting controls both the aspects (bias & variance), and is considered to be more effective.
GradientBoostingClassifier(max_depth=4, learning_rate = 0.7) It creates an instance of the Gradient Boosting classifier with maximum depth of the individual as 4 and the contribution of each 0.7.

	precision	recall	f1-score	support
-1	0.95	0.95	0.95	976
1	0.96	0.96	0.96	1235
accuracy			0.96	2211
macro avg	0.96	0.96	0.96	2211
weighted avg	0.96	0.96	0.96	2211

	precision	recall	f1-score	support
-1	0.97	0.95	0.96	976
1	0.96	0.98	0.97	1235
accuracy			0.97	2211
macro avg	0.97	0.97	0.97	2211
weighted avg	0.97	0.97	0.97	2211

	precision	recall	f1-score	support
-1	0.99	0.96	0.97	976
1	0.97	0.99	0.98	1235
accuracy			0.97	2211
macro avg	0.98	0.97	0.97	2211
weighted avg	0.97	0.97	0.97	2211

Fig. 2 Classification of decision trees, random forest gradient boosting respectively

The website is built using Flask for URL checking, requiring users to log in for authentication. After logging in, users can check whether a URL is legitimate or a phishing attempt. A Gradient Boosting classifier is utilized for this task due to its high accuracy and performance.

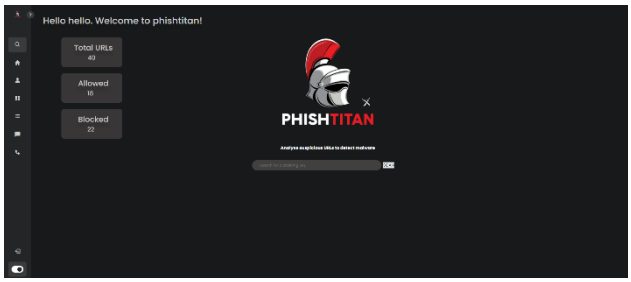

Fig. 3 Phishtitan website dashboard for checking URLs

Once a URL is analyzed, if it is deemed legitimate, the user can access the website. If the URL is detected as phishing, it will be blocked. The website also features dynamic counting of both blocked and safe URLs, and the dashboard displays the top seven most searched URLs. This system provides an effective and user-friendly method for ensuring online safety by preventing access to malicious websites.

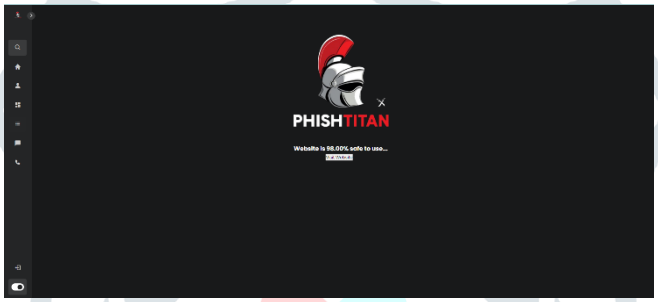

Fig. 4 Result after checking a legit URL

In Fig. 4, the legit URL is checked, and the website indicates that it is safe to use.

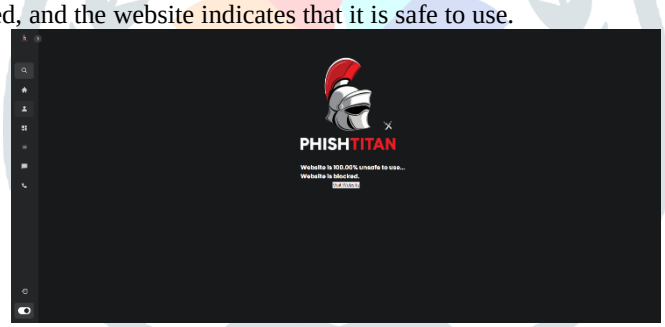

Fig. 2 Result after checking a legit URL

In the Fig. 5 the phished URL is used for checking and it gives the result that it is unsafe, therefore blocking the website.

IV. RESULTS

The dataset with 30 features is extracted and split in 20-80 for testing and training. The training data is used in logistic regression, gradient boost classifier, support vector machine, naive bayes classifier, decision tree and random forest machine learning algorithms and to find the accuracy and performance, testing data is used. Comparative analysis is performed to get the best accuracy model i.e. Gradient boost classifier with 97.4 accuracy.

V.

	ML Model	Accuracy	f1_score	Recall	Precision
0	Gradient Boosting Classifier	0.974	0.977	0.994	0.986
1	Random Forest	0.967	0.970	0.994	0.989
2	Support Vector Machine	0.964	0.968	0.980	0.965
3	Decision Tree	0.960	0.964	0.991	0.993
4	Logistic Regression	0.934	0.941	0.943	0.927
5	Naive Bayes Classifier	0.605	0.454	0.292	0.997

Fig. 3 Comparative analysis on different machine learning algorithms

VI. Conclusion

This paper will help to minimize the phishing attacks as the machine learning algorithm are used to detect them effectively. The dataset used, is from Kaggle consisting of 31 features and 11054 rows of data with a train and test split of 80:20. The six different learning classifiers are used to train the dataset that are naive bayes classifier, decision tree, logistic regression, gradient boost classifier, support vector machine and random forest. Gradient boost classifier has the highest accuracy of 97.4 with the highest performance data.

REFERENCES

[1] Mandadi, A., Boppana, S., Ravella, V., & Kavitha, R. (2022). Phishing Website Detection Using Machine Learning. In 2022 IEEE 7th International Conference for Convergence in Technology (I2CT) IEEE.

[2] Alrefaai, S., Özdemir, G., & Mohamed, A. (2022). Detecting Phishing Websites Using Machine Learning. In 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) (pp. 1-6). IEEE.

[3] Garje, A., Tanwani, N., Kandale, S., Zope, T., & Gore, S. (2021). In-depth study of detection of phishing URLs using machine learning. International Research Journal of Modernization in Engineering Technology and Science, 03(04), 1-6.

[4] Nishitha, U., Kandimalla, R., Mourya Vardhan, R. M., & Kumaran, U. (2023). Phishing detection using machine learning techniques. In 2023 3rd Asian Conference on Innovation in Technology (ASIANCON) (pp. 1-6). IEEE.

[5] T. Nathezhtha, D. Sangeetha and V. Vaidehi, ” WC-PAD: Web Crawling based Phishing Attack Detection ” 2019 International Carnahan Conference on Security Technology (ICCST).

[6] AD. Kulkarni and L. L. Brown III, "Phishing Websites Detection using Machine Learning," International Journal of Advanced Computer Science and Applications, vol. 10, no. 7, pp. 8-14, 2019.

[7] Mahesh, Ananth, and Dheepthi. "Using Machine Learning to Detect and Classify URLs: A Phishing Detection Approach." In 4th International Conference on Electronics and Sustainable Communication Systems (ICESC), July 6-8, 2023. IEEE, 2023.

[8] Deshpande, Atharva, Omkar Pedamkar, Nachiket Chaudhary, and Dr. Swapna Borde. "Detection of Phishing Websites using Machine Learning." International Journal of Engineering Research & Technology (IJERT), vol. 10, no. 05, May 2021, pp. 430-432. ISSN: 2278-0181.