International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307

OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2410228

Phishing Detection Using Machine Learning Algorithm

Vishesh Bharuka, Allan Almeida, Sharvari Patil

Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai, Maharashtra, India

ARTICLEINFO

Article History:

Accepted: 15 March 2024

Published: 30 March 2024

Publication Issue

Volume 10, Issue 2

March-April-2024

Page Number

343-349

ABSTRACT

Phishing is a criminal scheme to steal the user's personal data and other credential information. It is a fraud that acquires victim's confidential information such as password, bank account detail, credit card number, financial username and password etc. and later it can be misuse by attacker. The use of machine learning algorithms in phishing detection has gained significant attention in recent years. This research paper aims to evaluate the effectiveness of various machine learning algorithms in detecting phishing URL's/website. The algorithms tested in this study are Decision Tree, Random Forest, Multilayer Perceptron, XGBoost, Autoencoder Neural Network, and Support Vector Machines. A dataset of phishing URLs is used to train and test the algorithms, and their performance is evaluated based on metrics such as accuracy, precision, recall, and F1 Score. The paper takes in data of phished URL from Phishtank and legitimate URL from University of New Brunswick. The results of this study demonstrate that the Random Forest and XGBoost algorithms outperforms other algorithms in terms of accuracy and other performance metrics and the system has an overall accuracy of 98 %.

Keywords : Phishing Detection, Feature Collection, Feature Selection, Classification, Machine Learning, Explainable AI, Data Sets.

I. INTRODUCTION

In everyday life interacting with website has turned into a typical way of communication. It operates over the Internet or any computer networks. Spam websites contain the link to the phishing sites or any malwares. Phishing is the attempt to obtain sensitive information such as usernames, passwords, and credit card details (and sometimes, indirectly, money), often for malicious reasons, by masquerading as a trustworthy entity in an electronic communication [12].

Phishing URL's may contain links to websites that are infected with malware. Phishing is typically carried out by email spoofing or instant messaging and it often directs users to enter details at a fake website whose look and feel are almost identical to the legitimate one [3]. Phishing, is an example of social engineering techniques used to deceive users, and it exploits the poor usability of current web security technologies.

Attempts to deal with the growing number of reported phishing incidents include legislation, user training, public awareness, and technical security measures. Many websites have now created secondary tools for applications, like maps for games, but they should be clearly marked as to who wrote them, and users should not use the same passwords anywhere on the internet.

Phishing is a continual threat, and the risk is even larger in social media such as Facebook, Twitter, and Google. Hackers could create a clone of a website and tell you to enter personal information, which is then emailed to them. Hackers commonly take advantage of these sites to attack people using them at their workplace, homes, or in public in order to take personal and security information that can affect the user or company (if in a workplace environment) [11]. Phishing takes advantage of the trust that the user may have since the user may not be able to tell that the site being visited, or program being used, is not real; therefore, when this occurs, the hacker has the chance to gain the personal information of the targeted user, such as passwords, usernames, security codes, and credit card numbers, among other things.

However, phishing has become more and more complicated and sophisticated so that phishers can bypass the filter set by current anti-phishing techniques and cast their bait to customers and organizations. A possible solution is to create a web extension to enhance the phishing URL's detection. By analysing phished websites, it is observed that phishing websites/messages often include certain phrases, for example, 'security', 'verify your account', 'if you don't update your details within 2 days, your account will be closed', 'click here to access to your account' and so on. Such terms are useful to classify if a website is a phished website. In addition, Phishing URL's often alert customer to click links to other websites which the real link is not the same as it is shown in the page.

II. METHODS AND MATERIAL

A. Literature Review

In the existing papers, the authors have used methods which are a combination of various machine learning algorithms. Following are the various methodologies and approaches that are used to achieve desired results.

In paper [7] the author presents a real-time anti-phishing system using seven classification algorithms and NLP based features. It offers language independence, detects new websites, and achieves 97.98% accuracy in identifying phishing URLs. The system overcomes limitations of existing anti-phishing methods, providing effective protection against cyber threats in electronic commerce.

The paper [1] focuses on combating phishing sites by employing machine learning algorithms and a new dataset consisting of 5000 legitimate web-pages and 5000 phishing ones. After testing various machine learning algorithms, J48, Random Forest, and Multilayer Perceptron were selected. The researchers utilized feature selection tools to enhance model efficiency, and the best results were obtained using 20 features out of 48 with the Random Forest algorithm, achieving an impressive accuracy of 98.11% in phishing detection.

In paper [2] the author introduces a framework for detecting phishing websites using a stacking model. Phishing, a fraudulent practice aimed at stealing users' credentials and personal information for financial gain, impacts various fields like e- commerce, online business, banking, and digital marketing. The proposed approach involves analysing the phishing dataset using feature selection techniques like information gain, gain ratio, Relief-F, and recursive feature elimination (RFE). Two features, combining the strongest and weakest attributes, are suggested. Principal component analysis is then applied, followed by different machine learning algorithms (random forest, neural network, bagging,

support vector machine, Naïve Bayes, and k-nearest neighbour) to process the proposed and remaining features. Finally, two stacking models, Stacking1 (RF + NN + Bagging) and Stacking2 (kNN + RF + Bagging), are employed to combine the highest-scoring classifiers and enhance the overall classification accuracy.

The paper [11] provides a comprehensive review of research in the field of feature selection for phishing detection. It was published in the International Journal of Intelligent Systems Technologies and Applications in 2016. This paper explores various methodologies for feature selection in the detection of phishing attacks, which are fraudulent attempts to acquire sensitive information by pretending to represent a trustworthy entity. Through analysing past research, it intends to identify effective strategies and gaps in the current knowledge. Specific authors' research interests include areas like cyber-security, software engineering, and soft computing, which contextualize this work.

The survey paper [12] discusses techniques in machine learning for handling large data sets with an extensive amount of irrelevant information. It focuses on two main challenges: relevant feature selection and suitable example selection. The paper outlines progress made in this area from both an empirical and theoretical standpoint. A common framework for comparing various methods is also presented. Future work regarding these challenges is proposed for exploration.

B. Architecture

Several procedures are followed in the design of a model for phishing detection. It is important to use a suitable design process while creating a system that produces accurate outcomes. The details of these actions are provided below. To create an effective model, it is usually advisable to follow these procedures.

1.) Feature Collection:

Features like IP address, URL length, use of "-," age of domain, DNS record, and other characteristics are gathered when building one's own dataset. These options are stored in a file in a CSV format for identifying novel features that can be used for phishing detection, and developing more robust and accurate phishing detection systems.

Limitations: The effectiveness of these approaches may depend on the choice of features, datasets, and machine learning algorithms used.[6] There is still room for further research in this area, particularly in the development of more robust and accurate machine learning models for phishing detection.

Some people have created datasets with just binary or ternary values using rules. According to these criteria, which compared the feature, each URL's value in the dataset is either 0,1, or -1, with 1 denoting phishing, 0 suspicious activity, and -1 genuine activity. This procedure is repeated until the model receives a URL for validation and we need to gather the features in order to test it. So, if using a pre-made dataset, this step is only necessary during the testing stage.

2.) Feature Selection:

When utilising machine learning to detect phishing, a set of attributes that have been retrieved from phishing emails or websites are analysed to identify the malicious sources. In feature selection, the most pertinent and instructive traits are chosen from a potentially huge pool of attributes. Its main goal is to improve the accuracy, effectiveness, and interpretability of machine learning models. The selection of features is a crucial phase in the classification process because it eliminates the extraneous information that add little to the classification process and instead raise the false negative rate by producing results that are improperly categorized. On this basis we have used explainable AI

to select the most important 17 features of a URL from the pool of various features.

3.) Classification:

The ultimate stage of the system flow, which produces the output result, is this one. On the basis of the 17 features which are as follows domain length of URL , IP Address in URL, length of URL, depth of URL, Redirection "//" in URL, http/https" in Domain name, Using URL Shortening Services “Tiny URL”, Prefix or Suffix "-" in Domain, DNS Record, Website Traffic, Age of Domain, End Period of Domain, IFrame Redirection, Status Bar Customization, Disabling Right Click, Website Forwarding and after data processing is done on the model the various ml algorithms such as Decision Tree, Random Forest, Multilayer Perceptron, XGBoost, Autoencoder Neural Network, and Support Vector Machines are applied on the model. After which the model is extracted and converted into a web based website using flask where user can enter the URL of the site to check if it is phished or not.

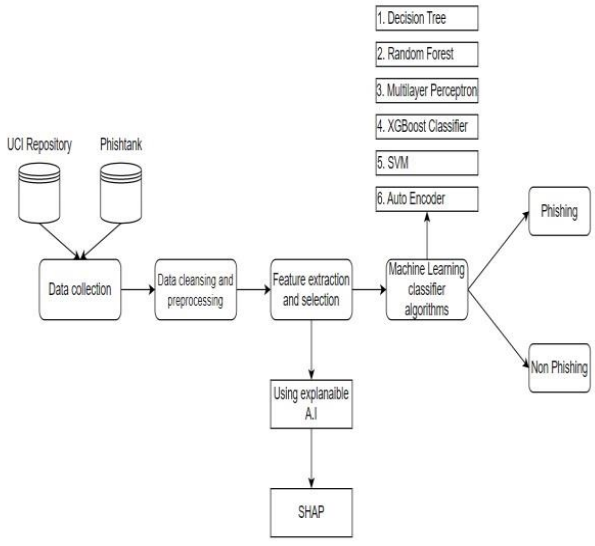

Figure 1. System Architecture

Figure 1 shows the architecture diagram for detecting phishing website. We initially collected data regarding phishing URL’s and legitimate URLs from various online sites.

Then we worked on preprocessing the data by removing null values from the data and also removed duplicate data from our dataset which was degrading our model’s performance. We used various techniques for detecting essential features and various machine learning algorithm to find whether URL is phished or not.

4.) Implementation:

In the following project the set of phishing URLs are collected from an opensource service called PhishTank. From this dataset, 5000 random phishing URLs are collected to train the ML models. The legitimate URLs are obtained from the open datasets of the University of New Brunswick. This dataset has a collection of benign, spam, phishing, malware & defacement URLs. From this dataset, 5000 random legitimate URLs are collected to train the ML models in an 80-20 test ratio. After which we have used explainable AI to detect 17 features from the URL of the phished site. We have basically combined the heuristic and ml approach for our phishing feature detection.

III.RESULTS AND DISCUSSION

We studied all the different kinds of approaches available to detect phishing and also the in-use Anti-phishing tools. The ML technique for detecting spam turned out to be less useful for newly registered domains. The corresponding bar graph shows the dependency of our program on the various feature that we have selected according to this URL length is one of the most important features.

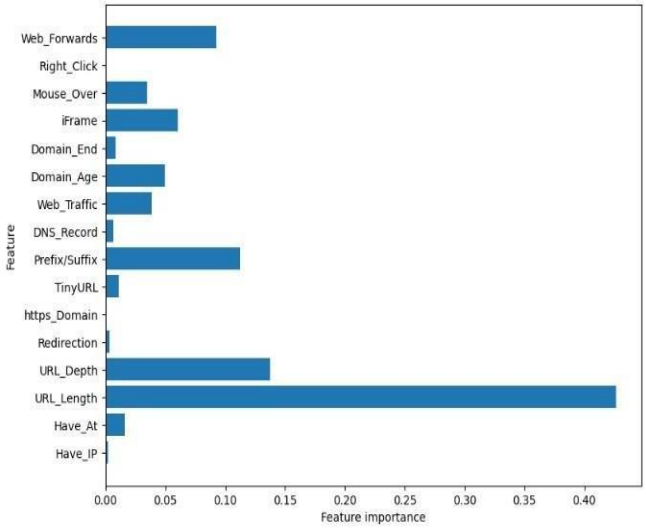

Figure 2. Different features in a URL and their importance

Figure 2 provides us with the contribution of a particular feature in determining whether a URL is phishing one or not. From the figure we can conclude that URL_Length contributes maximum in making a decision.

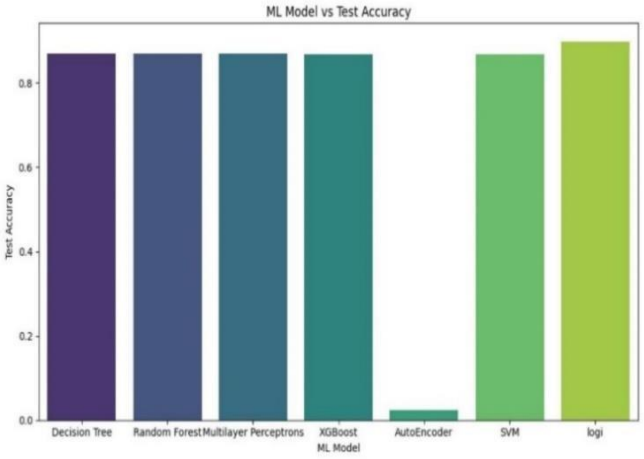

Figure 3. Results

Hence, we have compared the various machine learning algorithms and found that XG boost provide the best result. In figure 3 we can see comparison of accuracy of various machine learning algorithms on the basis of training and testing.

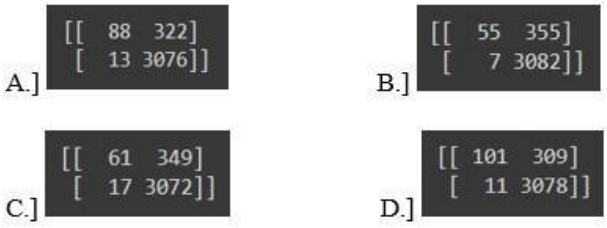

Figure 4. Confusion Matrices of various algorithms

A confusion matrix is a table that visualizes and summarizes the performance of a classification algorithm. Figure 4 represents the performance of algorithms like A) Multilayer Perceptron B) Random Forest C) Support Vector Machine D) XGBoost Classifier.

A confusion matrix is a tool used in machine learning and statistics to evaluate the performance of a classification model. It provides a summary of the model's predictions by comparing them to the actual values in a dataset. The matrix is particularly useful when dealing with binary classification problems, where we try to classify the data into two categories, typically labelled as "positive" and "negative."

Here's how a confusion matrix is structured:

- True Positives (TP): The number of correctly predicted positive instances.
- True Negatives (TN): The number of correctly predicted negative instances.
- False Positives (FP): The number of negative instances incorrectly predicted as positive (Type I error).
- False Negatives (FN): The number of positive instances incorrectly predicted as negative (Type II error).

To calculate a confusion matrix:

1. Make predictions using your classification model on a dataset.
2. Compare the model's predictions to the actual labels in the dataset.

3. Count the number of true positives, true negatives, false positives, and false negatives.

Once you have these values, you can construct the confusion matrix. Here is how it looks:

	Actual Positive	Actual Negative
Predicted	TP	FP
Positive	FN	TN

To calculate metrics like accuracy, precision, recall (sensitivity), specificity, and F1-score, you can use the values from the confusion matrix:

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Precision = TP / (TP + FP)
- Recall (Sensitivity) = TP / (TP + FN)
- Specificity = TN / (TN +FP)

IV.CONCLUSION

The definitive results highlight the exceptional effectiveness of the XGBoost algorithm, producing a remarkable accuracy rate of 98.4% in identifying phishing websites. They also demonstrate how explainable AI can be utilized to enhance feature selection. This significant accomplishment confirms the effectiveness of machine learning in thwarting cyber threats and establishes XGBoost as a strong candidate for improving detection accuracy. The study's ramifications include improving machine learning-based defences against phishing attempts, which calls for additional investigation into parameter optimization to increase real-world adaptability. These findings support the ability of machine learning approaches to successfully handle and prevent the ongoing problem of phishing assaults in the ever-changing landscape of cyber security.

V. REFERENCES

[1]. A. Almseidin, Mohammad, AlMaha Abu Zuraiq, Mouhammd Al-Kasassbeh, and Nidal Alnidami. "Phishing detection based on machine learning and feature selection methods." (2019): 171-183.

[2]. Zamir, Ammara, Hikmat Ullah Khan, Tassawar Iqbal, Nazish Yousaf, Farah Aslam, Almas Anjum, and Maryam Hamdani. "Phishing web site detection using diverse machine learning algorithms." The Electronic Library 38, no. 1 (2020): 65-80.

[3]. Jain, Ankit Kumar, and Brij B. Gupta. "Phishing detection: analysis of visual similarity based approaches." Security and Communication Networks 2017 (2017).

[4]. Gandotra, Ekta, and Deepak Gupta. "An efficient approach for phishing detection using machine learning." Multimedia Security: Algorithm Development, Analysis and Applications (2021): 239-253.

[5]. Jain, Ankit Kumar, and Brij B. Gupta. "Towards detection of phishing websites on client-side using machine learning based approach." Telecommunication Systems 68 (2018): 687-700.

[6]. Yadav, Neelam, and Supriya P. Panda. "Feature selection for email phishing detection using machine learning." In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 2, pp. 365-378. Springer Singapore,2022.

[7]. Sahingoz, Ozgur Koray, Ebubekir Buber, Onder Demir, and Banu Diri. "Machine learning based phishing detection from URLs." Expert Systems with Applications 117 (2019): 345-357.

[8]. Abdulraheem, Rana, Ammar Odeh, Mustafa Al Fayoumi, and Ismail Keshta. "Efficient Email phishing detection using Machine learning." In

2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0354-0358. IEEE, 2022.

[9]. Tanimu, Jibrilla, and Stavros Shiaeles. "Phishing Detection Using Machine Learning Algorithm." In 2022 IEEE International Conference on Cyber Security and Resilience (CSR), pp. 317-322. IEEE, 2022.

[10]. Mithra Raj, Mukta, and J. Angel Arul Jothi. "Website Phishing Detection Using Machine Learning Classification Algorithms." In International Conference on Applied Informatics, pp. 219233.Cham: Springer International Publishing, 2022.

[11]. Zuhair, H., Selamat, A. & Salleh, M. (2016). Feature selection for phishing detection: a review of research. International Journal of Intelligent Systems Technologies and Applications, 15(2), 147-162.

[12]. A.L. Blum and F. Langley, "Methods for Handling Large Amounts of Irrelevant Information in Machine Learning" in Artificial Intelligence, vol. 97, pp. 245271, 1997, Elsevier Science B.V.

[13]. Abu-Nimeh, Saeed, Dario Nappa, Xinlei Wang, and Suku Nair. "A comparison of machine learning techniques for phishing detection." In Proceedings of the anti-phishing working groups 2nd annual eCrime researchers' summit, pp. 60-69. 2007.

[14]. Shahrivari, Vahid, Mohammad Mahdi Darabi, and Mohammad Izadi. "Phishing detection using machine learning techniques." arXiv preprint arXiv:2009.11116 (2020).

[15]. Crawford, Michael, Taghi M. Khoshgoftaar, Joseph D. Prusa, Aaron N. Richter, and Hamzah Al Najada. "Survey of review spam detection using machine learning techniques." Journal of Big Data 2, no. 1 (2015): 1-24.

[16]. Rashid, Junaid, Toqeer Mahmood, Muhammad Wasif Nisar, and Tahira Nazir. "Phishing detection using machine learning technique." In 2020 first international conference of smart systems and emerging technologies (SMARTTECH), pp. 43-46. IEEE, 2020.

[17]. Yi, Ping, Yuxiang Guan, Futai Zou, Yao, Wei Wang, and Ting Zhu. "Web phishing detection using a deep learning framework." Wireless Communications and Mobile Computing 2018 (2018).

[18]. Kumar, Nikhil, and Sanket Sonowal. "Email spam detection using machine learning algorithms." In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 108-113. IEEE, 2020.

[19]. Abdelhamid, Neda, Fadi Thabtah, and Hussein AbdelJaber. "Phishing detection: A recent intelligent machine learning comparison based on models content and features." In 2017 IEEE international conference on intelligence and security informatics (ISI), pp. 72-77. IEEE, 2017.

[20]. Yadollahi, Mohammad Mehdi, Farzaneh Shoeleh, Elham Serkani, Afsaneh Madani, and Hossein Gharaee. "An adaptive machine learning based approach for phishing detection using hybrid features." In 2019 5th International Conference on Web Research (ICWR), pp. 281-286. IEEE, 2019.