e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

PHISHING DETECTION APPROACH USING MACHINE LEARNING

Deepak Pathak*1, Mohammad Ammar*2, Mohit Bhandari*3

*1,2,3Department Of Computer Science & Engineering, Meerut Institute Of Engineering & Technology, Meerut, India.

ABSTRACT

In today's world, we are witnessing that the ever-growing advancement of technologies like e-commerce, e-banking, e-registration, etc had vast impact in a lot of factor on ours life. Attacks caused by Phishing have promptly manifested as a prime issue of cybersecurity. Fake web pages or phishing websites developed by attackers to fool and rob vital information of users such as username and password. However there are a number of methods to predict phishing, phisher's tactics were developed to avoid being detected. The most suitable way of predicting phishing is machine learning as maximum phishing attacks have common attributes which can be easily identified by machine learning algorithm. However, the precise capturing of fake webpages is a difficult topic as being directly proportional to dynamic aspects. Our study unfolds the Decision Tree (DT) classifier consisting significant attributes selection, to identifying fake websites with the aim of enriching the classification of webpages as fake or legal webpages. To perform the experiments we have used a publically accessible phishing website dataset from the UCI machine learning repository, which contains 4899 phishing webpages and 6158 legal webpages. In our study, our team firstly gather attributes form the dataset and then we train our DT model and at last we test it and we have achieved 98%(approx) accuracy by our feature selection technique, which surpassed the DT classification when compared to other feature selection techniques.

Keywords: Phishing Websites, Machine Learning, Decision Tree, Feature Extraction.

I. INTRODUCTION

Phishing is a social engineering approach targeted at-> Gathering the confidence of the victim to share his/her details such as username, email address, financial information or password, etc. The attackers adopt these details to harm the victim. [2, 3].

The "blacklist" method is a comprehensive strategy to differentiate phishing websites which are updating anti-virus databases with Internet Protocol (IP) addresses and blacklisted URLs. The vital con which is the part of this technique is that it fails to recognize 0-hour phishing attacks. Analytic-based identification, with attributes prominent in legitimate-world phishing attacks, is able to recognize 0-hour phishing attacks, but the attributes were not guaranteed. Frequent attacks and the rate of false positives were detected is very high [4, 5].

To overcome the inabilities of heuristic-based and blacklist processes, many cyber-security groups/individuals are focusing on machine learning methods. Machine learning is a collection of approaches that account already existing techniques to anticipate upcoming outcomes. This activity is used by the professional to review a huge amount of banned URLs, their features to rightly identify fake webpages, together with 0-hour fake webpagess [4, 6].

Personal desktop users are exposed to phishing attacks. For five basic reasons:

(1) Clients are not aware of the Uniform Resource Locator (URL).

(2) The entire location of the page due to redirects or hidden URLs.

(3) No specific idea of which pages to trust.

(4) The URL consists of too many options or several pages may have been entered incorrectly.

(5) End users may not differentiate the fake webpage from the real/original ones [1].

A well-known company named Markmonitor which deals in brand insurance that certifies innovations. In Q3 2019, the main centered targeted victim by phishing will be web-mail websites and software as a services (SaaS). Phishers continued to collect evidence for these types of webpages, then start a business email contracts (BECs) & login to SaaS platforms [1].

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

Continuing with remaining research paper containing: Section 2 describes the related work on fake webpage identification; Section 3 provides proposed methodology (our approach) and briefly explain about the data, attributes and technique and at last Section 4 is part of our conclusion.

II. RELATED WORK

In this fast pace digital life, identity theft has arise as a major root of agitation for cyber-security professionals as it is relatively easy to generate a fake webpage by way of looks like an official webpage. Although experts cannot identify Fake webpages, everyone doesn't have the ability, resulting in a scenario of becoming a victim of cybercrime attacks. The only thought of the phisher is to steal the crucial credentials of the victim. Attacks on the crime of identity theft are on the rise effectively due to lack of user information. It is difficult to fight the crime of theft of sensitive information as it happens advantage of user vulnerability; however what all is important is that to develop crime catching environment for sensitive information [9, 10].

The ongoing part of the paper will reveal some the latest research reports on the topic of "Phishing Websites Detection Approach Using Machine Learning".

In Mahajan 2018, the authors letdown Phishing Websites Detection Approach of URLs/webpages by examining distinct Machine Learning Models such as: Random forest, Decision Tree, and Support Vector Machine (SVM) to reveal fake webpages. Their approach started with a collection of data having around 36,711 webpages consisting of 19653 phishing webpages and 17058 legitimate webpages. The URLs of fake webpages are picked from www.phishtank.com and URLs of legitimate webpages are picked from www.alexa.com. Datafile is divided into training set and testing set in different ratios. Experiments were conducted by using extracted features. It was resulted that the models Decision Tree, Random Forest, and SVM provided a performance prediction accuracy of 96.71%, 97.14% and 96.51% respectively. The best division result was accomplished using Random Forest algorithm having lowest and false negative [11].

In Kulkarni & Brown, 2019 authors recommended a phishing website detection model that uses number of classification stratergies which includes a decision tree, a Naive Bayes'classifier, SVM, and a Neural Network. This model was used in a data set containing nine features from The University of California, Irvine Machine Learning Repository. Among such datasets it includes elements from 1352 URLs. 702 are phishing attempts, 548 are legitimate, and 103 are suspect. The collected data also includes nine features appropriated from each URL. Experiments were carried out for each classifier. The finest classification results were obtained using Decision Tree, having classification accuracy of 90.39 % [12].

In Shahrivari, 2020 The authors puposed a model to categorize websites as phishing or legitimate by applying a various classification methods, including Support Vector Machine, Logistic Regression, Decision Tree, Gradient Boosting, Ada Boost, Neural Networks, KNN, XGBoost and Random Forest. This exemplary was applied on a dataset of phishing websites captured from the UCI Machine Learning Repository, which contains 4898 phishing websites and 6157 legitimate websites. This Experiments contains 30 features , and ten-fold cross-validation was engaged for training, verification, and testing. The model displayed performance prediction accuracy of 96.59 %, 97.26 %, and 98.32 %, for Decision Tree, Random Forest, and XGBoost respectively. As a result, we received satisfactory performance in ensemblaged classifiers such as XGBoost and Random Forest in terms of computation continuation and accuracy [13].

In Gadge, 2017 the authors Introduced an approach for identifying phishing URL websites. This technique inspects the websites and calculates heuristic values. By using C4.5 decision tree approach, these features used to determine whether the site was a phishing or not. Data from Google and PhishTank were used to contruct the dataset. This program contains two stages: pre-processing and detection. In the pre-processing step, features are obtained using rules, and then the features and their accompliced values are fed into the C4.5 algorithm, which produced an accuracy of 89.40 % [14].

III. PROPOSED METHODOLOGY

Data Collection: The very first step in building the introduced phishing website detection model is to choose a convenient training dataset which consists of both legitimate and phishing websites that are used to support and test the proposed system to appraise its performance. In this study, we evaluate the effectiveness of the proposed phishing website detection procedure using a publicly accessible phishing website dataset from the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

UCI Machine Learning Repository ("Phishtank,"2016.). This dataset consist of 6157 legitimate websites and 4898 phishing websites from by different website features were extracted. A collection of phishing websites was mostly taken from the Phishtank and MillerSmiles archives. Table 1 presents the key details of the phishing website dataset used in the tests and assessment.

Table 1: The description of the dataset of phishing websites utilized in the experiments

Attributes	Value
Number of features (attributes)	30
Number of websites (instance)	11057
Number of phishing websites	4899
Percentage of phishing websites	44%
Number of legitimate websites	6158
Percentage of legitimate websites	56%

Features Selection: Selecting the most appropriate features for the test will give a better result. Features are the important aspect of deal with phishing website detection study. The following are some of the features of our dataset:

Table 2: The list of features used

1. Having an IP Address	11. Using Non-Standard Ports	21. Disabling Right Click
2. Length of URL	12. HTTPS token	22. Using Pop-up Window
3. URL Shortening Service	13. Request URL	23. Iframe
4. Using the @ symbol	14 Anchor URL	24. Domain Age
5. Double Slash Redirection	15. Links in Tags	25. DNS Record
6. Prefix Suffix	16. SFH	26. Web Traffic
7. Using a Sub-domain	17. Submitting Information Via Email	27. Page Rank
8. SSL Status	18. Incorrect URL	28. Google Index
9. Domain Registration Length	19. Website Redirect Count	29. Number of Links Pointing To Page
10. Favicon	20. Status Bar Customization	30. Statistical Report

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

Machine learning (ML) for Phishing Attack Detection:

ML technologies are well-known for finding Phishing websites/ web-pages for stealing delicate information and this can lead to general segregation confusion. Teaching a machine learning model of learning-based value system, immediate data should have linked features, phishing misleads and authentic website classes. Dissimilar dividers are used to reveal criminal attacks for delicate information. Previous research suggests that the accuracy of detection is high as ML strategies are robust used. A few methods of selecting features are used to minimize features. Input data collection is provided as training input machine learning model for predicting the crime of stealing sensitive information or official traffic.

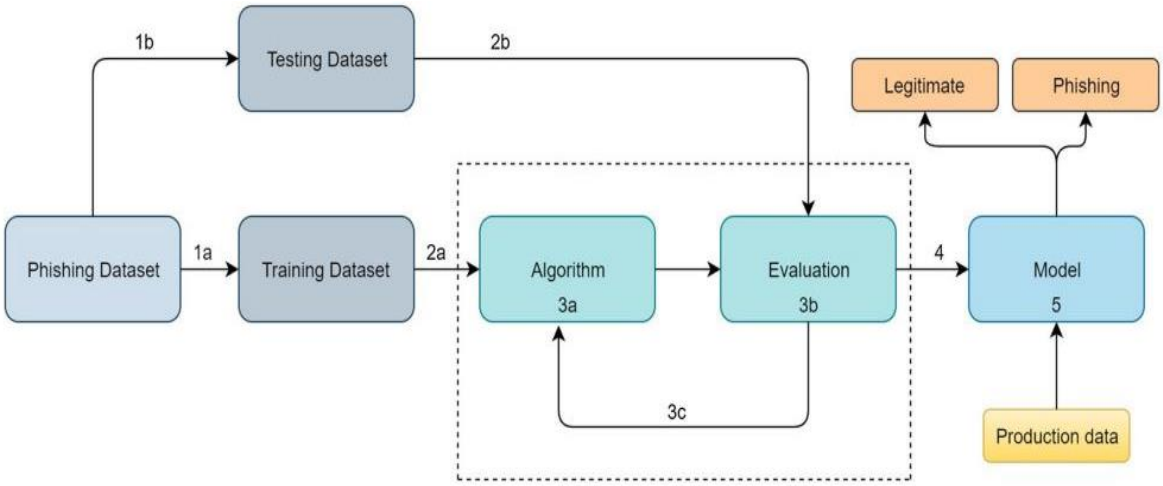

Fig 1: Proposed Methodology Flow Chart.

Decision Tree (DT)

- DT is a decision taking approach which consist a tree shaped structure having different and feasible result, such as Yes/No, True/False, Having/Not having. DT is the method of displaying the algorithms as complete conditional control expressions.
- DT is a method which falls under the category of Supervised Learning; DT is suitable for solving problems of both types of Supervised Problems that is (classification problems and Regression problems).
- But DT is promptly suggested for dealing with the Classification problems.
- The decisions or the tests are evaluated considering the features of the provided dataset.
- The General Structure of Decision Tree is given below-:

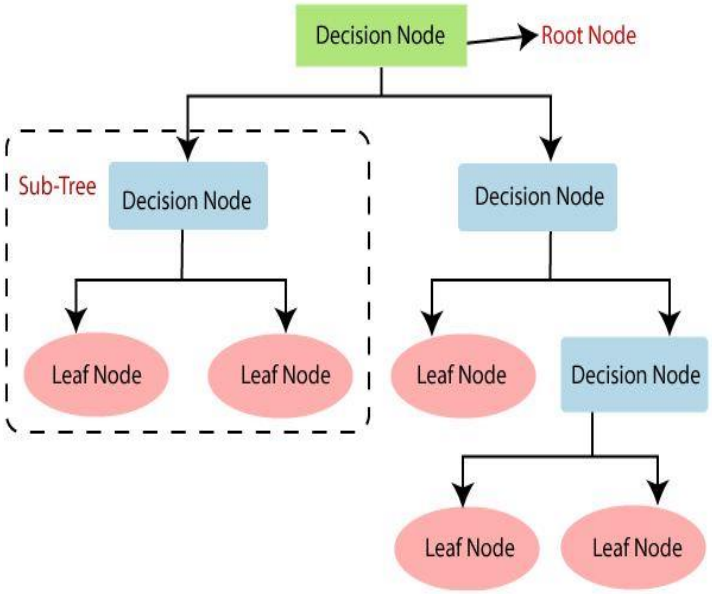

Fig 2: General structure of Decision Tree

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

Flow Chart of the Proposed Methodology

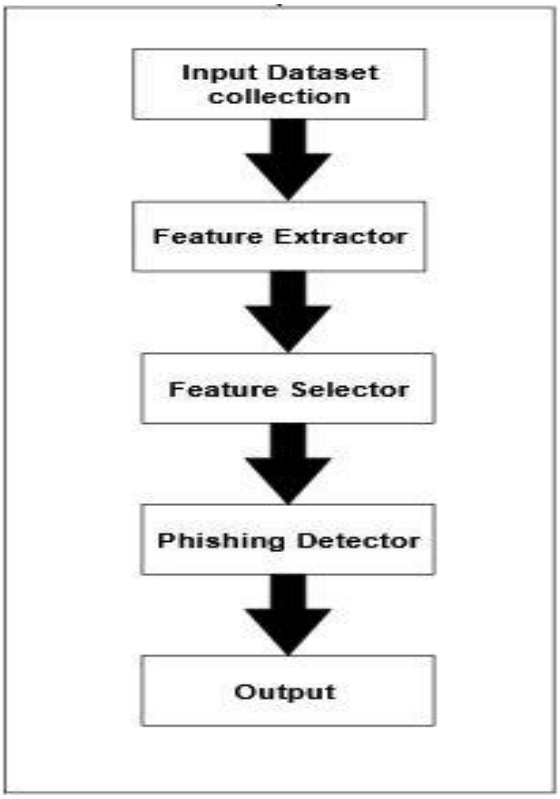

Fig 3: Flow Chart of Proposed Methodology.

Input Dataset Collection:

This module acquires the legitimate and phishing websites datasets collected from the UCI Machine Learning Repository. This dataset has 4898 phishing websites and 6157 legitimate websites from which different website features were drawn out.

Feature Extractor:

To detect fake websites from original ones, a number of attributes (features) may be accumulated from the website. The effectiveness of the extracted attributes are crucial for the completion of fake website detection approach.

Feature Selector:

It gives the procedure of identifying which features are comparatively essential among extracted features. A number of features are crucial compared to others, because a certain number of features have less or no impact. So it is important to select features for our ML model.

Phishing Detector:

The advised technique of classification i.e. DT, is put into the group of features in your ML module. It will use 30 attributes (features) extracted from the data set, which is used to identify whether the website's coordination is fake or legitimate.

Output:

Based on the applied classification methods and on the election of attributes, the ML model provides the output as phishing = “Is Malicious” or legitimate =”Is not Malicious”.

IV. CONCLUSION

The prolonged development of technology in networks has contributed to the unrestricted acceptance of electronic banking, e-commerce, e-health, social media and e-learning in different aspect of our lives. And financial associations continuing to experience huge loss and phishing websites becoming tougher to spot, it is important to build adequate self-identification strategies for detecting them. The DT classifier algorithm was

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:05/May-2022 Impact Factor- 6.752 www.irjmets.com

tested on all chosen features to detect the accuracy of the phishing website detection model. We used a 10-fold verification method to train and assess the model to avoid over-fitting. According to the findings, the selection of adequate attributes has an effect with the accuracy of the task of locating phishing websites. As a result, when using a DT classifier based on the elected features, we obtained a high accuracy of 98.80%.

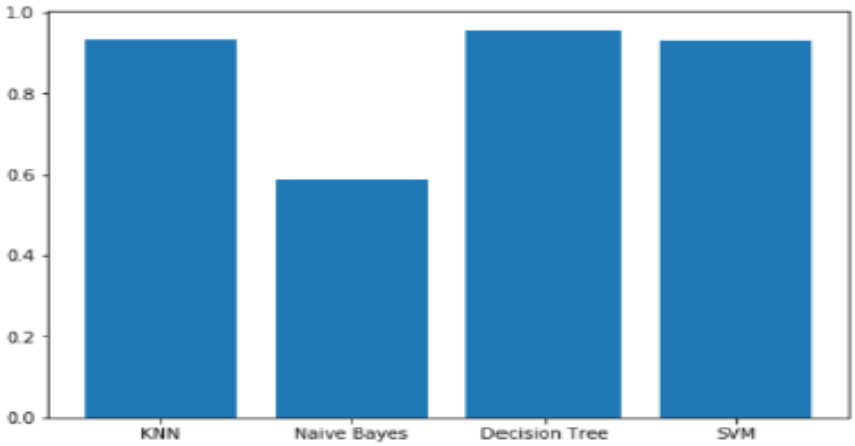

Fig 4: Accuracy comparison Graph

V. REFERENCES

[1] Basit, Abdul, Maham Zafar, Xuan Liu, Abdul Rehman Javed, Zunera Jalil, and Kashif Kifayat. "A comprehensive survey of AI-enabled phishing attacks detection techniques."

[2] Alabdan, Rana. "Phishing attacks survey: Types, vectors, and technical approaches." Future Internet 12, no. 10 (2020): 168.

[3] Stavroulakis, P.; Stamp, M. (Eds.) Handbook of Information and Communication Security; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2010.

[4] Ahmed, Dalia Shihab, Assist Prof Dr Karim Q. Hussein, and Hanan Abed Alwally Abed Allah. "Phishing Websites Detection Model based on Decision Tree Algorithm and Best Feature Selection Method." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 13, no. 1 (2022): 100-107.

[5] Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning, (April). Khonji, M., Iraqi, Y., Member, S., & Jones, A. (2013). Phishing Detection : A Literature Survey, 15(4), 2091– 2121.

[6] Lokesh, G. H., & Boregowda, G. (2020). Phishing website detection based on effective machine learning approach. Journal of Cyber Security Technology, 5(2374-2917), 1–14.

[7] Mao, J., Bian, J., Tian, W., Zhu, S., Wei, T., Li, A., et al. (2018). Detecting phishing websites via aggregation analysis of page layouts. Procedia Computer Science, 129, 224–230.

[8] Li, Y., Yang, Z., Chen, X., Yuan, H., & Liu, W. (2019). A stacking model using url and html features for phishing webpage detection. Future Generation Computer Systems, 94, 27–39.

[9] Gupta, Surbhi, Abhishek Singhal, and Akanksha Kapoor. "A literature survey on social engineering attacks: Phishing attack." In 2016 international conference on computing, communication and automation (ICCCA), pp. 537-540. IEEE, 2016.

[10] Chawla, Minal, and Siddarth Singh Chouhan. "A survey of phishing attack techniques." International Journal of Computer Applications 93, no. 3 (2014).

[11] Mahajan, Rishikesh, and Irfan Siddavatam. "Phishing website detection using machine learning algorithms." International Journal of Computer Applications 181, no. 23 (2018): 45-47.

[12] Kulkarni, Arun D., and Leonard L. Brown III. "Phishing websites detection using machine learning." (2019).

[13] Shahrivari, Vahid, Mohammad Mahdi Darabi, and Mohammad Izadi. "Phishing Detection Using Machine Learning Techniques." arXiv preprint arXiv:2009.11116 (2020).

[14] Machado, Lisa, and Jayant Gadge. "Phishing sites detection based on C4. 5 decision tree algorithm." In 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), pp. 1-5. IEEE, 2017.