IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 3, March 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

||Volume 13, Issue 3, March 2024||

|DOI:10.15680/IJIRSET.2024.1303200|

Phishing Website Detection Based on URL Using Machine Learning

V Samba Siva[1],K Meghana Reddy[2], G Likith Sai[3], P Mahesh[4], V Madhuri[5]

[1]Assistant Professor, Department of CSE, AITS, Tirupati, India

[2, 3,4,5]Student, Department of CSE, Annamacharya Institute of Technology and Sciences, Tirupati, India

ABSTRACT: Phishing attacks have emerged as a significant threat to online security, posing risks to both individuals and organizations. The development of efficient detection mechanisms is crucial to mitigate these threats. Phishing Website Detection based on URL using machine learning presents a comprehensive approach to identify phishing websites based on their URLs. This paper deals with machine learning technology for detection of phishing URLs by extracting and analysing various features of legitimate and phishing URLs. Prominent machine learning algorithms such as Decision Tree (DT) employed to achieve robust results. A decision tree is a tree like model where each node represents a decision based on a particular feature. In phishing detection, a decision tree makes decisions based on features like URL length, presence of certain keywords or the use of HTTPS. Its utilization highlights effectiveness in accurately distinguishing between legitimate and malicious URLs. It helps to improve detection accuracy combining the outputs of multiple decision trees. To evaluate the proposed approach different evaluation parameters were adopted, such as the precision, accuracy, recall, F1-score and specificity to illustrate the effects and efficiency of the models.

KEYWORDS: Decision Tree, Phishing attack, Machine Learning.

I.INTRODUCTION

Internet is not only important for individual users but also for organizations doing business online, these organizations normally offer online trading [1]. Nevertheless, internet-users may be vulnerable to different types of web-threats that may cause financial damages, identity theft, loss of private information, brand reputation damage and loss of customer's confidence in e-commerce and online banking. Therefore, internet suitability for commercial transactions becomes doubtful. Phishing is considered a form of web-threats that is defined as the art of impersonating a website of an honest enterprise aiming to acquire private information such as usernames, password's and social security numbers [2]. Phishing websites are created by dishonest persons to impersonate web pages of genuine websites. These websites have high visual similarities to the legitimate ones in an attempt to defraud the honest internet-users.

Phishing attacks lead to theft of sensitive information such as e-commerce accounts, confidential bank account details and other personally identifiable information of an Internet user [3]. Traditionally, phishing attacks target email users, however, with the unprecedented explosion in popularity of Online Social Media (OSM) like Facebook, Twitter, YouTube and Foursquare, adversaries also use these media to spam and phish[4].There are many authors [5] who have also investigated problem of phishing attacks. Feature selection is also widely accepted and applied by the authors in order to develop efficient anti phishing tool. Feature selection, find out relevant features from large feature space and creates new feature space with minimum number of features and is an integral part in the model development due to high dimensionality data [6].

Spear Phishing [7] target a specific group of people or community belonging to an organization or a company. They send emails which pretend to be sent by a colleague, manager or a higher official of the company requesting sensitive data related to the company.

Phishing detection approaches [8] are broadly classified into two types: user education based techniques and software-based techniques. Software-based detection is further classified into heuristic based, blacklist-based and visual similarity based techniques. User education based approach is to classify phishing and non-phishing email, which is developed by two embedded training designs to teach users. After this training, users can identify phishing emails by themselves. Software-based detection is further classified into following sub-categories as Blackl-based approach, Heuris-based approach, Visual similarity-based approach. In Blackl-based approach [9] the suspicious domain is matched with a predefined phishing domain called blacklist. The negative aspect of this scheme is that it usually does not cover all phishing websites because a freshly launched fraud website takes some time to add to the blacklist record. The heuristic design of suspicious websites matches with the feature set, which are generally found in phishing websites [10]. Zero-day attack can be identified using heuristic approach. Visual similarity-based approaches compare the visual appearance of a suspicious website and its corresponding legitimate site. Visual similarity-based techniques

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

||Volume 13, Issue 3, March 2024||

|DOI:10.15680/IJIRSET.2024.1303200|

[11] use features set like text content, HTML Tags, Cascading Style Sheet (CSS), image processing, etc., to make decision. The proposed machine learning algorithm precisely identifies the attacks created for website phishing.

II.RECENT WORKS

Phishing is a fraudulent attempt to get you to provide personal information, including but not limited to, account information. Whittaker et al. [12], which defines a phishing page as "any web page that, without permission, alleges to act on behalf of a third party with the intention of confusing viewers into performing an action with which the viewer would only trust a true agent of the third party." The design of the Google's phishing classifier used to automatically maintain Google's phishing blacklist. This classifier uses a wide variety of features: from lexical features, such as whether the URL contains an IP address, to URL metadata, such as Google PageRank, as well as features extracted from the page content and hosting information. While this work describes the classifier used to maintain blacklists at the server side, our work focuses on the design of an on-the-fly classifier at the client side. A. Bhavani (2023) proposed a phishing detection approach based on machine learning algorithms such as random forest, XG Boost. These algorithms are aimed to get rid of phishing attack, predicting the threat and it simplifies the feature extraction process and reduces processing overhead of URLs and domain names. It detects the phishing websites but training process for phishing website detection system is slow and time-consuming [13]. Adarsh Mandadi (2022) proposed detection of legitimate and phishing websites based on support vector machine, neural networks, decision tree, random forest. These algorithms aimed to recognize and examine attributes in phishing sites and suggests several features to increase detection performance and effectively classified the phishing and legitimate URLs with an accuracy of 87.0% [14] but hard to find high quality phishing URL's that falls on legitimate and illegitimate websites so that could degrade the performance of phishing website detection system.To overcome these issues, different machine learning models were employed to achieve robust results.

III.MACHINE LEARNING MODEL

Various machine learning algorithms were used in phishing website detection based on url using machine learning to achieve robust results.

3.1Decision Tree Algorithm

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome**.** The decisions or the test are performed on the basis of features of the given dataset. Decision tree algorithm is easy to understand and also easy to implement. Decision tree begins its work by choosing best splitter from the available attributes for classification which is considered as a root of the tree. Algorithm continues to build tree until it finds the leaf node. Decision tree creates training model which is used to predict target value or class in tree representation each internal node of the tree belongs to attribute and each leaf node of the tree belongs to class label. The decision tree algorithm is a non-parametric method which is used for classification and regression. A decision tree is a tree like model where each node represents a decision based on a particular feature. In phishing detection, a decision tree makes decisions based on features like URL length, presence of certain keywords or the use of HTTPS.

3.2 Proposed Methodology Architecture

The proposed approach presented in Figure 2. The canopy centroid selection method is used in clustering as a pre-processing step. Here, the canopy is used as feature selection method as a feature engineering step to select the most effective feature in the detection of phishing URLs. The Ensemble model is based on three different machine learning models such as linear Regression, Support vector Machine, and Decision Tree with Hyper parameter tuning technique to select the best parametric values for the training process of the model. The cross fold validation is used for the effective train test split which improves the training of the model.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

||Volume 13, Issue 3, March 2024||

|DOI:10.15680/IJIRSET.2024.1303200|

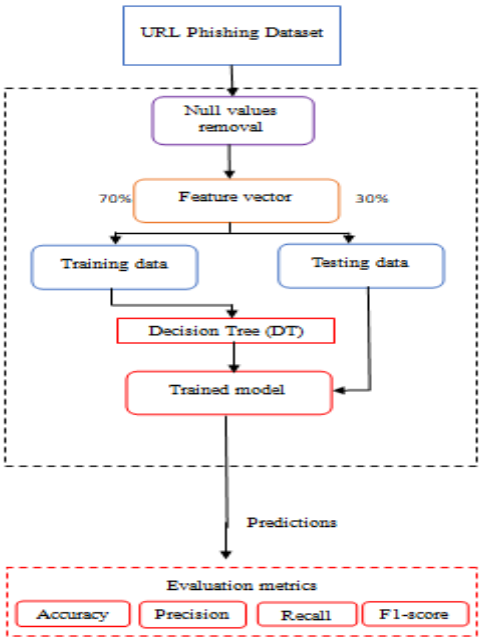

FIGURE1: Proposed approach based on canopy feature selection

3.3 Evaluation Parameter

Machine learning performance must be evaluated using several evaluation parameters. The machine-learning algorithm provides results in the form of predictions. The evaluation parameters measure the number of true and false predictions made by the model in both legitimate and phishing classes.

Parameters such as accuracy, precision, recall, specificity and the F1-score were used. Accuracy measures model performance in terms of the number of accurate predictions made by the model as shown in Equation.

$$\text{Accuracy}= \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

Precision measures the positive rate of the model to the extent to which the model predicts the positive values and indicates the extent to which the model classifies the phishing URLs.

$$\text{Precision}= \frac{TP}{(TP+FP)}$$

A metric used for the analysis of the classification models that answer out of all the likely positive labels

$$\text{Recall}= \frac{TP}{(TP+FN)}$$

The F1 score is the harmonic mean of precision and recall, where the F1 score reaches its best value (perfect precision and recall). The general formula is as follows:

$$\text{F1 Score}=2* \frac{(Precision*Recall)}{(Precision+Recall)}$$

Table1: Result for the performance of the Random Forest model

Max Depth	Accuracy	F1 Score	Recall	Precision
0	94.9	94.46	95.41	95.44
5	92.07	89.67	96.97	93.18
10	94.3	94.59	95.23	94.92
20	95.38	95.7	96.06	95.88
30	95.41	95.8	96	95.91

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

||Volume 13, Issue 3, March 2024||

|DOI:10.15680/IJIRSET.2024.1303200|

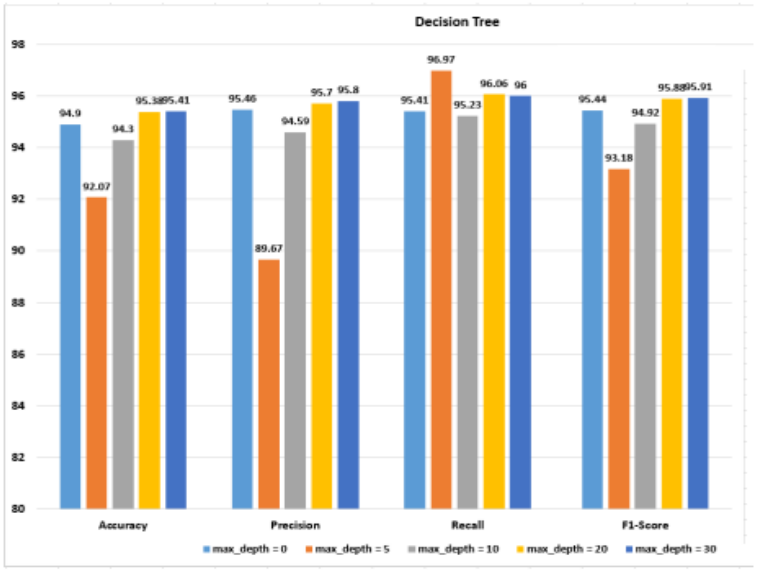

FIGURE 2: Experimental results of the Machine Learning Models

IV.CONCLUSIONS

The Internet consumes almost the whole world in the upcoming age, but it is still growing rapidly. With the growth of the Internet, cybercrimes are also increasing daily using suspicious and malicious URLs, which have a significant impact on the quality of services provided by the Internet and industrial companies. Currently, privacy and confidentiality are essential issues on the internet. To breach the security phases and interrupt strong networks, attackers use phishing emails or URLs that are very easy and effective for intrusion into private or confidential networks. Phishing URLs simply act as legitimate URLs. A machine-learning-based phishing system is proposed in this study. The proposed approach is evaluated in this study by experimenting with a separate evaluation parameters and then further evaluation of the study was carried out. The proposed approach successfully achieves its aim with effective efficiency.

REFERENCES

[1] J. Liu and Y. Ye, "Introduction to E-Commerce Agents: Marketplace Solutions, Security Issues, and Supply and Demand," in E-Commerce Agents, Marketplace Solutions, Security Issues, and Supply and Demand, London, UK., 2001.

[2] R. M. Mohammad, F. Thabtah, and L. McCluskey, ''Predicting phishing websites based on self-structuring neural network,'' Neural Comput. Appl., vol. 25, no. 2, pp. 443–458, Aug. 2014.

[3] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, ''PhishAri: Automatic realtime phishing detection on Twitter,'' in Proc. eCrime Res. Summit, Oct. 2012, pp. 1–12.

[4] M. Zouina and B. Outtaj, ''A novel lightweight URL phishing detection system using SVM and similarity index,'' Hum.-Centric Comput. Inf. Sci., vol. 7, no. 1, p. 17, Jun. 2017.

[5] H. S. Hota, A. K. Shrivas, and R. Hota, ''An ensemble model for detecting phishing attack with proposed remove-replace feature selection technique,'' Proc. Comput. Sci., vol. 132, pp. 900–907, Jan. 2018.

[6] G. Sonowal and K. S. Kuppusamy, ''PhiDMA—A phishing detection model with multi-filter approach,'' J. King Saud Univ., Comput. Inf. Sci., vol. 32, no. 1, pp. 99–112, Jan. 2020.

[7] Hong J (2012) The state of phishing attacks. Commun ACM 55(1):74–81.

[8] A. K. Jain and B. Gupta, ''PHISH-SAFE: URL features-based phishing detection system using machine learning,'' in Cyber Security. Switzerland: Springer, 2018, pp. 467–474.

[9] Sheng S, Wardman B, Warner G, Cranor LF, Hong J, Zhang C (2009) An empirical analysis of phishing blacklists. In: CEAS 2009.

[10] Almomani A, Gupta BB (2013) Phishing dynamic evolving neural fuzzy framework for online detection zero-day phishing E-mail. IJST 6(1):122–126.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

||Volume 13, Issue 3, March 2024||

|DOI:10.15680/IJIRSET.2024.1303200|

[11] Zhang Y, Hong JI, Cranor LF (2007) Cantina: a content-based approach to detecting phishing web sites. In: Proceedings on WWW, ACM, New York, pp 639–648

[12] Whittaker, B. Ryner, and M. Nazif. Large-Scale Automatic Classification of Phishing Pages. In Proc. of NDSS '10.

[13] A. Bhavani, R. Sai Lakshmi, P. Harshavardhini, P. Vijay Prakash, N. Vamsi Behara and V. Ajay Kumar,"Detection of Legitimate and Phishing Websites using Machine Learning", IEEE Xplore ,ICSCSS ,2023,366-371.

[14] Adarsh Mandadi, Saikiran Boppana, Vishnu Ravella, and Dr. R. Kavitha, "Phishing Website Detection Using Machine Learning" ,IEEE Xplore,2022.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details