DETECTING PHISHING WEBSITE USING MACHINE LEARNING

Pakala Shireesha[1],Vuturi SaiRani[2] ,Thirunahari Venkata Narashimha Charyulu[3],Pothapa Anirudh[4], Dr.D.S Sameena Begum[5]

[1,2,3,4]*B.Tech. Student, Department of Computer Science and Engineering,*

shireeshapakala22@gmail.com,vuturisairani@gmail.com, tnarasimha444@gmail.com,

pothapaanirudh8@gmail.com, shashasameena@gmail.com

[5]*Associate Professor, Department of Computer Science and Engineering,*

Nalla Malla Reddy Engineering College, Hyderabad, India

***Abstract*-**Phishing attacks continue to evolve and pose significant security risks to online users. The proposed approach incorporates a range of features, including URL, domain, page content, user behavior, and network traffic, to build a classifier model. The model is trained on a large dataset of known phishing and legitimate websites and evaluated on a separate test set to assess its performance. The study also examines the impact of different feature combinations and model selection on the performance of the proposed approach. The experimental results demonstrate that the proposed approach achieves high accuracy, precision, and recall in detecting phishing websites. The research also compares the proposed approach with existing state-of-the-art techniques and highlights its superiority in terms of detection rates. The findings suggest that the proposed approach has practical implications for improving online security and protecting users from phishing attacks.

Keywords-Phishing,Legitimate, MachineLearning, Detection,URL,Accuracy,Precision, Recall.

1. INTRODUCTION

Nowadays phishing attacks have become a major threat to online security, as they can lead to the loss of sensitive information and financial loss. Traditional approaches to detecting phishing websites rely on manual inspection and rule-based systems, which are limited in their ability to keep up with the constantly evolving nature of these attacks.

Machine learning has emerged as a promising approach to detecting phishing websites, as it can analyze large amounts of data and learn patterns that can be used to identify malicious behavior. In recent years, researchers have developed various machine learning models to detect phishing websites,includinggradient boosting,neural networks, decision trees, and support vector machines.However, despite the promising results achieved by these models, there are still many challenges that need to be addressed. One of the main challenges is the lack of a standardized dataset for evaluating the performance of different models. Additionally, the dynamic and complex nature of

phishing attacks requires the development of more sophisticated models that can adapt to new attack vectors.

This research paper aims to address these challenges by proposing new machine learning models and algorithms for detecting phishing websites. We will also develop a new dataset for evaluating the performance of different models and compare the effectiveness of our proposed models with existing approaches. By advancing the state-of-the-art in machine learning-based phishing detection, we hope to provide better protection to users and organizations against these increasingly sophisticated attacks.

2.PROPOSED SYSTEM

The proposed phishing detection system utilizes machine learning models and deep neural networks. The system comprises two major parts, which are the machine learning models and a web application. These models consist of Decision Tree, Support Vector Machine, XG Booster, Multilayer Perceptions, Auto Encoder Neural Network, and Random Forest. These models are selected after differentcomparison-based performances of multiple machine learning algorithms. Each of these models is trained and tested on a website content-based feature, extracted from both phishing and legitimate dataset.

The main purpose of the project is to detect the fake or phishing websites who are trying to get access to the sensitive data or by creating the fake websites and trying to get access of the user personal credentials. We are using machine learning algorithms to safeguard the sensitive data and to detect the phishing websites who are trying to gain access on sensitive data.

3.LITERATURE REVIEW

Aiding Phishing Website Detection With a Feature-Free Tool In this paper, we propose a feature-free method for detecting phishing websites using the Normalized Compression Distance (NCD), a parameter-free similarity measure which computes the similarity of two websites by compressing them, thus eliminating the need to perform any feature extraction. It also removes any dependence on a specifific set of website features. This method examines the HTML of webpages and computes their similarity with known phishing websites, in order to classify them. We use the Furthest Point First algorithm to perform phishing prototype extractions, in order to select instances that are representative of a cluster of phishing webpages. We also introduce the use of an incremental learning algorithm as a framework for continuous and adaptive detection without extracting new features when concept drift occurs. On a large dataset, our proposed method signifificantly outperforms previous methods in detecting phishing websites, with an AUC score of 98.68%, a high true positive rate (TPR) of around 90%, while maintaining a low false positive rate (FPR) of 0.58%. Our approach uses prototypes, eliminating the need to retain long term data in the future, and is feasible to deploy in real systems with a processing time of roughly 0.3 seconds.When deciding which approach to take when performing clustering and classifification, we take into consideration

the main goals or characteristics which are important to have for the detection to be accurate and sustainable. Firstly, we aim for the system to be feature-free, meaning that it can learn directly from the data without the need to perform manual feature extraction.For the reasons mentioned above, a prototype-based clustering approach has been selected, where clusters are represented by actual samples in the dataset instead of centroids.

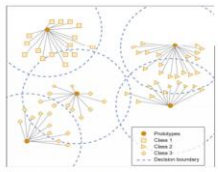
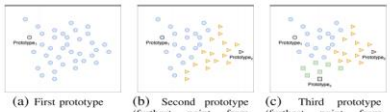


Fig. 1. Classification using Prototypes.



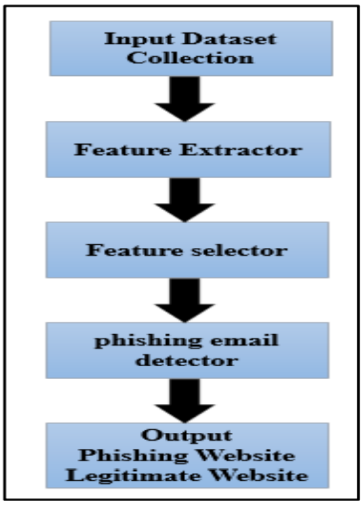
(a) First prototype (b) Second prototype (furthest point from $prototype_1$) (c) Third prototype (furthest point from $prototype_2$)

Multilayer Stacked Ensemble Learning Model to Detect Phishing Websites

Now a days, the ease in communication over internet has brought revolutionary changes. This positive transformation drastically increases the number of internet users. At the same time adversaries make use of this opportunity to steal sensitive credentials of an internet user by creating phishing websites or sending fake emails to naive online users.There exists many antiphishing techniques that use whitelists,
blacklists, heuristics, visual descriptors, third party services and machine learning algorithms for the detection of phishing sites. Also, the techniques use the source code or URL for the feature extraction to classify the phishing sites. Out of all the antiphishing techniques, machine learning based techniques played a vital role in detection of phishing sites. Moreover, the existing literature demonstrates that machine learning based solutions could achieve performance of atleast 99%. Hence, we also attempted to use machine learning approach for the phishing detection.There exists many antiphishing techniques that use whitelists, blacklists, heuristics, visual descriptors, third party services and machine learning algorithms for the detection of phishing sites. Also, the techniques use the source code or URL for the feature extraction to classify the phishing sites.

Measures	EW1[36]	EW2[37]	EW3[40]	EW4[42]	EW5[43]	EW6[44]	MLSELM
Accuracy	97.16	97.33	93.41	95.0	97.4	95.4	**97.76**
Precision	96.86	97.0	92.99	95.0	96.0	93.5	**97.34**
Recall	96.83	98.3	91.98	95.0	98.1	95.9	**98.07**
F-Score	–	97.6	92.48	95.0	97.0	97.0	**97.70**

4.METHODOLOGY



The phishing website detection architecture is made up of five components that serve as a task assembly. The following are the functions of each module:

Input Dataset Collection: This module is in charge of acquiring the phishing and legitimate website datasets obtained from the UCI Machine Learning Repository ("UCI Machine Learning Repository: Phishing Websites Data Set.," 2016). This dataset contains 4898 phishing websites and 6157 legitimatewebsites from which many website features were extracted. Feature Extractor: To identify phishing websites from legitimate ones, many features may be gleaned from a website. The quality of the retrieved features is critical to the performance of phishing website detection techniques. The UCI Machine Learning Repository ("UCI Machine Learning Repository: Phishing Websites Data Set.," 2016) has a dataset of phishing websites that contains 30 essential features of websites that have been shown in (Mccluskey & Mccluskey, 2012) to be effective and influential in predicting phishing and legitimate websites. The following table covers the important features that may aid in the successful prediction of phishing websites.Table (3) presents the main features that can contribute in the effective prediction of phishing sites.

Feature Selector: It is the process of determining which features are meaningful from the retrieved features. Certain features are more significant than others, since some of the other features have little or no influence. As a result, attribute selection is crucial in our machine learning architecture. The feature set is selected using two feature selection methods: Wrapper Features Selection (WFS) and Correlation-based Feature Selection (CFS).

Phishing Email Detector: The suggested classification method, DT, is applied to the collection of features in this module. DT method will use 30 features derived from the data set to determine if a website's synchronization is a phish or not. In this paper, we use 10-fold cross-validation to train and assess our classifier. For the 10-fold cross-validation process, divide the data set into 10 parts; 9 of the 10 parts are used to train the classifier, and the information gained from the training phase is used to validate (or test) the 10th part; this is repeated 10 times, with each part serving as both training and test data at the end of the training and testing phase. The accuracy of each run is calculated. As a result, the ultimate accuracy of learning from this dataset is the average of the n accuracies for all runs. The use of cross-validation ensures that the training and test data are both varied. In machine learning, the cross-validation method is well-known for providing a very accurate estimate of a classifier's generalization error.

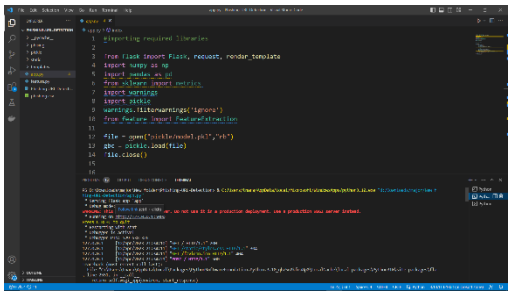
Output: Based on the selection of features and classification techniques employed, this module generates results. The outcome is generated using phishing website detection accuracy, which is used to identify the unclassified website as either legitimate = 1 or phishing = -1.

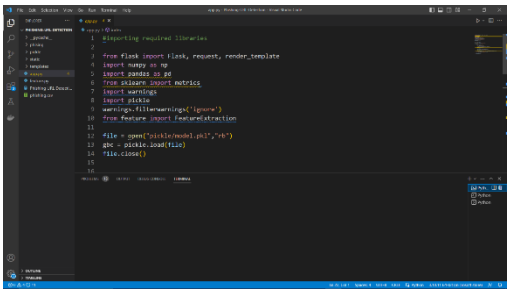
5. RESULTS

The results for the proposed system is obtained from the developed Machine Learning Model in the form of Binary Classification output.It gives an output as whether the website is a phishing or legitimate.

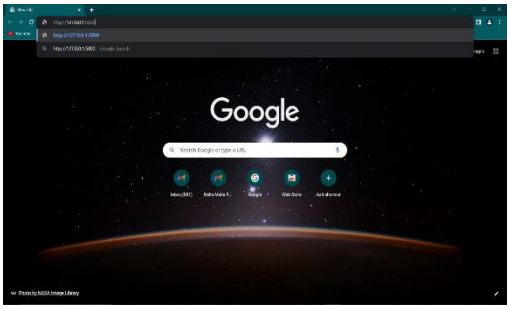
5.1 Visual Studio Code

5.2 Running the code and copying the generated URL

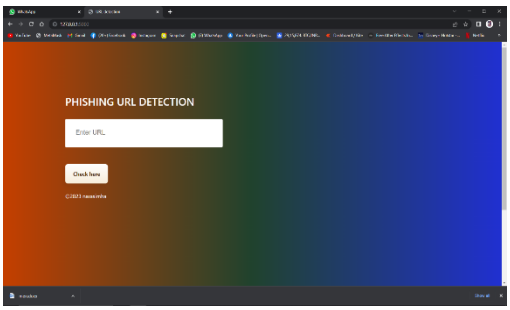




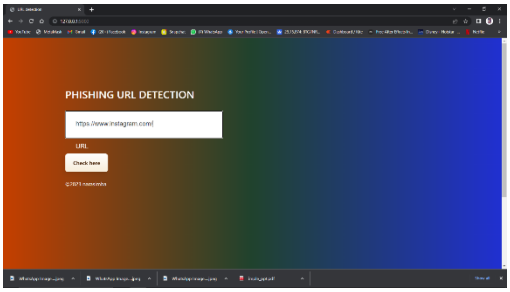
5.3 Opening the URL detection



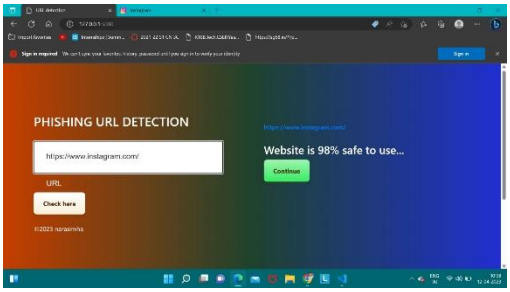
5.4 HOME PAGE



5.5 Checking the URL site



5.6 Result of the URL site



6. DISCUSSION

Phishing is a common tactic used by cybercriminals to steal sensitive information from unsuspecting users. Detecting phishing websites is crucial in preventing these attacks, and machine learning can be a valuable tool in this effort.

Machine learning algorithms can be trained to identify patterns and characteristics commonly found in phishing websites, such as:

1. URL and domain similarity: Phishing websites often use URLs and domain names that are similar to legitimate websites to trick users into thinking they are on a trusted site.
2. Content and language: Phishing websites often contain language and content that is designed to elicit an emotional response, such as fear or urgency, in order to persuade the user to take action.
3. Website design: Phishing websites often have poor design and layout, with low-quality images and inconsistent formatting.

By analyzing these and other factors, machine learning algorithms can identify potential phishing websites and flag them for further investigation or block access to them altogether.

learning with other measures, such However, it is important to note that machine learning algorithms are not foolproof and may sometimes produce false positives or fail to detect sophisticated phishing attacks. Therefore, it is important to supplement machine as user education and awareness, to ensure comprehensive protection against phishing attacks.

7. Conclusion

The system developed detects if a URL link is phishing or legitimate by using machine learning models and deep neural network algorithms. The feature extractionand the models used on the dataset helped to uniquely identify phishing URLs and also the performance accuracy of the models used. It is also surprisingly accurate at detecting the genuineness of a URL link this developed project is to explore various machine learning models, perform exploratory data analysis on phishing records, and understand their capabilities a This phishing records have some characteristics such as 'HTTPS', 'Anchor URL', 'Website Traffic' are more important for classifying whether a URL is classified as a phishing URL. Overall, a well-designed and executed phishing URL detection project using machine learning can provide insight into the latest techniques and technologies to combat cyberthreats. It will also be a valuable learning experience for those interested in

cybersecurity and machine learning, contributing to the development of practical and effective solutions to improve online security.

8.References

[1] W. Zhuang, Q. Jiang and T. Xiong, "An Intelligent Anti-phishing Strategy Model for Phishing Website Detection," *2012 32nd International Conference on Distributed Computing Systems Workshops*, Macau, China, 2012, pp. 51-56, doi: 10.1109/ICDCSW.2012.66.

[2] Tupsamudre, H., Singh, A. K., & Lodha, S. (2019). Everything is in the name–a url based approach for phishing detection. In *Cyber Security Cryptography and Machine Learning: Third International Symposium, CSCML 2019, Beer-Sheva, Israel, June 27–28, 2019, Proceedings 3* (pp. 231-248). Springer International Publishing.

[3] Jain, A. K., & Gupta, B. B. (2018). PHISH-SAFE: URL features-based phishing detection system using machine learning. In *Cyber Security: Proceedings of CSI 2015* (pp. 467-474). Springer Singapore.

[4] Chawla, A. (2022). Phishing website analysis and detection using Machine Learning. *International Journal of Intelligent Systems and Applications in Engineering*, *10*(1), 10-16.

[5] Rashid, J., Mahmood, T., Nisar, M. W., & Nazir, T. (2020, November). Phishing detection using machine learning technique. In *2020 first international conference of smart systems and emerging technologies (SMARTTECH)* (pp. 43-46). IEEE.

[6] Yerima, S. Y., & Alzaylaee, M. K. (2020, March). High accuracy phishing detection based on convolutional neural networks. In *2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)* (pp. 1-6). IEEE.

[7] Liu J, Ye Y (2001) Introduction to E-business operators: commercial center arrangements,

[8] APWG, Aaron G, Manning R (2013)) APWG phishing reports. APWG, 1 February 2013.

[9] [Online].Accessble: http://www. antiphishing.org/assets/apwg-reports/. Gotten to 8Feb2013

[10] Liu J, Ye Y (2001) Introduction to E-business operators: commercial center arrangements, security issues, and market interest. In: E-business specialists, commercial center arrangements, security issues, and market interest, London, UK.