Grenze International Journal of Engineering and Technology, June Issue

Detection of Phishing Website using Machine Learning

T. Keerthana[1], P.A. Mathina[2], A. Ramathilagam[3] and K. Valarmathi[4]
[1]Electronics and Communication Engineering, P.S.R. Engineering College, Sivakasi, India,
Email: tkeerthana40@gmail.com
[2]Assistant professor, Electronics and Communication Engineering, P.S.R. Engineering College, Sivakasi, India
Email: mathina@psr.edu.in
[3-4]Professor, Electronics and Communication Engineering, P.S.R. Engineering College, Sivakasi, India
Email: {valarmathi, ,ramathilagam}@psr.edu.in

***Abstract*— Detecting phishing websites is crucial in the ongoing battle against cyber threats, as malicious actors continually evolve their tactics to exploit unsuspecting individuals and organizations. We address the escalating risks associated with deceptive online practices, specifically focusing on the insidious form of cyber-attack known as phishing. The objective is to develop robust detection mechanisms capable of identifying and thwarting fraudulent websites before they compromise user and organizational security. Utilizing advanced technologies such as machine learning, pattern recognition, and real-time analysis, our detection system scrutinizes websites for telltale signs of phishing. These signs include suspicious URLs, misleading content, and attempts to emulate authentic login pages. Beyond individual protection, effective phishing detection plays a crucial role in safeguarding financial assets, preventing identity theft, and mitigating the potential for widespread data breaches. It serves as a critical line of defense, preserving the integrity of personal information and bolstering the resilience of entire digital ecosystems. As cyber threats continue to evolve, the sophistication of phishing detection systems must also advance. Proactive monitoring, rapid response to emerging threats, and ongoing education and awareness efforts are essential components of a comprehensive strategy to combat the pervasive and ever-evolving menace of phishing websites. In a world where online interactions are integral to daily life, the importance of robust phishing detection cannot be overstated, as it acts as a vital bulwark in the protection of individuals and organizations against cyber threats.**

***Index Terms*— Cyber threats, Phishing, Malicious, URL**

I. INTRODUCTION

In the dynamic landscape of cybersecurity, our project assumes a pivotal role by addressing the escalating threat posed by phishing websites. Phishing, a malicious stratagem, continually evolves, employing sophisticated techniques to replicate authentic platforms and deceive unsuspecting users. Our research focuses on the development of robust detection mechanisms to counteract this insidious form of cyber-attack[2]. By harnessing advanced technologies such as machine learning, pattern recognition, and real-time analysis, our system scrutinizes websites for telltale signs of phishing activities ranging from suspicious[18]URLs to attempts at emulating genuine login pages. Acting as a critical line of defense, our project not only preserves the integrity of personal information but also bolsters the resilience of entire digital ecosystems.

As cyber threats continue to evolve, the sophistication of our phishing detection systems is paramount. Proactive monitoring, rapid responses to emerging threats, and ongoing education and awareness efforts constitute integral

Grenze ID: 01.GIJET.10.2.404

components of our comprehensive strategy. In a world where online interactions are integral to daily life, the importance of robust phishing detection cannot be overstated. Our project serves as a vital bulwark against the pervasive and ever-evolving menace of phishing websites, contributing significantly to the protection of individuals and organizations in the digital realm.

II. Literature Survey

To improve the accuracy of phishing website identification, Kalabarige et al. [1] propose a unique Boosting-Based Hybrid Feature Selection and Multi-Layer Stacked Ensemble Learning model. Castono.F et al use a multi-layer stacked model to capitalize on the synergy between ensemble learning and boosting techniques [2]. A detailed analysis of clever detection systems for phishing attempts using HTML URLs and a state-of-the-art method for identifying phishing websites is stated by Asiri et al. [3]. The Authors examined adversarial techniques against a system intended to identify harmful URLs for advertisements.

Insights on the weaknesses of malicious advertisement URL detection systems are undertaken by Zieni et al. [4], furthering the understanding of cybersecurity difficulties. A low-cost, efficient XG Boost technique designed to identify dangerous URLs in unbalanced datasets. emphasizes their commitment to advancing security measures against malicious URLs. The work, conducted at China Mobile Research Institute, aligns with the ever-evolving landscape of cybersecurity challenges. [5-6]. Ariyadasa [7] et al examined the categorization of attack kinds and carried out an exhaustive examination of attack techniques, particularly phishing mail attack groups profiling, underscoring their commitment to advancing the understanding of phishing mail attacks

Hereby integrating Long-Term Recurrent Convolutional and Graph Convolutional Networks, the authors leverage a multifaceted approach using both URL and HTML information. A noteworthy contribution to the rapidly developing field of cybersecurity, Shen et al.'s proposed Effectual cost-sensitive XGBoost [8] technique for the detection of malevolent URL detection in an imbalanced data set may have implications for the enhancement of phishing site detection mechanisms.

Lee et al [9] classified the many forms of phishing emails into three categories: spoofing emails, using email body vulnerabilities, and spoofing attached files. They also provided in-depth evaluations of their attack techniques. Liu, X et al [10] proposed a similar learning technique called SPWalk uses a network-based approach for phishing detection.

III. Methodology

The approach we employed in our research to identify phishing websites combines cutting-edge technologies to produce a reliable and efficient system. The key components include machine learning, pattern recognition, and real-time analysis, strategically orchestrated to scrutinize websites for signs of phishing activities. Machine learning algorithms[16] play a central role in training the system to recognize patterns indicative of phishing.

The model is trained on diverse datasets comprising both legitimate[17] and phishing websites, enabling it to learn and discern subtle distinctions between authentic and deceptive platforms. This training facilitates the algorithm's ability to make accurate predictions when exposed to new, unseen data. Improve accuracy, precision, and recall of learning groups using features selected from each feature.

A. Verification of IP address

If the IP address is found in the URL, this property is set to 1; otherwise it is set to 0. Most websites do not use IP addresses as URLs to store web pages. The use of an IP address in a URL indicates that an attacker is trying to steal confidential information.

B. Authentication of @ symbol in URL

If the IP address is found in the URL, this property is set to 1; otherwise it is set to 0. Most websites do not use IP addresses as URLs to store web pages. Use of IP addresses in URLs indicates an attacker is trying to steal confidential information.

C. Check dots in Hostname

Phishing URLs have multiple dots in the URL. For example, the URL http://shop.fun.amazon.phishing.com. Phishing.com is a real domain[14] name, and the use of the word "amazon" is intended to trick users into clicking on it. The average number of elements in a good URL is 3. If the number of elements in the URL exceeds 3, the property is set to 1, otherwise - 1.

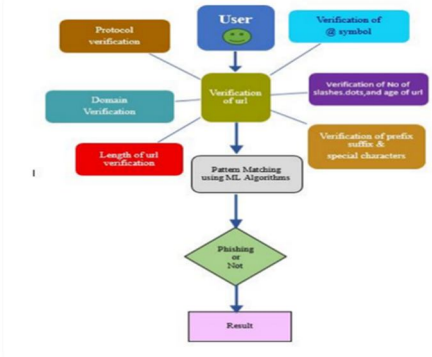

Figure 1. Proposed Frame work for Detection

D. Verification of URL Length

Phishing URLs frequently have longer names to resemble legitimate [11] websites. Analyzing the length of URLs might be useful in identifying too long or suspicious domain names.

E. Verification of Number of Slashes

It becomes apparent that benign URLs have five slashes; if this number is higher than five, the feature is set to one, otherwise to zero.

F. Verification Number of Hyphens in Host Address

Phishing websites may try to imitate genuine domains by using hyphens. Phishing [12] signs may be found by examining the host address's hyphen count.

IV. Machine Learning Models

A subfield of artificial intelligence known as "machine learning" involves developing models using algorithms for machine learning. This model is accomplished on some data and then used to process more data to provide predictions. Models related to this classification problem include: A subfield of artificial intelligence known as "machine learning" deals with building models using algorithms that are trained on certain data and then utilized to process further data in order to provide predictions.

A. Decision Tree

It is a tracking machine learning algorithm built on a tree of data structures and the decision is made at each point of the tree. Decisions are then made for which child node to go to next, and when a traversal terminates at a leaf node, the class is predicted. The features in the input data are used to select the decisions, and a feature is the only one chosen to split if the system's entropy decreases. Highest depths of 18 provided the best accuracy.

$$E(S) = \log 2pi \sum - pi \qquad (1)$$

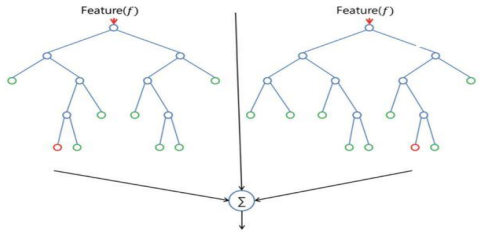

Figure 2. Decision tree

Information Gain can be defined as the difference between the entropy before and after splitting based on a feature. A feature should be used to maximize the amount of information obtained.

$$E(W) - E(W/Z) = IG\ (Z,W) \qquad (2)$$

Z is the starting entropy, and E(W/Z) is the entropy following some information being obtained.

B. Random Forest

This machine learning algorithm is supervised and uses an ensemble approach to generate numerous Decision Trees during training. The output of the algorithm is the average of the individual trees' predictions. Maximum depth of 16 produced the best accuracy.

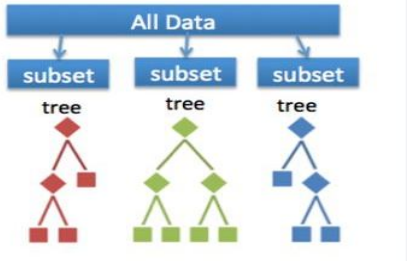

Figure 3: Random Forest

C. Support Vector Machine

The goal of this machine learning approach is to define an n-dimensional plane[13] that divides two types of data points. The program takes n features as input. g(x) = wTx + b is the equation of the plane that divides the data; w and b are weights that the model has learned.

$$Y = \begin{cases} 1, & if \ \ g(x) \geq 1x < 0 \\ 0, & if \ \ g(x) \geq -1 \end{cases} \qquad (3)$$

Points above g(x) = 1 plane are categorized as class 1, and points below g(x) = −1 are classified as class 0. This is because the plane serves as the division between them.

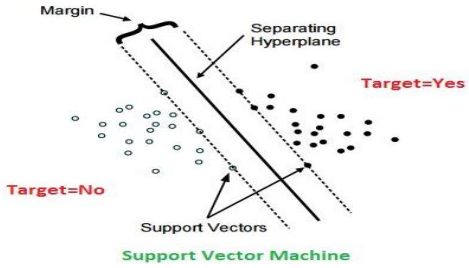

Figure 4.Support Vector Machine

D. Logistic Regression

It is a statistical model-based supervised machine learning technique that outputs the likelihood of a given class. It calculates the probability using its logistic function. It calculates the likelihood as follows:

$$P(X) = e^{(a+bX)} + 1 \qquad (4)$$

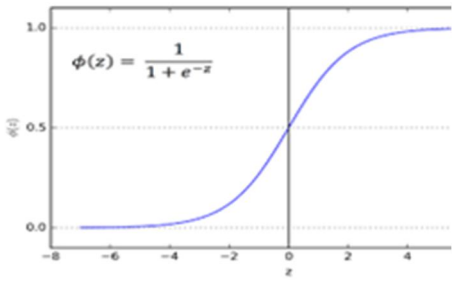

Figure 5 Logistic Regression

The input variable is denoted by X. The weights of the logistic regression model are a and b, while e is the base of the natural logarithm.

E. Voting Classifier

Polling is a supervised machine learning technique in which a model is trained using multiple sets of models, each of which provides a prediction. The final output of the voting classifier is the prediction with the highest vote total. It works by reducing false positives, called false negatives and false positives. It can also improve the overall performance of the model by reducing the likelihood of inaccurate predictions when only a single model is used to obtain the final result. The combination of Decision Tree and Random Forest achieved[20] the highest accuracy. The voting classifier primarily uses two kinds of voting, which are

Hard Voting

The final output of this classifier is generated by voting for the class which was correctly predicted by the models the most times or with the highest probability; in other words, it is basically the mode of the predictions made by the models

Soft Voting

The final probability of the model is predicted by this classifier using the mean/ave Testing Set (80%): Some of this data is used to evaluate machine learning models. In this stage, the model learns the patterns, relationships, and features[15] that exist in the data..

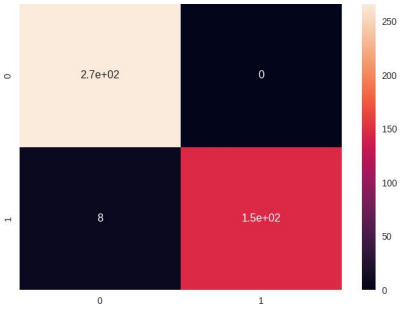

Figure 6: voting classifier confusion matrix

IV. Results And Discussion

A comparison and description of the various machine learning models trained are shown in Table 1. To use specific characteristics chosen from all available features to increase the ensemble learning's recall, accuracy, and precision.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$\text{Precision(P)} = \frac{TP}{TP+FP} \quad (6)$$

$$\text{Recall(R)} = \frac{TP}{TP+FN} \quad (7)$$

A quality of binary (two-class) classifications is assessed using the Matthews Correlation Coefficient (MCC). It is especially important with unequal information, including positive, negative, negative, and negative.

$$MCC = \sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)(TN + FN)} \quad (8)$$

where:
TP - Number of true positives,
TN - Number of true negatives,
FP - Number of false positives, FN- Number of false negatives
Both the Voting Classifier and the Decision Tree indicate great accuracy; the Voting Classifier achieves an excellent accuracy of 99%. While Random Forest's precision is just 87.6%, its recall is a high 100%, signifying that it recognizes all positive instances. SVM has good accuracy, precision, and recall rates and performs well on all measures. whereas not as accurate as other models, Logistic Regression maintains a respectable level of precision and recall. To obtain a balanced performance with high accuracy, precision, and recall, the Voting Classifier integrates multiple models.
An equivalent test of classification efficiency is the Matthews correlation coefficient, which adjusts for true positives, true negatives, false positives, and false negatives. With a maximum MCC score of 96.40, the Voting

Classifier demonstrated exceptional overall performance concerning both sensitivity and specificity. With an MCC of 95.9, the Decision Tree Classifier similarly attained good performance.
With MCC scores of 90.45 and 91.80, respectively, Random Forest and SVM exhibited strong performance. Despite considering its slightly lower MCC of 80.3 than the other models, Logistic Regression is still rather good. The classifier you choose will rely on the specific goals and specifications. Overall performance is normally better when the MCC is higher, but it's important to take into consideration the particular situation and classification problem.
Using the confusion matrix, we can evaluate the classification performance of the iris dataset. Points outside the diagonal are points mislabeled by the classifier; The diagonal shows the number of points where the prediction [19] and the reality are equal. The diagonal of the confusion matrix represents the number of correct guesses; higher values are better. MCC ranges from -1 to 1, where 1 is the decision tree classifier's highest level of precision. the heat map calculation's output performance.

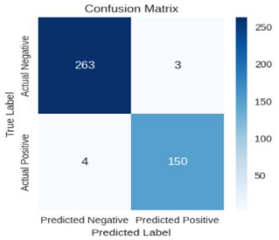

Figure 7 .Decision Tree

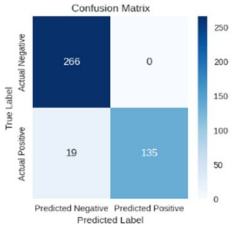

Figure 8. Random Forest

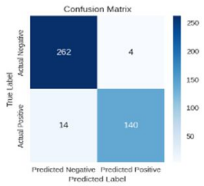

Figure 9. SVM

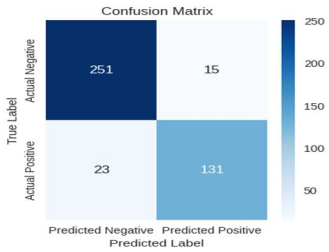

Figure 10. Logistic regression

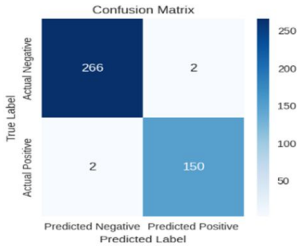

Figure 11. voting classifier

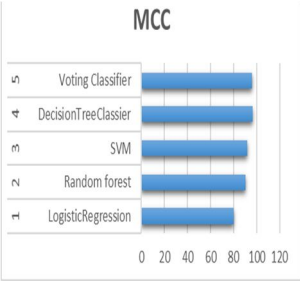

Figure 12. MCC Representation

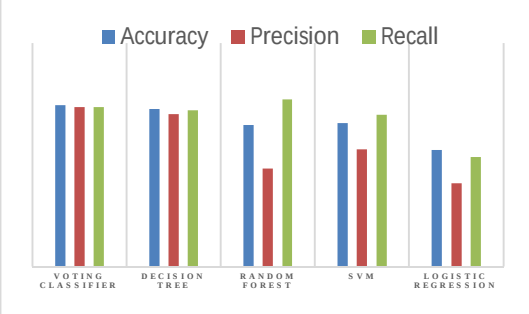

Figure 13. Performance analysis

TABLE I MATTHEWS CORRELATION COEFFICIENT

S.No	Classifier	Sample
1	Logistic Regression	80.3
2	Random forest	90.45
3	SVM	91.8
4	Decision Tree Classifier	96.4
5	Voting Classifier	95.9

TABLE II PERFORMANCE ANALYSIS

Classifier	Accuracy	Precision	Recall
Voting Classifier	99	98.6	98.6
Decision Tree	98.3	97.4	98.2
Random Forest	95.4	87.6	100
SVM	95.7	91.2	97.2
Logistic Regression	90.9	85.5	89.7

V. CONCLUSIONS

In conclusion, the imperative role of robust phishing detection in safeguarding against evolving cyber threats is evident. The utilization of advanced technologies, such as machine learning and real-time analysis, underscores the commitment to proactive identification and thwarting of fraudulent websites. Beyond individual protection, the impact extends to securing financial assets, preventing identity theft, and averting widespread data breaches, reinforcing its critical position as a defense line in the digital landscape. As cyber threats evolve, continuous advancements in the sophistication of phishing detection systems become paramount for efficacy. Looking to the future, Additionally, Continuous research and development in the field of network security, especially to improve the adaptation of detection systems, are essential to stay ahead of cyber adversaries. Collaborative initiatives between industry, academia, and government entities can further strengthen the resilience of digital ecosystems against the persistent and ever-evolving menace of phishing websites.

REFERENCES

[1] G Lakshmana Rao Kalabarige, Routhu Srinivasa Rao Alwyn R. Pais, And Lubna, Abdelkareim Gabralla ,"A Boosting-Based Hybrid Feature Selection and Multi-Layer Stacked Ensemble Learning Model toDetect Phishing Websites",IEEE Access on June 2023 Digital Object Identifier 10.1109/ACCESS.2023.3293649

[2] Castano. F., Fidalgo, E., Aláiz-Rodríguez, R., & Alegre, E. (2023). "PHIKITA: Phishing kit attacks dataset for phishing websites identification." IEEE Access, 11, 40779–40789.

[3] Asiri, S., Xiao, Y., Alzahrani, S., Li, S., & Li, T. (2023). A survey of intelligent detection designs of HTML URL phishing attacks. IEEE Access, 11, 6421–6443.

[4] Zieni, R., Massari, L., & Calzarossa, M. C. (2023)". Phishing or not phishing? A survey on the detection of phishing websites." IEEE Access, 11, 18499–18519.

[5] Altamimi, A. B., Ahmed, M. Z., Khan, W., Alsaffar, M., Ahmad, A., Khan, Z., & Alreshidi, A. (2023). PhishCatcher: Client-Side Defense against web spoofing Attacks using Machine Learning. IEEE Access, 11, 61249–61263

[6] Nowroozi, E., Abhishek, Mohammadi, M., & Conti, M. (2023). An Adversarial attack analysis on Malicious Advertisement URL Detection Framework. IEEE Transactions on Network and Service Management, 20(2), 1332–1344F..

[7] Ariyadasa, S., Fernando, S., & Fernando, S. (2022). Combining Long-Term Recurrent convolutional and Graph convolutional networks to detect phishing sites using URL and HTML. IEEE Access, 10, 82355–82375

[8] Shen, H., Li, B., Peng, H., Xin, J., & Zhang, E. (2021). An effective Cost-Sensitive XGBOOST method for malicious URLs detection in imbalanced dataset. IEEE Access, 9, 93089–93096

[9] Lee, J., Lee, Y. H., Lee, D., Kwon, H., & Shin, D. (2021). Classification of attack types and analysis of attack methods for profiling phishing mail attack groups. IEEE Access, 9, 80866–80872.M.

[10] Xiuwenliu and Jianming Fu (2020). SPWalk: similar property oriented feature learning for phishing detection. IEEE Access, 8, 87031–87045R.

[11] Zhu, E., Chen, Y., Ye, C., Li, X., & Feng, L. (2019). OFS-NN: An effective phishing websites detection model based on optimal feature selection and neural network. IEEE Access, 7, 73271–73284..

[12] J Alsariera, Y. A., Adeyemo, V. E., Balogun, A. O., & Alazzawi, A. K. (2020). AI Meta-Learners and Extra-Trees algorithm for the detection of phishing websites. IEEE Access, 8, 142532–142542.

[13] Gandotra E., Gupta D, "An Efficient Approach for Phishing Detection using Machine Learning", Algorithms for Intelligent Systems, Springer021, 10.1007/978-981-15-8711-5_12.

[14] Chiew KL, Chang EH, Tiong WK,"Utilisation of website logo for phishing detection", Computer Security, pp.16–26, 2015

[15] Tan CL, Chiew KL, Wong K, "PhishWHO: phishing webpage detection via identity keywords extraction and target domain name finder", Decision Support Systems, vol. 88, pp 18–27, 2016

[16] Jain A.K., Gupta B.B. "PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning", Cyber Security. Advances in Intelligent Systems and Computing, vol. 729, 2018

[17] Srinivasa Rao R, Pais AR, "Detecting phishing websites using automation of human behavior", In: Proceedings of the 3rd ACM workshop on cyber-physical system security, ACM, pp 33–42, 2017.

[18] Rao RS, Pais AR. Jail-Phish: An improved search engine based phishing detection system. Computers & Security. 1;83:246–67 2019

[19] Ramalakshmi, K. , Srinivasa Raghavan, V., Enhanced prediction using deep neural network-based image classification', Imaging Science Journal, 2023, 71(5), pp. 472–483

[20] Suleman, M. T., & Awan, S. M. (2019). Optimization of URL-Based Phishing Websites Detection through Genetic Algorithms. Automatic Control and Computer Sciences, 53(4), 333–341.