Phishing URL detecting using Machine Learning

M. G. K. Nayanathara
Dept. of Computer Systems and Network Engineering
Sri Lanka Institute of Information Technology
Malabe,Sri Lanka
IT20174644

***Abstract*— Phishing attacks, which exploit human vulnerabilities and compromise sensitive information, continue to represent serious hazards in today's linked digital world. This paper describes a novel method for detecting phishing URLs using machine learning models. The report recognizes the important connection between phishing assaults and cyber insurance, emphasizing the need of limiting financial losses through comprehensive cybersecurity risk management measures. Our suggested detection technique obtains good accuracy rates in identifying phishing websites by extracting and analyzing numerous URL and HTML elements. The research also looks at other studies that highlight the effectiveness of supervised deep learning classification, optimization approaches, and machine learning algorithms for phishing detection. The methodology section describes the feature extraction and dataset preparation processes, as well as the evaluation of multiple models, including Decision Trees, Random Forests, Multilayer Perceptrons, XGBoost Classification, and Support Vector Machines, and how the XGBoost classifier was chosen as the best model. The results show that these models function well, offering insights into feature relevance and correct classification of phishing URLs. This study contributes to the field of cybersecurity by providing a practical and effective way to resist phishing assaults, hence improving the security of persons and organizations in the digital realm.**

Keywords—Phishing, Cyber Insurance, Decision Trees, Random Forests, Multilayer Perceptrons, XGBoost Classification, and Support Vector Machines

I. INTRODUCTION

This research paper contains information regarding a phishing URL detection model. The issue of phishing has become more pervasive in today's interconnected digital world, where we rely largely on email, online banking, and social networking. Phishing is a type of cyber-attack in which hostile actors seek to trick people into disclosing sensitive information such as passwords, credit card numbers, or personal information. These attackers masquerade as trustworthy entities, frequently through skillfully written emails, websites, or communications, with the goal of exploiting human frailty and gaining unauthorized access to sensitive information. Phishing assaults can take many different shapes, but they all share some characteristics. Email phishing is a widespread practice in which attackers send false emails that look to be from reputable organizations or individuals [1]. These emails frequently use social engineering strategies to create a sense of urgency or importance in order to compel the receiver to take quick action. The email could include a link to a bogus website made to look like a trusted organization's login page, deceiving consumers into entering their credentials [1]. Spear phishing is another strategy that targets specific persons or organizations by personalizing the assault to abuse their personal or professional relationships [2]. Falling prey to a phishing assault can have serious implications. Attackers can gain unauthorized access to personal or financial accounts, resulting in identity theft, financial loss, or reputational damage. Phishing attacks not only target people, but they also pose a substantial risk to corporations and organizations [1]. A successful phishing assault can lead to data breaches, compromised networks, or unauthorized access to critical corporate information, resulting in financial loss and reputational harm [2]. To defend oneself against phishing assaults, you must employ a combination of alertness, education, and technological precautions [2]. To begin, it is critical to cultivate a healthy skepticism towards unwanted emails or communications. Be wary of emails that ask for personal information or demand fast action. Look for phishing signals like generic greetings, spelling or grammatical problems, or questionable URLs [2]. Always confirm the sender's identity by contacting the organization directly through official methods.

The idea behind this project originated from the topic cyber insurance. The link between phishing and cyber insurance is an important part of individual and organizational cybersecurity risk management. Phishing attacks pose a substantial risk to the security and privacy of sensitive information According to the statistics provided by the Internet Crime Complaint Center of Federal Bureau of Investigation the rate of social engineering attacks such as phishing attacks have increased due to the pandemic [3]. Figure 1 contains the statistics in a graph comparing to prior the covid-19 pandemic.

Type of cyberattack	2019	2020	Percentage of increase/decrease
Social attacks (phishing, vishing, pharming, etc.)	114,702	341,342	197.60%
Credit card fraud	14,378	17,614	22.50%
Investment scams	3,999	8,788	119.80%
Malware	2,373	1,423	-40%
Identity theft	16,053	43,330	169.90%
Ransomware	2,047	2,474	20.90%
Denial of service (including TDoS)	1,353	2,018	49.20%

Figure 01 [3]

Cyber insurance can provide financial protection if an attack is successful. Phishing attacks frequently target individuals or organizations in order to steal sensitive information or obtain unauthorized access to networks and systems. Financial losses can arise from these assaults, which include monies taken from bank accounts, fraudulent transactions conducted with compromised credit card information, and even ransom demands from ransomware operations [3]. Furthermore, phishing attempts can result in reputational harm, legal liabilities, and the costs of investigating and combating the attack. Cyber insurance is intended to assist in mitigating the financial effect of cyber-related incidents such as phishing attacks. It often covers costs associated with data breaches, network security breaches, and other cyber catastrophes [3]. While specific coverage and terms vary depending on the insurer and policy, cyber insurance typically covers costs such as legal fees, forensic investigations, notification and credit monitoring services for affected individuals, public relations efforts, and even ransom payments in the case of ransomware attacks [4]. When it comes to phishing, cyber insurance can offer financial security by covering the costs of minimizing the attack and recovering from its aftermath. For example, if a company is the victim of a phishing assault that results in a data breach, cyber insurance can assist cover the costs of investigating the incident, alerting affected individuals, offering credit monitoring services, and

dealing with any legal obligations that may emerge [4].

This project creates a website that can detect phishing URLs by using machine learning models.

II. LITERATURE REVIEW

A. *An efective detection approach for phishing websites using URL and HTML features*

Ali Aljofey, Qingshan Jiang, Abdur Rasool1, Hui Chen, Wenyin Liu, Qiang Qu & Yang Wang has emphasized the growing menace of phishing websites, which pose serious threats to internet users by deceiving them into disclosing personal information [5]. The study suggests a novel strategy to addressing the anti-phishing problem by extracting and analyzing several aspects of suspected webpages in order to effectively identify large-scale phishing offences [5]. The proposed strategy outperformed the existing baseline approaches, achieving an accuracy of 96.76% with only 1.39% false-positive rate on the dataset and an accuracy of 98.48% with 2.09% false-positive rate on the benchmark dataset [5]. The study suggests a novel strategy to addressing the anti-phishing problem by extracting and analyzing several aspects of suspected webpages in order to effectively identify large-scale phishing offences. The proposed strategy outperformed the existing baseline approaches, achieving an accuracy of 96.76% with only 1.39% false-positive rate on the dataset and an accuracy of 98.48% with 2.09% false-positive rate on the benchmark dataset. The study suggests a novel technique to addressing the anti-phishing problem by combining URL and HTML properties [5]. The proposed strategy outperformed the existing baseline approaches, achieving an accuracy of 96.76% with only 1.39% false-positive rate on the dataset and an accuracy of 98.48% with 2.09% false-positive rate on the benchmark dataset. The authors believe that the suggested method detects phishing websites effectively and can be used as a dependable tool to safeguard internet users from phishing attempts [5].

B. *Smart Phishing Detection in Web Pages using Supervised Deep Learning Classifcation and Optimization Technique ADAM*

L. Lakshmi,M. Purushotham Reddy,Chukka Sant haiah, U. Janardhan Reddy had examined the rise in online web activity and the related rise in online fraud rates, with a focus on online transaction websites such as banking and commercial business

[6]. Phishing attacks, in which hackers seek to obtain sensitive information from users such as bank account data and credit card passwords, pose a severe danger to web security. The research concludes that the proposed method of detecting phishing websites using supervised deep learning classification and optimization algorithms with a feature vector of 30 parameters is effective [6]. To distinguish between phishing and real websites, the suggested deep learning model with Adam Optimizer employs a Listwise technique. When compared to other classic machine learning algorithms such as SVM, Adaboost, and AdaRank, the suggested approach performs well [6]. The results suggest that the proposed approach detects phishing websites more accurately. The research also emphasises the significance of employing deep learning techniques for detecting phishing websites, as the size of the web is rapidly rising and present methods have not delivered adequate accuracy for large datasets [6].

C. *Detection and Prevention of Phishing Websites using Machine Learning Approach*

Vaibhav Patil, Pritesh Thakkar, Chirag Shah, Tushar Bhat, Prof. S. P. Godse had addressed the notion of phishing, which is a social engineering assault that exploits vulnerabilities in the user's system [7]. The study seeks to address this issue by providing a layer of user protection. Phishing attacks occur when a criminal sends an email or URL posing as someone or something, they are not in order to obtain sensitive information from the victim. Because of curiosity or a sense of urgency, the victim may provide their password, login, credit card number, and so on.
Using machine learning techniques and algorithms, the research provides three methodologies for detecting phishing websites. The first way examines various elements of the URL; the second examines the legality of the website by learning where it is hosted and who manages it; and the third examines the website's genuineness using visual appearance-based analysis [7]. The paper finds that these approaches can be helpful in detecting phishing websites and providing additional user security [7].

- ".25". Use "cm3", not "cc". (*bullet list*)

D. *Detecting Phishing Websites Using Machine Learning*

Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Dr.Aram Alsedrani had examined the internet's growing importance in people's lives and

how it has evolved into a helpful platform for public transactions such as e-banking and e-commerce [8]. This has, however, resulted in security risks targeting private information, which has become a huge security issue. The research focuses on applying machine learning approaches to detect phishing websites [8]. According to the article, a machine learning-based system can detect phishing websites with high accuracy. When a phishing website is found, the system alerts the user via a browser extension. Because of its high performance, the Random Forest approach was utilized for categorization [8]. The research tested several feature combinations and concluded that a combination of 26 features resulted in an accuracy of 98.8%. The study also examines the normal distribution curve of feature combinations and includes a table of maximum and minimum accuracy values for various feature combinations [8].

E. *Detecting Phishing Website Using Machine Learning The word "data" is plural, not singular.*

Mohammed Hazim Alkawaz, Stephanie Joanne Steven, Asif Iqbal Hajamydeen mentions that phishing is a type of cybercrime in which attackers spoof a reputable website in order to deceive consumers into disclosing personal information [9]. This study addresses the frequency of phishing assaults in a variety of industries, including online payment and webmail, and suggests a phishing detection system to inform users when they reach blacklisted URLs or phishing websites. The paper concludes that the suggested phishing detection system is effective at warning users when they reach blacklisted URLs or phishing websites [9]. The system offers unique capabilities such as directly recording blacklisted URLs from the browser, informing users via pop-ups and email, and offering significant monitoring functionality [9]. The system is intended to assist users in being vigilant when visiting blacklisted websites.

III. METHODOLOGY

The project starts by extracting the features of two data sets including malicious and non-malicious datasets. A random sample strategy is performed to both genuine and phishing URL Data Frames to ensure a balanced dataset. A total of 5000 entries are chosen at random from each Data Frame, resulting in the creation of two new Data Frames containing a representative subset of the original

data. Following that, many URL features are extracted in order to differentiate between legitimate and phishing URLs. Custom functions are defined to do various checks and computations. These functions examine various aspects of URLs, such as the domain, the presence of an IP address or the '@' symbol, URL length, depth, redirection, HTTPS token, prefix/suffix, URL shortening, DNS record, web traffic, domain age, domain end, iframe usage, mouse over behavior, right-click disabling, and web forwards. These functions take a URL as input and return a binary value indicating the presence or absence of properties. It iterates over each URL in the legitimate and phishing URLs Data Frames using these feature extraction algorithms. Then a collection of features for each URL is computed by running the appropriate routines and appending them to a feature list. In addition, each URL is labelled, with 0 indicating legal URLs and 1 denoting phishing URLs. The retrieved features and labels are organized into two separate Data Frames: one for valid URLs and one for phishing URLs. These Data Frames are referred to as 'legitimate' and 'phishing,' respectively. Each Data Frame is made up of rows that represent individual URLs and columns that reflect the extracted features and the label. Finally, the 'legitimate' and 'phishing' Data Frames are concatenated into a single Data Frame called 'Finaldata', which combines the balanced dataset of valid and phishing URLs. Based on the retrieved features, this Data Frame is then ready for additional analysis and model training to detect and classify URLs as real or phishing.

Now in the main function the data distribution is visualized by creating histograms for each feature using the 'hist()' function from matplotlib and heat map was a correlation heatmap is generated using seaborn's 'heatmap()' function to examine the correlation between different features in the dataset.

Figure 2 shows the histograms while figure 3 shows the heat map.

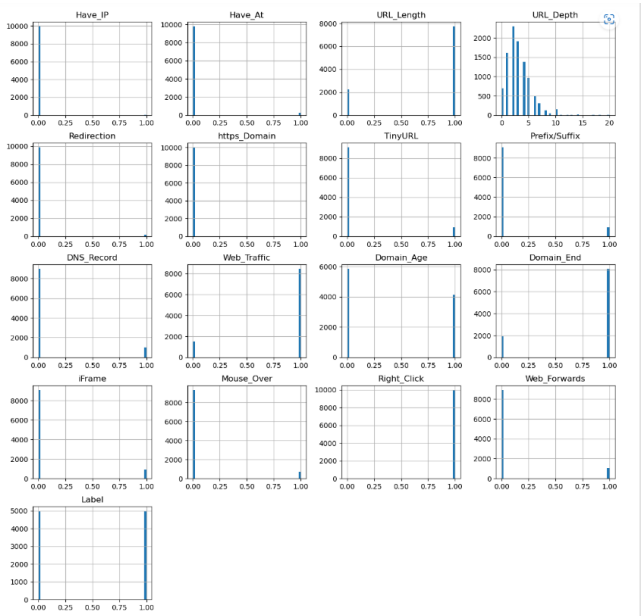

Figure 2

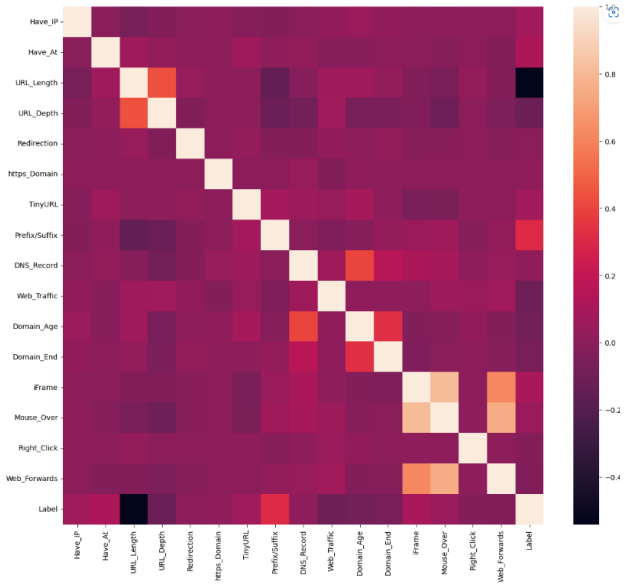

Figure 3

Some columns were removed from the Data Frame to prepare the data for modelling, and the resulting Data Frame is assigned to a new variable named 'data'. In the 'data' DataFrame, null or missing values are tested using the 'isnull()' method, which gives the sum of missing values for each column. When separating the dataset into training and test sets, the rows in the 'data' DataFrame are shuffled at random to ensure equitable distribution. For shuffling, use the 'sample()' function with the 'frac' argument set to 1 and the 'reset_index()' method. The characteristics and target columns are separated and assigned to variables 'X' and 'y', respectively. 'X' denotes the features omitting the 'Label' column, while 'y' denotes the target column ('Label').Using the 'train_test_split()' method from sklearn's'model_selection' module, the dataset is divided into training and test sets. For reproducibility, the test set size is set to 20% of the total dataset, and a random state of 12 is utilised. The variables 'X_train', 'X_test', 'y_train', and 'y_test' are assigned to the training and test sets.The performance evaluation phase starts with importing the 'accuracy_score' function from sklearn's'metrics' module. To save the model performance results, a 'storeResults()' function is defined. The function accepts as parameters the model name, training accuracy, and test accuracy and appends them to distinct lists: 'ML_Model', 'acc_train', and 'acc_test'. The Decision Tree model is the first to be reviewed. A 'DecisionTreeClassifier' instance is created with the maximum depth set to 5. The 'fit()' method is used to fit the model to the training data. Following that, the model is utilized to predict the target values for both the training and test sets. The model's accuracy on training and test data is calculated using the 'accuracy_score()' function, which compares predicted values to actual target values. The precision is printed on the console. Using matplotlib's 'barh()' function, a bar plot is

constructed to visualise the feature importance in the Decision Tree model. The 'feature_importances_' element is used to retrieve the feature importances from the model. 'yticks()' is used to assign feature names to the y-axis ticks. The 'show()' method is used to display the plot.For the Random Forest, Multilayer Perceptrons, XGBoost Classification, and Support Vector Machine (SVM) models, the process is repeated. Each model is instantiated, fitted to the training data, used to forecast target values, and its accuracy is evaluated. The Random Forest model's feature importance's are plotted.For the Random Forest, Multilayer Perceptrons, XGBoost Classification, and Support Vector Machine (SVM) models, the process is repeated. Each model is instantiated, fitted to the training data, used to forecast target values, and its accuracy is evaluated. The Random Forest model's feature importance's are plotted. To organize the model performance data, a Data Frame named 'results' is generated, which includes the model names, training accuracies, and test accuracies.

	ML Model	Train Accuracy	Test Accuracy
3	XGBoost	0.867	0.864
2	Multilayer Perceptrons	0.856	0.855
1	Random Forest	0.814	0.818
0	Decision Tree	0.812	0.818
4	SVM	0.801	0.805

Figure 4

XGBoost was chosen as the best model after referring to the table shown in figure 4.

A. *Result*

The trained model was implemented into a website form where, when the URL is added to the form it generates the outcome in a string whether it was a malicious website or not.

CONCLUSION

This paper presented a phishing URL detection website that made use of machine learning approaches. In today's linked digital world, the increasing incidence of phishing assaults poses major hazards to individuals and organizations. The suggested algorithm demonstrated great accuracy in identifying phishing websites by extracting and analyzing numerous URL properties. The findings emphasize the usefulness of machine learning in preventing phishing threats, as well as the need of user education and technology safeguards. This study contributes to continuing efforts to improve cybersecurity and protect sensitive data in the digital realm.

ACKNOWLEDGMENT

I would like to thank Dr.Lakmal Rupasinghe and Mrs. Chethana Liyanapathirana, lecturers in charge of the module Secure Software Systems for giving me a chance to gain more knowledge regarding machine learning and malware detection system creation by completing this review paper. Next, I am thankful for all the authors of the literature references I have used, for researching thoroughly on these topics.

REFERENCES

[1] phishing.org, "What Is Phishing?," [Online]. Available: https://www.phishing.org/what-is-phishing.

[2] wikipedia, "Phishing," wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Phishing. [Accessed 25 04 2023].

[3] G. T. Aliza Vigderman, "Cyber Insurance Statistics," Security.org, 27 Jan 2023. [Online]. Available: https://www.security.org/insurance/cyber/statistics/. [Accessed 10 May 2023].

[4] CyberPolicy, "Phishing," CyberPolicy, [Online]. Available: https://www.cyberpolicy.com/threats/phishing. [Accessed 10 May 2023].

[5] Q. J. A. R. H. C. W. L. AliAljofey, "This article examines the rise in online web activity and the related rise in online fraud rates, with a focus on online transaction websites such as banking and commercial business. Phishing attacks, in which hackers seek to obtain sensitive information," 2022.

[6] M. P. R. S. L. Lakshmi, "Smart Phishing Detection in Web Pages using Supervised Deep Learning Classifcation and Optimization Technique ADAM," vol. 07, p. March , 2021.

[7] P. T. C. S. T. B. P. S. P. G. Vaibhav Patil, "Detection and Prevention of Phishing Websites using Machine Learning Approach," 2018.

[8] B. A. N. A. D. A. Amani Alswailem, "Detecting Phishing Websites Using Machine Learning," 2019.

[9] S. J. S. A. I. H. Mohammed Hazim Alkawaz, "Detecting Phishing Website Using Machine Learning," 2020.