DALARNA UNIVERSITY

Degree Thesis

Understanding and Mitigating Phishing Attacks

Author: Hani Shaker
Course name: Thesis report in Microdata analysis
Course code: GMI2C8
Credits: 15hp
Supervisor: Nausheen Saeed
School: Högskolan Dalarna
Examiner: Joonas Pääkkönen and Yves Rybarczyk
Date of examination: 2024/05/19

Abstract

Phishing is a prevalent cyber-attack method where attackers disguise themselves as trustworthy entities to deceive individuals into divulging sensitive information such as usernames, passwords, and financial details. This study aims to enhance phishing detection and prevention mechanisms by identifying the most significant features that distinguish phishing websites from legitimate ones and by evaluating the effectiveness of different machine learning models in phishing detection. A quantitative data analysis approach was employed, leveraging a dataset from the UCI Machine Learning Repository. The study focused on attributes such as SSL/TLS certificates, URL manipulation, and subdomain usage to identify potential phishing sites. The Random Forest model and Logistic Regression model were compared to determine their accuracy and reliability in detecting phishing websites. The results indicate that the Random Forest model outperforms the Logistic Regression model in terms of precision, recall, F1 score, and accuracy, making it a more robust tool for phishing detection. This research highlights the need for continuous updating of detection models and datasets to keep pace with the evolving tactics of phishing attacks, ultimately contributing to the development of more effective anti-phishing strategies.

Keywords

Phishing, Identity Theft, Phishing Detection, Machine Learning Detection.

Contents

1 INTRODUCTION

1.1 Background

Phishing is a cyber-attack technique where attackers masquerade as trustworthy entities to deceive individuals into divulging sensitive information such as usernames, passwords, and financial details. These attacks typically involve fraudulent emails, websites, or messages that appear legitimate, aiming to trick users into providing personal data. Despite numerous defense mechanisms, phishing remains a significant challenge due to the evolving tactics of attackers and the continuous development of new phishing methods (Jain & Gupta, 2022).

Phishing has developed significantly in terms of how it is carried out. Initially, phishing started as deceptive emails sent out by unknown individuals masquerading as known entities. Over time, it has grown more sophisticated, utilizing modern web development tools to create realistic-looking fake pages and messages.

Phishing methods now include social media messages in addition to emails and even deceptive surveys promising rewards for account information. These methods often come with URLs that lead to fake internet pages appearing like genuine ones, making it hard to notice them.

To counter these evolving threats, detection systems have also advanced. Traditional rule-based systems blocklist known malicious sites but face difficulties in managing new attacks. Machine learning (ML)-based systems inspect historical datasets to observe patterns of activities that can help identify such attempts more effectively, although they are constrained by their reliance on prior data. Continual learning (CL)-based techniques have been developed to adapt to new phishing tactics while still being effective against older ones, providing better protection from the constantly changing nature of phishing attacks (Ejaz, Mian, and Manzoor, 2023).

Various forms of phishing attacks:

Email Phishing: This is where phony emails mimic credible sources to obtain sensitive data from recipients.

Spear Phishing: Targets individuals or organizations with customized messages, making the scam more personal and challenging to detect.

Whaling: Focuses on executives to gain access to sensitive company systems.

Vishing: Uses SMS (smishing) and voice calls (vishing) with the intent to deceive people by pretending to be institutions.

Pharming: Redirects users from legitimate websites to fake ones by tampering with a computer's host file or exploiting DNS server vulnerabilities.

These methods are aimed at convincing clients to give away their information, which can lead to financial loss or identity theft. Understanding these techniques can enable individuals and companies to enhance their protection against phishing threats (Alabdan, 2020).

1.2 Problem Formulation

Phishing attacks have evolved significantly, utilizing phishing kits to deploy attacks more efficiently and on a larger scale. These kits enable cybercriminals to create fake websites quickly, making phishing attacks harder to detect and prevent. The lack of comprehensive datasets associating phishing kits with actual phishing websites has hindered the development of effective detection methods. To enhance phishing detection and prevention, it is crucial to understand the key features that differentiate phishing websites from legitimate ones and evaluate the effectiveness of machine learning models in identifying phishing attacks (Castano et al., 2023).

Phishing websites significantly threaten online security, particularly within the realm of online shopping, where they are designed to steal sensitive user information such as login credentials and payment data by mimicking legitimate e-commerce sites. Despite advances in machine learning (ML)-based classifiers for phishing detection, these systems remain vulnerable to evasion attacks. These attacks can manipulate the input features of phishing websites to bypass detection while maintaining the site's deceptive appearance and functionality, making them particularly dangerous in an online shopping context where users frequently enter sensitive information (Song et al., 2021).

1.3 Purpose

The primary purpose of this research is to thoroughly understand phishing attacks by identifying the key features that differentiate phishing websites from legitimate ones and evaluating the effectiveness of various machine learning models in detecting phishing websites. This study aims to enhance the detection and prevention mechanisms against phishing attacks, thereby developing more robust strategies to protect users from falling

victim to phishing. By addressing these aspects, the research seeks to provide a comprehensive approach to improving online security measures against phishing threats.

1.4 Research Questions

1-Which features are the most important for identifying phishing websites?

The research question focuses on identifying the most significant attributes that can help in detecting phishing attacks. Understanding which specific characteristics differentiate phishing sites from legitimate websites allows for enhancing detection and prevention mechanisms. This question is crucial for developing strategies that can protect users from falling victim to fake websites.

2- How effective is the Random Forest model compared to Logistic Regression model in detecting phishing websites?

This research question aims to evaluate the effectiveness of different machine learning models, Random Forest model and the Logistic Regression model, in identifying phishing websites. Comparing these models helps determine which algorithm provides better accuracy and reliability in phishing detection, ultimately aiding in the development of more robust anti-phishing tools.

2 Literature Review

2.1 Machine Learning Detection

Machine learning (ML) is a scientific study of algorithms and statistical models that allow computer systems to perform specific tasks with no programming given, this definition shows how machine learning systems learn and improve by themselves through datasets, using algorithms to analyze, identify patterns, and make decisions without of human efforts (Mahesh, 2020).

A study done recently by Maru Kawale, Varghese, Dagu and Gupta (2024) has shown how effective the models of machine learning are for discovering phishing websites. It illustrates that Random Forest is far better than previous ways because it can deal with big data sets and use them to make more correct predictions.

There are many reasons why this research is important. The first one is that it explores some advanced machine learning techniques such as Random Forest which boasts 96.7 percent accuracy rate in detecting Phishing sites compared to other models. Secondly, high accuracy levels are crucial in arresting the situation where there are few malicious sites that would have gone undetected thereby preventing a lot of damage to users. To do this, a feature set was used containing 32 features from customizing status bar to forwarding websites, which were critical in differentiating safe versus unsafe sites.

Almousa et al (2022) wrote the paper “Phishing website detection: How effective are deep learning-based models and hyperparameter optimization?” which confirmed that performance issues of traditional machine learning (ML) models for detecting phishing websites over time occur due to evolving phishing tactics. This study seeks to investigate the use of continual learning (CL) techniques, including Learning Without Forgetting (LWF) and Elastic Weight Consolidation (EWC), for maintaining high detection accuracy over time.

Traditional ML models were compared to transfer learning (TL), and CL techniques with data from 2018 to 2020. CL methods outperform traditional ML and TL, by adapting to new data but retaining knowledge about earlier types of phishing. It reduces the need for frequent retraining hence sustainable detection ability able to identify robustly potential CL methods that could help in long-term phishing detection.

2.2 Phishing Website Detection

In "Phishing-Inspector: Detection & Prevention of Phishing Websites," Doke, Khismatarau, Jambhale and Marathe discuss new browser extensions designed to mitigate the growing threat of phishing attacks Traditional blacklisting and whitelisting techniques not enough due to the ease of creating fake websites. The proposed extension enhances detection by checking SSL certificates and tracking domain creation and expiration dates is important indicators of potential fake sites. the authors provide tools and extensions that help users better identify and avoid phishing websites.

The article “Why Phishing Still Works: User Strategies for Combating Phishing Attacks” by Alsharnouby, Alaca, and Chiasson examines how detection of phishing websites can be studied through a user study. The participants were shown some web pages and then asked to find instances of phishing with their eye movements being tracked to know what visual cues influenced their decisions most. The study reveals that even if users recognize the task at hand, they could only identify 53% of phishing websites. Moreover, it was found out that those users generally spent more time on looking at website content rather than browser security indicators like the URL or HTTPS status. It shows the need for better training on security cues for users coupled with stronger browser extenti0ons defenses against phishing.

2.3 Phishing Types

The Kang Leng Chiew, Kelvin Sheng Chek Yong and Choon Lin Tan article “A Survey of Phishing Attacks: Their Types, Vectors and Technical Approaches” gives a good view of various phishing methods that exist, how they are spread, and the techniques used by the attackers. It outlines how phishing has evolved over time highlighting different ways phishing attacks are carried out and explains what is done technically in these attacks.
They discuss important platforms that make it easy to do phishing attacks like, voice and sms among others. Additionally, they described some other vectors like social web site pages, emails etc. This also includes various other technical approaches like browser vulnerabilities, clickjacking attacks cloud computing vulnerabilities including cross site scripting (xss) and others which are many attackers use to trick people to get what they want.

Disguised email attacks are made up of portal website information mail phishing attacks, where the attackers will send emails that look like they are from trusted sites asking users to give their account information Another example of disguised email attack is secure email phishing attack that hide under the mask of being secure emails and trick users into inputting their account details in the phishing sites.

Email body vulnerability attacks refer to malicious script embedding that take place within an email body opening which as a result leads to phishing sites without necessarily clicking on any links. This means this method exploits vulnerability in HTML supporting webmail or e-mail client applications.

Attached file camouflage offences involve disguised attached files phishing site connections with mails retain linkages with attachments, that when clicked upon send individuals to phishing sites that seek for account credentials. Through one way or another, attackers also use this technique by inserting malicious code into hwp (Hangul Word Processing) documents among other attachments whose exploit vulnerability enables it to execute malicious activities after being opened by the user (Lee et al., 2021).

3 METHOD

In this study, quantitative data analysis is used to make sense of numerical information gathered from experiments and surveys. Quantitative data, based on numbers, is essential for getting conclusions and identifying patterns. Various techniques and methods help to organize and explain this data, providing good insights. These methods aid researchers in validating, cleaning, and analyzing data effectively to get good results (David, 2024).

3.1 Method approach

Literature studies were retrieved from Summon and Google Scholar. The following keywords were used to find relevant publications: "Phishing," "Identity Theft," "Phishing Detection," and "Machine Learning Detection." A quantitative data analysis approach has been used to answer the research questions and achieve the objectives outlined in the problem statement.

3.2 Data Collection

This dataset was retrieved from the UCI Machine Learning Repository, contains 11,055 instances, each described by 31 attributes, and is entirely free of missing values. The dataset has two class labels: Phishing, labeled as -1, and non-phishing, labeled as 1. Among the 11,055 instances, 4,898 belong to the Phishing class, while 6,157 are categorized as non-phishing (Mohammad & McCluskey, 2015). This dataset was chosen for its relevance to phishing detection, as it includes features specifically designed to identify phishing websites. The quality and completeness of the dataset, free of missing values, ensure accurate analysis without extensive data cleaning. The dataset's size and balanced distribution between phishing and non-phishing instances provide a comprehensive basis for model training. Additionally, its use in previous research indicates reliability and suitability for this study. The richness in features, capturing various aspects of website characteristics, allows for a thorough investigation into predictive attributes of phishing activity. These factors make this dataset an ideal choice for enhancing phishing detection methods.

3.3 Data Processing

The UCI Machine Learning Repository in 2019 was the source of all data used for this analysis. Data is provided as a CSV file and processed using Pandas within a Jupyter Notebook.
To optimize memory usage, numeric columns were converted to 32-bit floats and integers. The data set was examined for missing values to confirm that it is complete and ready for

analysis. Exploratory Data Analysis (EDA) utilized bar charts and pie charts to show how many phishing and none-phishing attempts are there in the data set from all instances.

Spearman Correlation analysis was employed to detect and visualize relationships between many features with the target class, which is phishing and non-phishing, and make a heatmap out of the correlation to show which features were positively or negatively correlated with target Class (phishing or non-phishing websites).

In the article by Jan Hauke and Tomasz Kossowski (2011), they discuss about Spearman correlation coefficient, which help us to see how closely two variables are related without saying that the relationship between them is linear. Spearman's method works well even when the data isn't on a typical scale like time or temperature, making it flexible and good for different types of data (Hauke & Kossowski, 2011).

Features and target variable were separated, and the data was split into training and test sets.
Random Forest classifier was trained, and its performance was evaluated using Precision, Recall, F1 Score, and Accuracy.
Logistic Regression model was trained, and its performance was evaluated using the same metrics and the performance metrics of both (Random Forest and Logistic Regression) models were plotted for comparison.

Random Forest (RF) is a widely used and effective machine learning method. It creates many decision trees during training and then combining the outputs to make prediction. This method helps improve accuracy and stability. RF high performance and it can rank features by importance. This method helps researchers understand which features are most important in their predictions (Devetyarov & Nouretdinov, 2010).

The second model is Logistic regression which is statistical method used to model the probability of a binary outcome. It uses the natural logarithm of the odds of the event to estimate how changes in predictor variables affect the likelihood of the outcome. The model outputs an odds ratio, which describes how the odds change with a one-unit increase in a predictor (LaValley, 2008).

Random Forest was chosen for its robustness and ability to handle large datasets with higher dimensionality. Its ensemble approach enhances prediction accuracy and reduces overfitting by averaging multiple decision trees. This method's capability to rank feature importance is particularly useful for identifying the most predictive attributes in phishing detection.

Logistic Regression was selected due to its simplicity and interpretability. It provides clear insights into the relationship between predictors and the binary outcome, making it easier to understand the effect of each feature on phishing likelihood. Additionally, Logistic Regression's efficiency in handling binary classification tasks complements the comprehensive analysis provided by Random Forest, offering a balanced evaluation of the dataset.

Precision measures the accuracy of positive predictions. It is defined as the ratio of true positives to the sum of true positives and false positives.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Positive}}$$

Recall measures the ability of a model to find all the relevant cases (true positives) within a dataset. It is defined as the ratio of true positives to the sum of true positives and false negatives.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}}$$

The F1 Score can be used when you are looking for balance between Precision and Recall, Which is a way to combine both Precision and Recall.

$$\text{F1 Score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}}$$ (Shung, 2018).

4 ANALYSIS

4.1 Phishing Instances

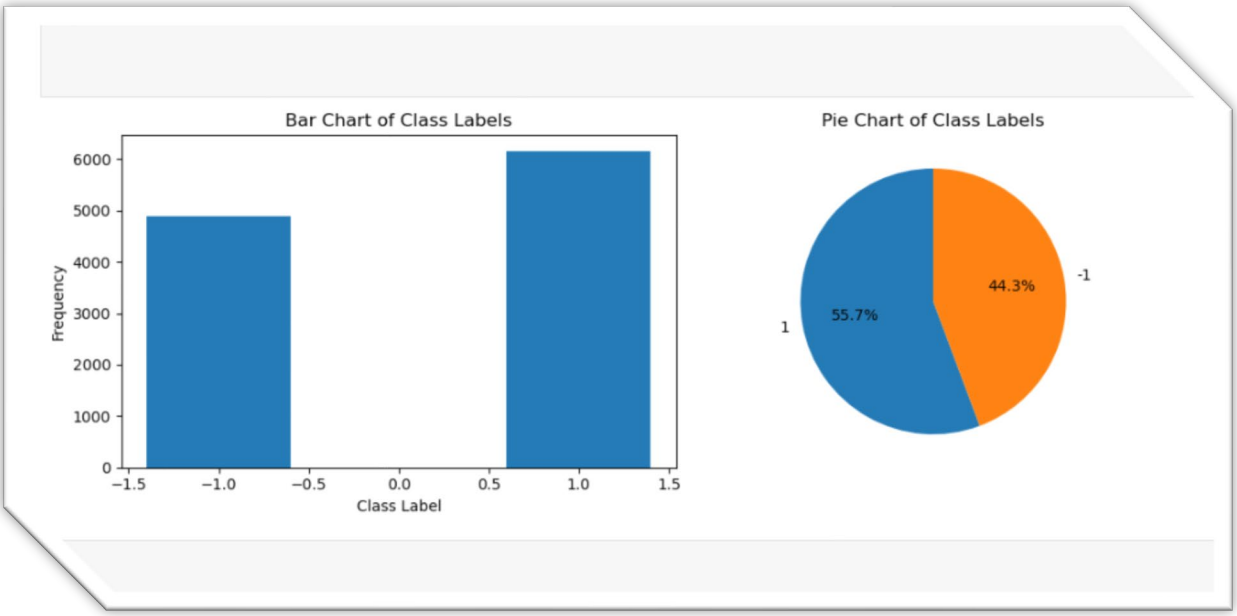

Figure 1 Phishing Instances

Figure 1 highlight the distribution of phishing versus non-phishing cases in both absolute (bar chart) and relative (pie chart) terms. These visualizations show whether the dataset is balanced or if one class is more prevalent than the other. As we see from the 11,055 instances, there are 55.7% non-phishing (1) instances and 44.3% phishing (-1) instances. This distribution indicates a relatively balanced dataset, which is beneficial for training machine learning models as it reduces the risk of bias toward either class.

4.2 Spearman Correlation

Figure 2, Figure 3 and Figure 4 drawn using heatmaps and they shows the relationships between the futures and target Class (which shows whether an instance is a phishing website or not) while the number of attributes in each of them tells us if positively/negatively correlated with phishing detection.

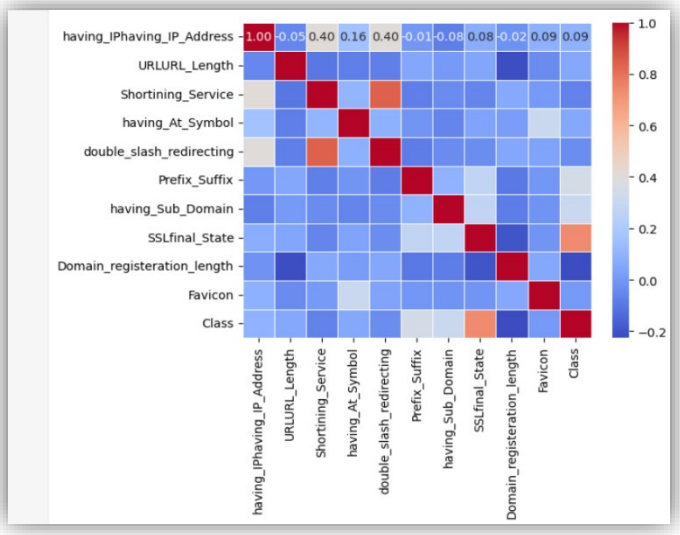

Figure 2 Spearman Correlation 1-10

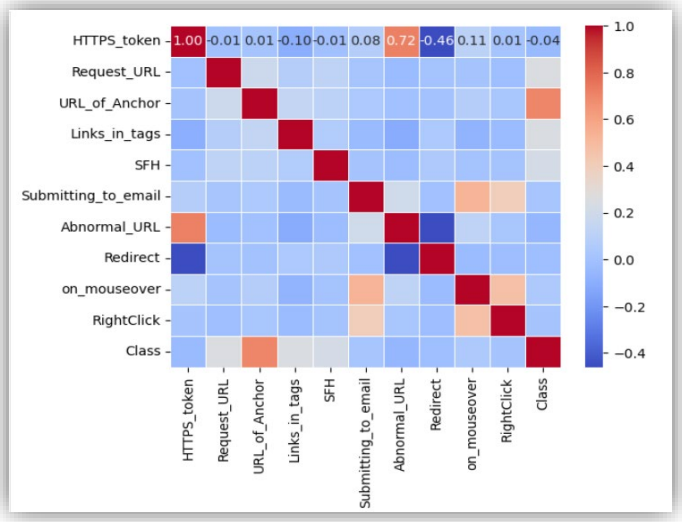

Figure 3 Spearman Correlation 11-20

Additionally, this visual representation of correlations also enables identification of potential redundancies and patterns that differentiate among phishing from non-phishing websites.

Positive Correlation:

Legitimate (non-phishing) sites tend to have more features that positively correlate with Class (where Class = 1 for non-phishing). If its value is higher for a given attribute, then such site will be indicated as non-phishing.

Negative Correlation:

Phishing websites mostly consist of characteristics that negatively correlate with Class (where Class = -1 for phishing). When such a feature dwindles it implies that the website is most likely a phishing site.

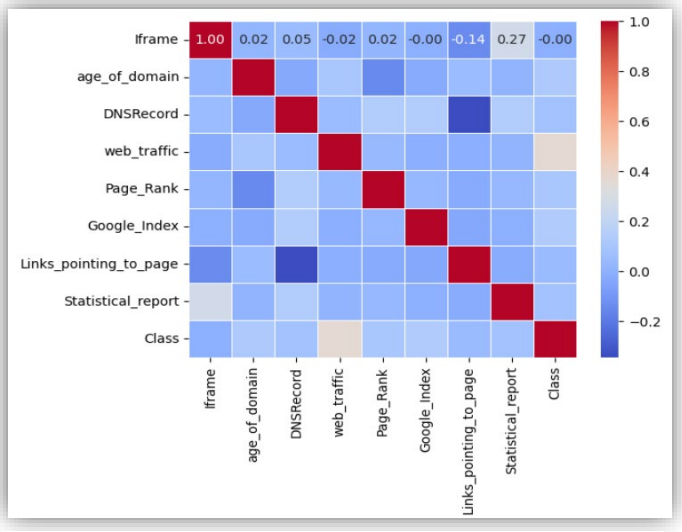

Figure 4 Spearman Correlation 21-28

4.3 Top Features

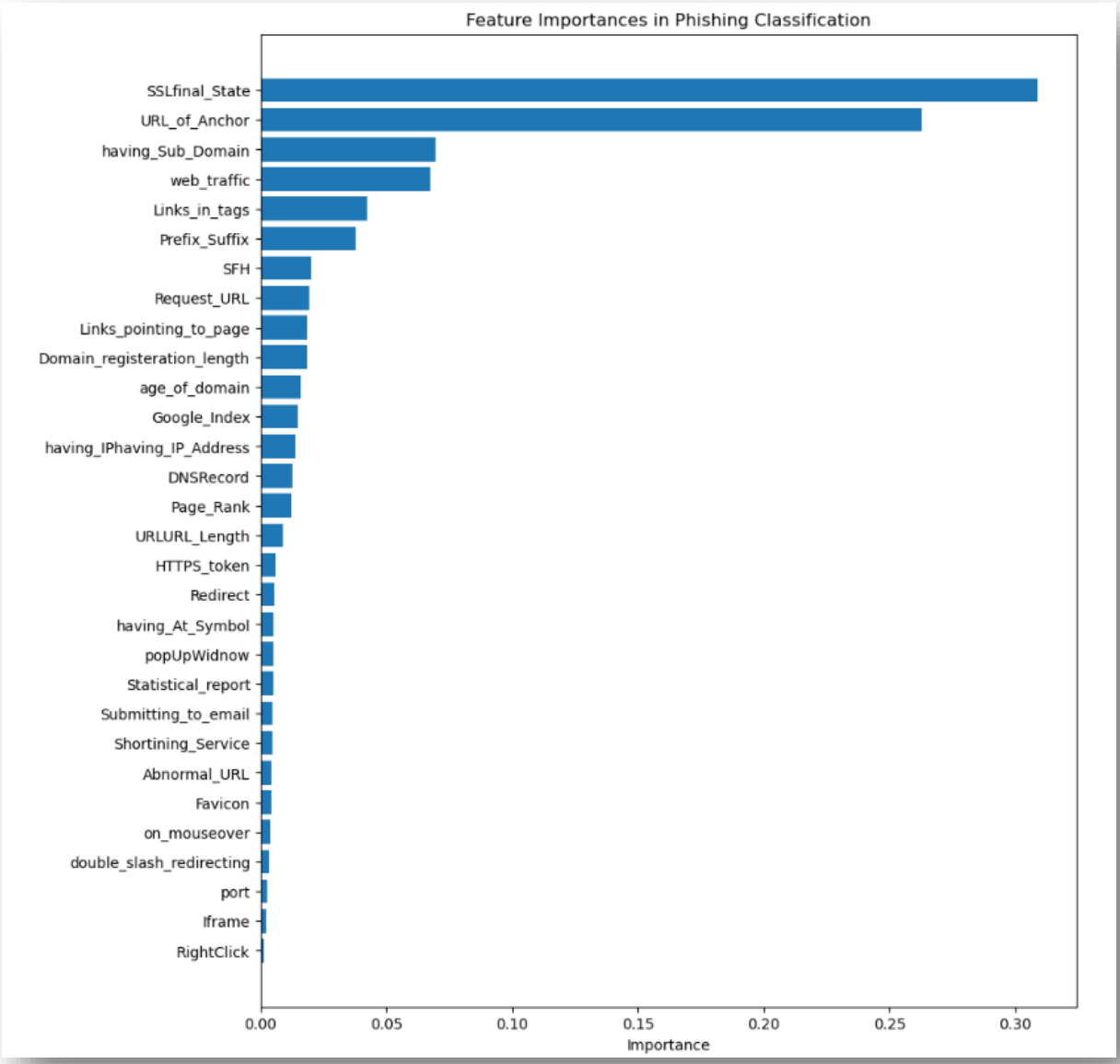

Figure 5 Features Importance

Figure 5 shows the relative importance of various features in distinguishing between phishing and non-phishing websites using a Random Forest classifier. The horizontal bar chart organizes features by their impact on classification, with the most important features at the top, for example:
web_traffic:

Importance: This feature measures the website's traffic. Legitimate websites typically have higher web traffic compared to phishing sites, which often have minimal or erratic traffic patterns.

Links_in_tags:

Importance: This feature assesses the number of links in HTML tags. Phishing sites often use numerous or hidden links to redirect users to malicious pages.

4.4 Precision, Recall, F1 Score and Accuracy

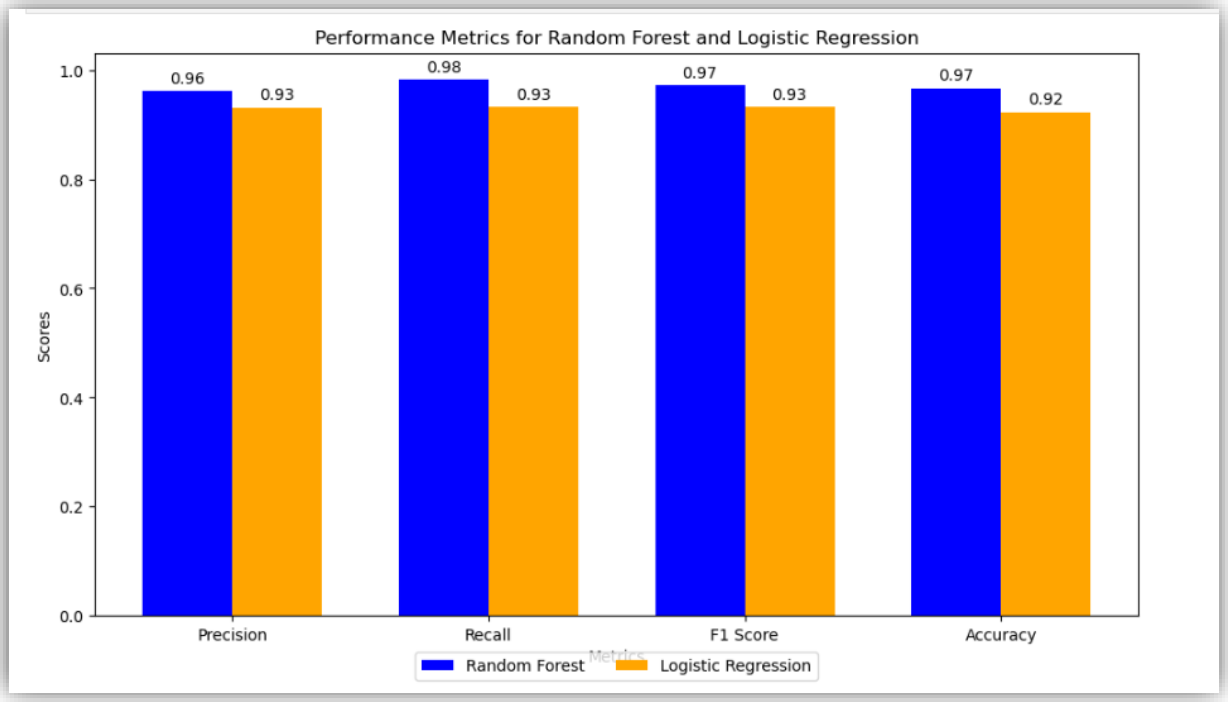

Figure 6 Comparison Between Two Machine Learning Models

Figure 6 compares the performance metrics of two machine learning models: Random Forest and Logistic Regression. It shows four key metrics as we can see, (Precision, Recall, F1 Score, and Accuracy) for each model.

X-Axis:
The x-axis lists those four-performance metrics: Precision, Recall, F1 Score, and Accuracy.
Y-Axis:
The y-axis shows the score values for each metric, ranging from 0.0 to 1.0, and higher values means better performance.
Blue Bars: shows the performance scores of the Random Forest model.
Orange Bars: shows the performance scores of the Logistic Regression model.

5 RESULTS

This section will look at analysis itself, clarify it and answer the research questions.

Spearman Correlation:
The heatmaps show correlations in example like future SSLfinal_State has a positive correlation with Class, which shows that a safe SSL status correlates positively with normal sites or non-phishing sites.

RightClick and Class: Phishing sites usually disable right click function so as not to allow users examine or copy links unlike valid websites. The correlation between RightClick feature and phishing class is negatively substantial.

Iframe and Class: Since the correlation coefficient is almost equal to zero, iframes alone are not enough to determine whether a website is a phishing site.

Research Question 1: Which features are the most important for identifying phishing websites?

To answer this question, we can refer to Figure 5 , which provides valuable information and a clear understanding. The attributes with the highest impact on predicting phishing attacks using the Random Forest Classifier.
SSLfinal_State:

SSLfinal_State determines if a website uses secure SSL/TLS certificates correctly or not. Generally, an authentic page employs valid SSL certificate to encrypt data transmission; however, fake sites often use invalid certificates resulting into pages having unsafe SSL settings.

URL_of_Anchor, amounts to a study of the URL links. One way in which phishing sites do this is by modifying these hyperlinks to lead users to fake websites but appearing as if they were authentic URLs. A mismatch between the link text and the actual URL can be an indication of a phishing attack while high manipulation of anchor tags is a strong sign of a phishing attempt.

having_Sub_Domain examines whether there is presence of subdomain details in the URL supplied. As such, phishing sites often include subdomains that resemble genuine websites'

domain names. This means that, for instance, a website may present itself as a well-known bank by use of "subdomain.website.com" when it is being hosted elsewhere under different domain name. This way, one can identify deceptive subdomains that are indicative of potential phishing attacks.

SSLfinal_State and the other features (URL_of_Anchor and having_Sub_Domain) plays important roles in identifying suspicious web pages which scammers use for illegal activities like stealing user's identity,

Research Question 2: How effective is the Random Forest model compared to the Logistic Regression model in detecting phishing websites?

To answer this question, we can refer to Figure 6 which compares four key performance metrics: Precision, Recall, F1 Score, and Accuracy for both the Random Forest and Logistic Regression models.

	Random Forest	Logistic Regression	Result
Precision	0.96	0.93	The Random Forest model has a higher precision, which means that it is more correct in identifying true positives among the predicted positives. This means fewer false positives compared to the Logistic Regression model.
Recall	0.98	0.93	The Random Forest model also has a higher recall, meaning it correctly identifies a higher proportion of actual positives. This means that the Random Forest model has fewer false negatives than the Logistic Regression model.
F1 Score	0.97	0.93	The F1 Score, which balances precision and recall, is higher for the Random Forest model. Which means that the Random Forest model achieves a better balance between precision and recall compared to the Logistic Regression model.
Accuracy	0.97	0.92	The Random Forest model has a higher accuracy, which means that it correctly predicts a higher proportion of all instances. This shows that in overall Random Forest model is giving better performance compared to the Logistic Regression model.

Table 1 Metrics

6 DISCUSSION

6.1 Comparison with Previous Research

The result of this study brings to the fore specific web features such as SSL certificates, URL manipulation and subdomain usage which make it easy to detect phishing websites that confirmed with previous research.

Common Findings:

SSL Certificates:

This research and previous research shows show important SSL/TLS certificates in proving website legitimacy and other like URL. phishers usually abuse such certificates to make their websites look real.

Use of Machine Learning:

Earlier research and this research both shows that machine learning techniques including Random Forest are effective in differentiating phishing sites. Additionally, this study examines how well these models perform with a focus on their precision and accuracy. And that random forest performs better than other models.

Differences from Previous Research

Focus on Continual Learning:
While previous research, primarily focuses on the use of continual learning (CL) techniques to maintain detection accuracy over time, this study does not incorporate continual learning methods. Instead, it focuses on evaluating traditional machine learning models like Random Forest and Logistic Regression in the context of phishing detection. The emphasis here is on comparing the immediate effectiveness of these models rather than their adaptability over time.

Evaluation Metrics and Model Comparison:
Another key difference is the comprehensive comparison of evaluation metrics between Random Forest and Logistic Regression. Previous studies often highlight the efficacy of one machine learning model or another without detailed comparative analysis. This research provides a side-by-side comparison of precision, recall, F1 score, and accuracy for both models, offering a nuanced understanding of their performance differences. This direct

comparison helps in identifying which model is more reliable and accurate for phishing detection in the given dataset.

6.2 Limitations

1-Evolving Nature of Phishing Attacks:
Phishing attacks continually evolve, and the strategies used by attackers change over time. This study's dataset may not encompass all current phishing tactics, limiting its ability to generalize findings to new or emerging types of attacks.

2-Scope of Machine Learning Models:
The study did not explore all possible machine learning models or advanced techniques that could be employed for phishing detection. This means that while Random Forest and Logistic Regression were effective, there could be other models that might perform better under different conditions

7 CONCLUSIONS

This study provides a comprehensive analysis of phishing detection, focusing on the identification of key features that differentiate phishing websites from legitimate ones and evaluating the effectiveness of machine learning models in detecting phishing attacks. The research utilized a dataset from the UCI Machine Learning Repository, which includes 11,055 instances described by 31 attributes, and employed quantitative data analysis techniques to derive meaningful insights.

The results indicate that certain features, such as SSLfinal_State, URL_of_Anchor, and having_Sub_Domain, play a crucial role in identifying phishing websites. The Random Forest classifier demonstrated superior performance in comparison to the Logistic Regression model, achieving higher precision, recall, F1 score, and accuracy. This highlights the effectiveness of the Random Forest model in accurately detecting phishing websites and reducing false positives and false negatives.

The study confirms that SSL/TLS certificates, URL manipulation, and subdomain usage are significant indicators of phishing activity, aligning with findings from previous research. However, unlike some prior studies that focus on continual learning techniques, this research primarily evaluated traditional machine learning models for immediate effectiveness.

Despite its contributions, the study acknowledges limitations, including the evolving nature of phishing attacks and the scope of machine learning models explored. Future research could benefit from incorporating continual learning methods and exploring a wider range of machine learning techniques to enhance detection capabilities further.

Overall, this research advances the understanding of phishing detection mechanisms and offers valuable insights into improving online security measures. By identifying key predictive features and demonstrating the superior performance of the Random Forest model, the study provides a solid foundation for developing more robust anti-phishing tools and strategies to protect users from phishing threats.

8 REFERENCES

Jain, A. K., & Gupta, B. B. (2022). A survey of phishing attack techniques, defence mechanisms and open research challenges. Enterprise Information Systems, 16(4), 527-565. https://doi.org/10.1080/17517575.2021.1896786

Castano, F., Fernandez, E. F., Alaiz-Rodriguez, R., & Alegre, E. (2023). PhiKitA: Phishing kit attacks dataset for phishing websites identification. IEEE Access, 11, 1-1. https://doi.org/10.1109/ACCESS.2023.3268027

Song, F., Lei, Y., Chen, S., Fan, L., & Liu, Y. (2021). Advanced evasion attacks and mitigations on practical ML-based phishing website classifiers. International Journal of Intelligent Systems, 36(9), 5210-5240. https://doi.org/10.1002/int.22510

Ejaz, A., Mian, A. N., & Manzoor, S. (2023). Life-long phishing attack detection using continual learning. Scientific Reports, 13(1), 11488-11488. https://doi.org/10.1038/s41598-023-37552-9

Mahesh, B. (2020). Machine Learning Algorithms - A Review. International Journal of Science and Research (IJSR), 9(1). Retrieved from https://www.researchgate.net/publication/344717762_Machine_Learning_Algorithms_-A_Review

Kawale, M., Maru, B., Dagu, S., Varghese, M., & Gupta, V. (2024). Machine learning based phishing website detection. Paper presented at the 833-837. https://doi.org/10.23919/INDIACom61295.2024.10498854

Almousa, M., Zhang, T., Sarrafzadeh, A., & Anwar, M. (2022). Phishing website detection: How effective are deep learning-based models and hyperparameter optimization? Security and Privacy, 5(6), n/a. https://doi.org/10.1002/spy2.256

Doke, T., Khismatrao, P., Jambhale, V., & Marathe, N. (2020). Phishing-inspector: Detection & prevention of phishing websites. ITM Web of Conferences, 32, 3004. https://doi.org/10.1051/itmconf/20203203004

Alsharnouby, M., Alaca, F., & Chiasson, S. (2015). Why phishing still works: User strategies for combating phishing attacks. International Journal of Human-Computer Studies, 82, 69-82. https://doi.org/10.1016/j.ijhcs.2015.05.005

Alabdan, R. (2020). Phishing attacks survey: Types, vectors, and technical approaches. Future Internet, 12(10), 168. https://doi.org/10.3390/fi12100168

Chiew, K. L., Yong, K. S. C., & Tan, C. L. (2018). A survey of phishing attacks: Their types, vectors and technical approaches. Expert Systems with Applications, 106, 1-20. https://doi.org/10.1016/j.eswa.2018.03.050

Lee, J., Lee, Y., 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. https://doi.org/10.1109/ACCESS.2021.3084897

David, E. (2024, April 1). Quantitative Data: Types, Analysis & Examples. ProProfs Survey. https://www.proprofssurvey.com/blog/quantitative-data/

Mohammad, R., & McCluskey, L. (2015). Phishing websites [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C51W2X

Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. Quaestiones Geographicae, 30(2), 87–93. https://doi.org/10.2478/v10117-011-0021-1

Devetyarov, D., & Nouretdinov, I. (2010). Prediction with Confidence Based on a Random Forest Classifier. In IFIP advances in information and communication technology (pp. 37–44). https://doi.org/10.1007/978-3-642-16239-8_8

LaValley, M. P. (2008). Logistic regression. Circulation, 117(18), 2395–2399. https://doi.org/10.1161/circulationaha.106.682658

Shung, K. P. (2018). Accuracy, precision, recall or F1? - towards data science. Medium. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9