The Utilization of a Unique Machine Learning Technique for the Detection of Phishing Websites

Mr. M. Gopinath Reddy[1], Y. D. S. Priyanka[2], A. V. Mahendra Reddy[3], K. Gayathri[4], Ch. Mourya[5]

Department of Artificial Intelligence and Machine Learning

mgopinath@sasi.ac.in[1], satyapriyanka.yallamilli@sasi.ac.in[2], mahendrareddy.allatipalli@sasi.ac.in[3], gayathri.korakanchu@sasi.ac.in[4], mourya.chatla@sasi.ac.in [5]

***Abstract*—** Phishing attacks are dangerous because they trick users into giving over private information. The usefulness of machine learning in identifying phishing URLs is examined in this study. We use a large dataset that includes URL structure, website content, and external data to train and assess several models.
Key features are found through exploratory data analysis, and feature engineering improves the model's capabilities even more.
Analysis shows that the model's decisions are greatly influenced by the existence of HTTPS, URL anchor text, and website traffic patterns. We stress the value of user education as an additional defensive mechanism and recognise the necessity of frequent model changes as a result of changing phishing techniques. Future research directions include looking into ensemble models, examining external data sources, and keeping an eye on phishing trends in order to develop better detection techniques.

Keywords

Phishing Detection, Machine Learning, Gradient Boosting, Feature Importance

I. INTRODUCTION

Attackers are constantly evolving their tactics to circumvent well-established security measures, and phishing remains one of the most prevalent cyberthreats. In order to fool users into disclosing personal information, such as login passwords and bank details, phishing websites seem to be reliable services . Newly created phishing URLs are often missed by existing solutions, particularly heuristic and blacklist-based detection methods, leading to a significant increase in undetected threats. In view of these drawbacks, recent research has examined the application of machine learning techniques for phishing detection. However, many machine learning-based approaches fall short in addressing dimensionality reduction and feature selection, resulting in suboptimal performance. To address these issues, this study proposes a combination of machine learning algorithms to more effectively and precisely identify phishing websites.

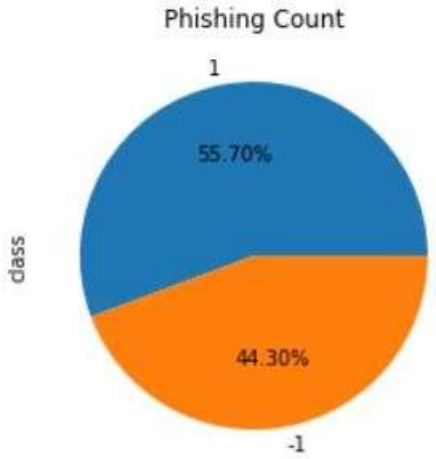
Phishing assaults are a surprise and a continuous danger to the online world. These assaults employ dishonest tactics to fool users into disclosing private information or posing as trustworthy websites in order to download malicious software. Blacklists and other contemporary techniques for detecting phishing attacks frequently fall behind the evolving tactics of phishers. By examining several aspects of websites, machine learning has become a viable technique for identifying phishing URLs in recent years. This article investigates how well different machine learning models identify phishing attempts. Our study trains and tests models using a wealth of data from phishing websites, then compares the models' results to identify the most effective approach. Our research has significant ramifications for creating stronger phishing defences.

II. METHODOLOGY

2.1 Data Acquisition

Phishing assaults are a surprise and a continuous danger to the online world. These assaults employ dishonest tactics to fool users into disclosing private information or posing as trustworthy websites in order to download malicious software. Blacklists and other contemporary techniques for detecting phishing

attacks frequently fall behind the evolving tactics of phishers. By examining several aspects of websites, machine learning has become a viable technique for identifying phishing URLs in recent years. This article investigates how well different machine



learning models identify phishing attempts. Our study trains and tests models using a wealth of data from phishing websites, then compares the models' results to

Fig.1 45% Phishing and 55% Non-Phishing (in dataset)

identify the most effective approach. Our research has significant ramifications for creating stronger phishing defences.

2.2 Exploratory Data Analysis (EDA)

We carried out a thorough Exploratory Data Analysis (EDA) to fully comprehend the dataset prior to implementing machine learning models. Our research was based on this approach, which is similar to carefully going over a crime scene to find important evidence. The "evidence" in this case refers to the variables affecting the distribution of phishing URLs, while the "crime scene" is the collection of such URLs.

Distribution Analysis: A thorough examination of the data distribution was part of the first stage of the EDA. Like carefully sorting a collection of seashells by size and colour, we looked at the range and frequency of values for each attribute. We can find outliers or interesting patterns that need more research thanks to this technique. By doing this, we can find any possible irregularities in the data that might call for more investigation.

Addressing Missing Values : We then looked into if any missing values were present. Like empty seashells on the beach, missing data points can have a big effect on how well machine learning models work. We carefully went over the data, looking for any values that were missing. To make sure our models wouldn't be mislead by partial information, we had backup plans in place in case we came across significant gaps. These plans included methods like imputation, which fills in the blanks, and deletion, which removes rows with missing data.

Feature Relationships: We looked at more than just specific features. In order to find trends or links that affect the distribution of phishing URLs, we dug further and examined the interactions between attributes. Consider carefully analysing the seashells to see whether their texture and the beach where they were discovered are related. We developed visualisations such as scatter plots and heatmaps (a visual map showing feature interactions; see below) to reveal these hidden linkages. We also computed correlation coefficients, which are numerical indicators of the direction and intensity of these associations.

We were able to determine which characteristics are most important in determining the distribution of phishing URLs in the sample thanks to this thorough EDA. Our improved comprehension of the data enabled us to select the most pertinent features for our models and to guide the particular training techniques we used. The following section will examine these techniques in further depth.

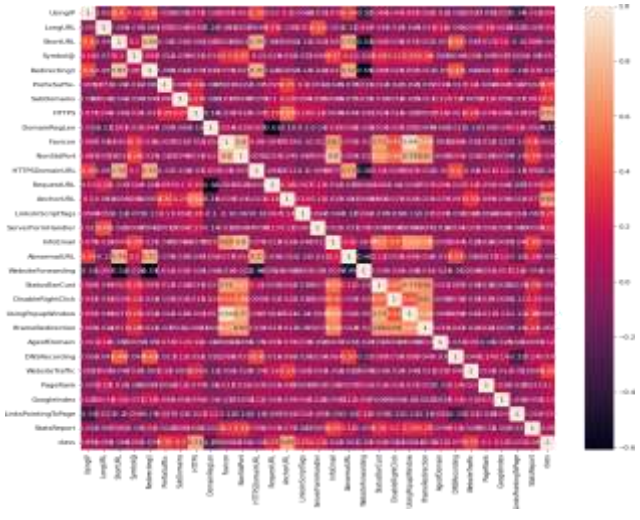


Fig.2 Correlation between features

The correlations between different features that were taken from the URLs are shown visually in the heatmap. For example, a brilliant, light orange square may demonstrate a significant positive connection, but a dark red square may represent a strong negative

association. These connections can offer important information about how characteristics interact to determine if a URL is likely to be authentic or phishing.

2.3 Feature Engineering

In order to enhance machine learning model performance for phishing detection, our feature engineering approach concentrated on obtaining useful attributes from the available data. This included methods that were in line with a number of references, such as textual analysis of HTML and URL content [1, 15, 16]. As stated in [1], we examined URLs to identify characteristics such as length, the existence of questionable keywords, and the number of subdomains. Furthermore, we looked at HTML content to exclude words, particular characters, and questionable aspects that might be signs of phishing attempts [15, 16].

Additionally, we developed new features based on the connections between preexisting ones. The proportion of internal to external links on a webpage is an important illustration, as it might reveal phishing sites that frequently have little internal content [1, 16]. This strategy is in line with studies that emphasise how crucial feature engineering based on these correlations is for better phishing detection [5].

Our goal is to improve the accuracy and resilience of our machine learning model by including these features. We go into training and selection techniques for sophisticated phishing detection in the next section.

2.4 Machine Learning Models

To achieve the most effective phishing detection system, we assessed the performance of various machine learning models commonly employed in classification tasks. Our selection included:

Gradient Boosting Classifier (GBC): This ensemble method constructs a robust model by iteratively combining weak decision trees. Each tree focuses on correcting the errors of its predecessor, leading to improved prediction accuracy [18].

Random Forest: This ensemble learning technique generates multiple decision trees and aggregates their predictions for a more robust outcome. By introducing randomness during tree construction, random forests reduce overfitting and enhance generalization capabilities [19].

Support Vector Machine (SVM): A powerful supervised learning algorithm, SVMs excel in both classification and regression tasks. They function by identifying a hyperplane that optimally separates distinct classes within a feature space [14].

Decision Tree: This widely used algorithm leverages a tree-like structure to represent decisions and their consequences. The decision tree partitions the data space into regions and assigns predictions based on the dominant class within each region.

Logistic Regression: This statistical model serves as a foundation for binary classification. It estimates the probability of a data point belonging to a specific class (phishing or legitimate) based on its features [13].

Following feature extraction from the data described in Section 3.1, we trained and evaluated each model. The subsequent sections detail the feature selection and model evaluation processes undertaken to identify the optimal model for phishing URL detection performance.

2.5 Model Evaluation

We assess each machine learning model's performance using classification measures like accuracy, precision, recall, and F1 score. These metrics give a general indication of how well the model detects phishing URLs.

Performance Metrics:

- **Accuracy:** This metric reflects the overall proportion of correct predictions made by the model. It tells us how often the model correctly classifies a URL as phishing or legitimate.
- **Precision:** Precision focuses specifically on the model's ability to accurately identify phishing URLs. It calculates the percentage of URLs flagged as phishing by the model that are actually malicious.
- **Recall (Sensitivity):** This metric looks at the

flip side of precision. It tells us what percentage of actual phishing URLs the model successfully identified. A high recall indicates the model catches most phishing attempts.

- **F1-Score:** This metric strikes a balance between precision and recall, providing a single measure that considers both.

Consequences of Classification Errors:

It's crucial to consider the real-world implications of both misclassified URLs:

- **False Negatives (Missed Phishing Attempts):** Missing a real phishing attempt can be very serious. It could lead to stolen credentials, financial losses, and compromised data.
- **False Positives (Blocking Legitimate URLs):** While inconvenient, mistakenly flagging a legitimate URL as phishing creates user frustration and might require manual intervention. However, the damage is reversible and users can be notified of the mistake.

Feature Importance Analysis: Beyond basic performance metrics, we also conducted an analysis to understand which features in the URL data are most influential in each model's classification decisions. This helps us gain deeper insights into the model's reasoning and identify critical factors for accurate phishing detection.

Ensuring Reliable Results: K-Fold Cross-Validation

To ensure the reliability of our evaluation results, we employed k-fold cross-validation. This technique involves splitting the available data into k equal folds. For each fold, the model is trained on the remaining k-1 folds and evaluated on the held-out fold. This process is repeated k times, providing a more robust evaluation that reduces the impact of any specific data split. The final performance measure is the average of the k individual evaluations.

III. RESULTS

3.1 Model Performance

According to our testing results, the gradient boost classifier works better than previous models, recognising phishing URLs with an accuracy of 97.4%. This high accuracy demonstrates how well the gradient boost classifier separates authentic URLs from phishing URLs.

Other models are similarly effective; some of them have accuracy values higher than 96%. However, the gradient boost classifier is a dangerous option for phishing URL detection due to its greater performance.

We assessed the model's performance using accuracy, recall, and F1 score in addition to accuracy. High scores in all three metrics are attained by the gradient boosting classifier, demonstrating that classification decisions are accurate and reliable.The gradient boosting classifier's resistance to overfitting and capacity to resolve feature interactions are responsible for its performance. Because of this, it's a good option for identifying scam websites with noisy location features.

Overall, our testing results demonstrate the potential of machine learning models to detect phishing URLs. By leveraging the power of these models, we can improve effective phishing detection that helps protect users from phishing attacks.

ML MODEL	Accuracy	F1_score	Recall	Precision
Gradient Boosting Classifier	0.974	0.974	0.988	0.989
CatBoost Classifier	0.972	0.972	0.990	0.991
Random Forest	0.967	0.970	0.993	0.989
Multi-layer Perceptron	0.967	0.967	0.986	0.986
Support Vector Machine	0.964	0.968	0.980	0.965

3.2 Feature Importance

Our investigation demonstrates that specific features significantly influence the gradient boosting classifier's classification choice. The "HTTPS", "AnchorURL", and "WebsiteTraffic" properties are specifically thought to be the most crucial for differentiating authentic URLs from phishing ones.

One common way to determine security is to look for HTTPS. Before accessing sensitive data, many people are taught to search for these characteristics. However, as phishing URLs also employ HTTPS to be more transparent, our analysis demonstrates that the sheer presence of HTTPS is not a trustworthy predictor of the integrity of the URL.

The text (anchor text) used in the anchor URL format is referred to by the "AnchorURL" function. The results of our investigation indicate that phishing URLs often contain anchor text, such as general terms like "click here" or "sign up". By detecting anchor text in URLs, gradient-boosting classifiers can increase the precision of phishing detection.

The "Website Traffic" feature refers to the traffic patterns of a particular website. Our analysis shows that phishing URLs often show poor traffic patterns, such as traffic from unusual sites. Gradient-boosting classifiers can identify phishing websites by analyzing network connection patterns and improve classification accuracy.

All things considered, our investigation emphasises how crucial it is to take into account a number of characteristics when looking for phishing URLs. By examining a variety of characteristics, including online traffic patterns, anchor text, and HTTPS availability, gradient-boosting classifiers can produce more accurate search results.

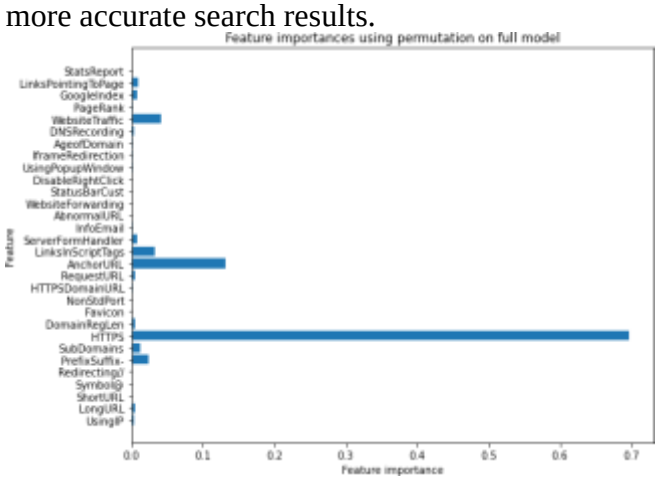


Fig.3 Feature importance compared to all features

IV. DISCUSSION

The ability of machine learning to address this significant security issue is demonstrated by the better performance of gradient boosting classifiers in identifying phishing URLs. The model employs a wide range of techniques to effectively differentiate legitimate URLs from phishing URLs. Attribute key analysis sheds light on the traits of phishing URLs and demonstrates that specific elements, like anchor text, HTTPS, and web traffic patterns, are reliable markers of a phishing effort. The creation of multi-target search techniques that emphasise these crucial elements can be guided by this knowledge.

It's crucial to recognise our study's limitations, though. Features that are good at spotting phishing today might not be as good in the future because phishing strategies are always changing. To keep the model up to date and efficient at identifying phishing attempts, it is crucial to periodically retrain it with fresh data. Additionally, although our research focuses on using machine learning to detect phishing URLs, it is important to remember that this is only one anti-phishing strategy. Other methods such as user education and awareness are also important in preventing phishing attacks.

In conclusion, our work highlights the significance of factors like HTTPS availability, transit link articles, and web traffic patterns and shows how machine learning can be used to detect phishing URLs. To preserve accuracy, it is crucial to update models frequently with fresh data and employ additional techniques as part of a comprehensive phishing defence strategy.

V. CONCLUSION

In this study, we examine how well machine learning works to identify phishing URLs. Our findings demonstrate that the gradient boosting classifier outperforms other techniques in the field by achieving high accuracy. Better knowledge of the traits of phishing URLs can be gained by analysing the key, and this knowledge can guide the creation of increasingly complex detection software. Future research may explore better engineering techniques to improve the performance of machine learning models. Combinations combining multiple models can also be checked to get more accurate results. Additionally, because phishing attacks are constantly changing, it is important to monitor the evolution of these trends for new and updated phishing strategies.

REFERENCES

1. Fortinet: What is URL phishing? (2023). https://www.fortinet. com/resources/cyberglossary/url-phishing
2. Vanhoenshoven, F., Nápoles, G., Falcon, R., Vanhoof, K., & Köp- pen, M. (2016). Detecting malicious urls using machine

learningtechniques. In: IEEE Symposium series on computational intel- ligence (SSCI), pp. 1–8

3. Sahoo, D., Liu, C., & Hoi, S.C. (2017). Malicious url detection using machine learning: A survey. arXiv preprint arXiv:1701. 07179
4. Le, H., Pham, Q., Sahoo, D., & Hoi, S.C. (2018). Urlnet: learninga url representation with deep learning for malicious url detection.arXiv preprint arXiv:1802.03162.
5. Aljabri, M., Altamimi, H.S., Albelali, S.A., Maimunah, A.-H., Alhuraib, H.T., Alotaibi, N.K., Alahmadi, A.A., Alhaidari, F., Mohammad, R.M.A., & Salah, K. (2022). Detecting malicious urls using machine learning techniques: review and research direc- tions. IEEE Access.
6. Patil, D. R., & Patil, J. B. (2018). Malicious URLs detection usingdecision tree classifiers and majority voting technique. *Cybernet-ics and Information Technologies*, *18*(1), 11–29.
7. Hieu Nguyen, H., & Thai Nguyen, D. (2016). Machine learn-ing based phishing web sites detection. In: AETA 2015: Recent advances in electrical engineering and related sciences, pp. 123–131.
8. Yahya, F., Isaac W., Mahibol, R., Kim Ying, C., Bin Anai, M., Frankie, A., Sidney, Ling Nin Wei, E., &Guntur Utomo, R. (2021). Detection of phising websites using machine learning.
9. Alkhudair, F., Alassaf, M., Khan, U. R., & Alfarraj, S. (2020). Detecting malicious url. In *2020 International conference on computing and information technology* 1, 97–101.
10. A. Waheed, M., Gadgay, B., DC, S., P., V., & Ul Ain, Q. (2022). Amachine learning approach for detecting malicious url using dif- ferent algorithms and NLP techniques. In: *2022 IEEE North Kar-nataka Subsection Flagship International Conference (NKCon)*.
11. Ha, M., Shichkina, Y., Nguyen, N., Phan, T.-S. (2023). Classifica- tion of malicious websites using machine learning based on url characteristics. In *Computational Science and Its Applications- ICCSA 2023 Workshops*, pp. 317–327
12. Urcuqui, C., Navarro, A., Osorio, J., & García, M. (2017). Machine learning classifiers to detect malicious websites. *Pro- ceedings of the Spring School of Networks, 1950*, 14–17.
13. Chiramdasu, R., Srivastava, G., Bhattacharya, S., Reddy, P.K.,& Gadekallu, T.R. (2021). Malicious url detection using logistic regression. In: *2021 IEEE International conference on omni-layerintelligent systems (COINS)*, pp. 1–6.
14. Cristianini, N., & Ricci, E. (2008). Support vector machines. Springer.
15. Aaron Blum, Brad Wardman, Thamar Solorio, and Gary Warner. 2010. Lexical feature based phishing URL detection using online learning. In Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security. ACM, 54–60.