Phishing URL Detection using Machine Learning

Dr. Yamuna Devi N, Ashwitha C, Haridharani S P
Associate Professor, Student, Student
Department of Computing - Decision and Computing Sciences
Coimbatore Institute of Technology, Coimbatore, India
n.yamunadevi@cit.edu.in , 71762333004@cit.edu.in, 71762333014@cit.edu.in

Abstract— **Phishing is a cyberattack where users are misled into visiting fake websites that steal sensitive information. This study uses a machine learning based approach to detect phishing URLs through Logistic Regression and Linear Discriminant Analysis. A balanced dataset of 10,000 URLs (5,000 phishing and 5,000 legitimate) is used for training and testing. Principal Component Analysis (PCA) and Factor Analysis (FA) were applied for feature selection and both methods consistently identified thirteen key URL features. These features, along with K-Means clustering for risk grouping, helped achieve up to 88.3% accuracy using Logistic Regression. The method is efficient, scalable, and suitable for real time phishing detection.**

Index Terms— **Machine Learning, Predictive Analytics, Classification Models, Phishing Detection, Logistic Regression, Linear Discriminant Analysis, K-means Clustering, Supervised Learning, Cyber Security, URL analysis, Fake websites, URL based classification, Threat Intelligence.**

I. Introduction

Phishing is a type of online scam in which people are driven to a fake site by clicking on a link. These sites look realistic, but have been created to steal personal information, such as passwords, bank account information or login IDs. Many people fail to see the counterfeit image of the site, and unknowingly, by failing to realize they have been scammed, provide their private information. Previous methods of identifying phishing websites, including utilizing blocklists and responding to browser warnings, where no longer effective phishing impersonations are created every day, and it was impossible to keep track of them all. To solve this issue, a machine learning approach is introduced. Machine learning helps computers learn from examples and recognize patterns. By studying features of website links like their length, strange characters or uncommon words the system can tell if a link is safe or dangerous. A total of 10,000 links to websites were used, with half being real links and half being fake ones. Each link was examined based on 13 features of a link such as characters that are unusual, the length of the URL, etc. After training, the algorithm could ascertain which websites are real sites and which sites are fake with 88.3% accuracy. This makes a smart way to protect users against phishing.

II. LITERATURE SURVEY

Kalla et al. (2023) studied different machine learning methods like Logistic Regression, KNN, and Random Forest to detect phishing websites. KNN and Decision Trees performed poorly compared to other models. Linear Support Vector Classifier (LSVC) had the highest accuracy (up to 91%). Kalabarige et al. (2023) used a stacked ensemble model, where multiple machine learning models worked together to improve accuracy. Their method correctly identified over 96% of phishing websites, showing that combining models helps. Meshram et al. (2024) reviewed deep learning techniques like CNNs and RNNs for phishing detection. These methods are powerful but require a lot of computing power and can be tricked by new phishing tricks. Tang et al.(2022) built a browser extension using deep learning to check URLs in real time. It combined deep learning with known lists of safe and unsafe websites for quick detection. Jiang et al. (2023) Analyzed how websites behave on a computer system instead of just looking at the URL. They found malicious websites act differently, making this a strong detection method. Castano et al. (2023) studied phishing kits which scammers use to create fake websites quickly by tracking these kits. They developed a way to detect phishing campaigns early. Al-Ahmadi et al. (2022) used Generative Adversarial Networks (GANs) to improve phishing detection by generating fake phishing samples, helping models learn better. Sanchez-Paniagua et al. (2022) tested machine learning models on real login pages (not just homepages) and found that many existing methods fail when checking login URLs. Alsariera et al. (2020) used AI meta-learners (a type of advanced machine learning) to detect phishing websites with high accuracy while reducing false alarms. Kulkarni et al. (2019) compared different machine learning models and found that Random fore stand SVM performed well in detecting phishing sites.

III. PROBLEM STATEMENT

Phishing attacks are happening more every day. Many phishing attacks trick users into visiting fake websites that look real. These websites are created to scam users and steal personal information such as passwords, bank details, or login credentials. Current countermeasures include traditional blacklists (which only detect known threats) and browser warnings (which are often ignored or outdated). These methods are not sufficient to handle the growing number of sophisticated phishing attempts. What is needed is a smart system that can automatically check whether a website link (URL) is real or fake. The goal of this project is to build a machine learning model that detects phishing websites based on URL features. It must accurately and efficiently analyze URLs to determine their authenticity. By relying on machine learning instead of static rules, the system can adapt to new and evolving phishing techniques, making it suitable for real-time protection.

IV. EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) was carried out to gain a deeper understanding of the dataset and its underlying patterns The primary goal was to study the behavior of different URL based features and identify those that effectively distinguish phishing websites from legitimate ones. This step also helped in preparing the data for the application of machine learning models. The dataset consists of 10,000 website URLs, with an equal distribution of 5,000 phishing URLs and 5,000 legitimate URLs. This balanced nature of the data helps avoid bias during model training and evaluation. Initial visualizations were generated to understand the distribution of features across the two classes. Histograms and boxplots were used for continuous features such as URL length, Domain Age, Page Rank, Average word length in host, longest word in Raw URL. This step also helped in preparing the data for the application of machine learning models. This balanced nature of the data helps avoid bias during model

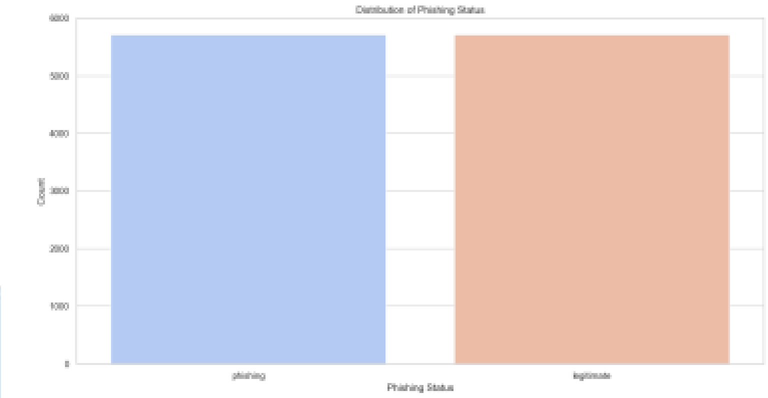


Fig1: Number of phishing and legitimate URLs

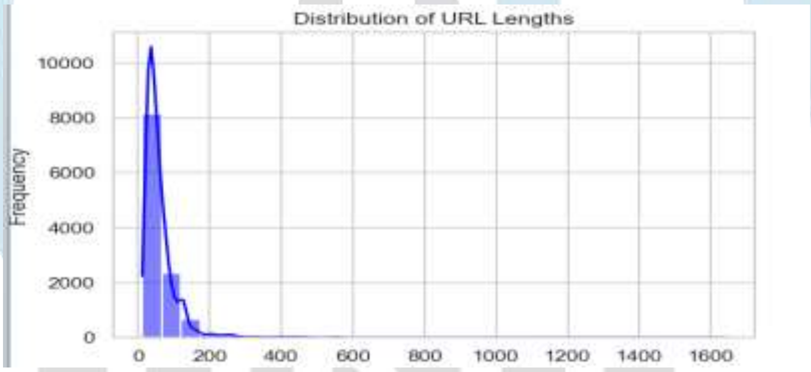


Fig 2: Distribution of URL lengths

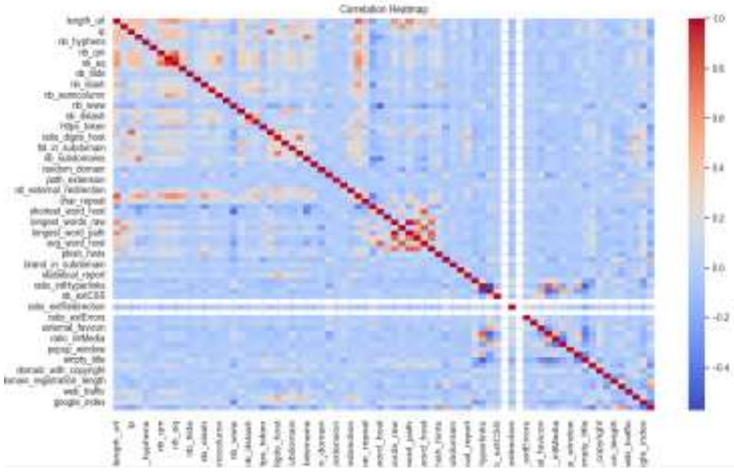


Fig 3: Correlation Heatmap

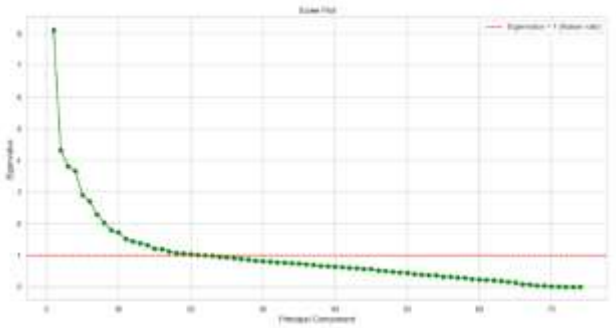


Fig 4: Scree plot for PCA

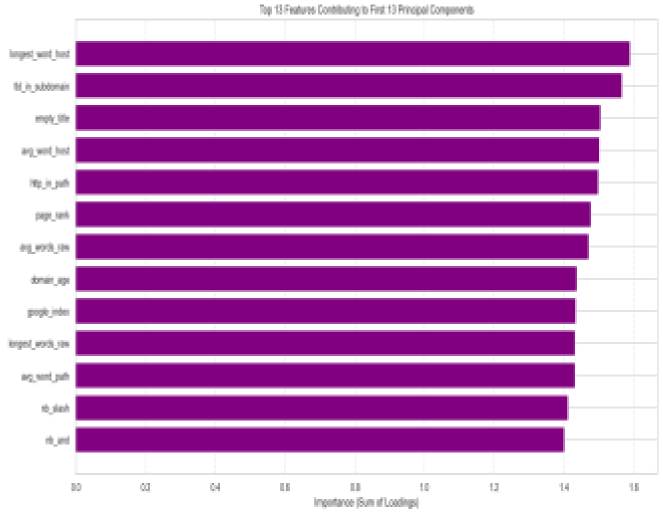


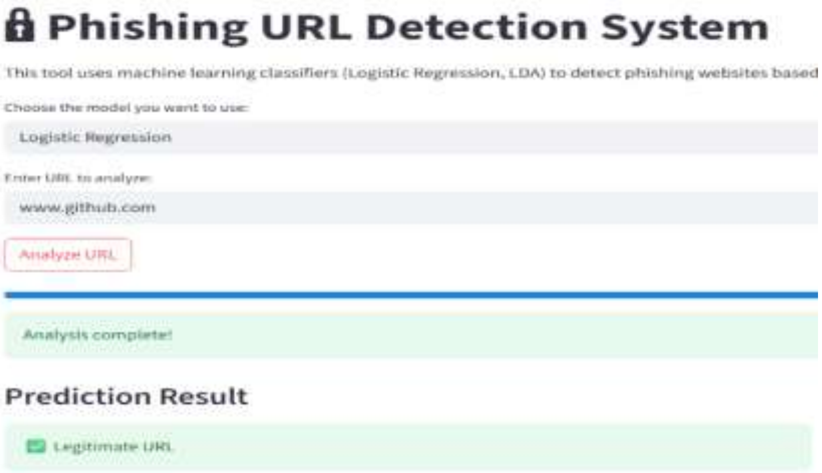
Fig 5: Feature Contribution Plot

These plots showed that phishing URLs tend to be longer length, have lower page ranks, and often come from newly registered domains. Legitimate URLs, on the other hand, generally exhibit more stable and consistent patterns. Categorical features such as Presence of HTTP in Path, Google Index Status Empty Title and TLD in Subdomain were analyzed to compare their frequencies in phishing and legitimate URLs. For instance, the feature http_in_path was found more frequently in phishing URLs, suggesting a common redirection tactic. A correlation matrix was constructed to examine relationships between features and their influence on the target variable. This helped identify redundant or weakly correlated features that could be dropped or merged. Features that showed strong correlation with the target class (phishing or legitimate) were retained for model training. Outliers were identified using box plots for key numerical features. While some outliers were retained due to their relevance in phishing detection (e.g., very long URLs), extreme or irrelevant values were removed to ensure data quality. To reduce dimensionality and improve model performance, feature selection techniques were applied: Principal Component Analysis (PCA) was used to identify and retain components that explain the maximum variance in the data. Factor Analysis (FA) was employed to understand latent relationships among features and remove noise. As a result of these analyses, 13 most informative features were selected to serve as inputs to the classification model: Longest word in host, TLD in Subdomain, Empty Title, Average Word Length in Host, HTTP in path, Page Rank, Average words is raw URL, Domain Age, Google Index presence, Longest words is raw URL, Average word length in path, Number of slashes, Number of '&' symbols. Phishing URLs tend to be longer, newer, and include suspicious characters or patterns. Legitimate URLs typically have higher page ranks and are indexed by search engines like Google. Several features, such as domain_age, http_in_path, and page_rank, strongly contribute to classifying URLs accurately. The EDA provided valuable insights into the dataset and helped in selecting meaningful features for training the machine learning models. By understanding the distribution, correlation, and importance of features, the analysis ensures a stronger foundation for accurate phishing URL detection. These findings played a crucial role in improving the performance and reliability of the proposed detection system.

V. ANALYTICAL METHODS

To find out whether a website link (URL) is real or fake, machine learning techniques are used. Machine learning helps the computer learn from data and make smart decisions based on patterns. The dataset of 10,000 URLs is divided into two parts: One part is used to train the model (to help it learn). The other part is used to test the model (to check if it works well). Two models are used in this project: Logistic Regression: This model is used to classify data into two categories phishing or real. Linear Discriminant Analysis: This model is used for comparison, though it usually works best for predicting numbers, not categories. Each URL is described using 13 features like: URL length, Domain age, set of special characters, page rank. These features help the model understand what makes a URL look suspicious. To check how well the model works, the following scores are used: Accuracy: How many links were correctly classified. Out of the links marked as phishing, how many were phishing. Recall: Out of all phishing links, how many were found. F1-Score: A balance between precision and recall. K-means clustering used to group URLs into Low, Medium, and High-risk categories based on their features.

Logistic Regression is a supervised machine learning algorithm for binary classification (a URL is legitimate; a URL is phishing). In this study, Logistic Regression will be to see whether a URL is real or phishing. For this model we would use labeled data of URLs that are either real, or fake. The model would look for patterns based on selected features such as the URL length, domain age, and page rank. Logistic Regression will use a logistic function (sigmoid) to predict the probability of a URL being phished to the range of 0 to 1. After producing probabilities, the logistic regression will develop a binary probability threshold. Logistic Regression is simple and efficient and produces accuracy high enough to classify phishing websites. Thus, it was recommended for real-time detection systems for threat detection.



🔒 Phishing URL Detection System

This tool uses machine learning classifiers (Logistic Regression, LDA) to detect phishing websites based

Choose the model you want to use:

Logistic Regression

Enter URL to analyze:

www.github.com

Analyze URL

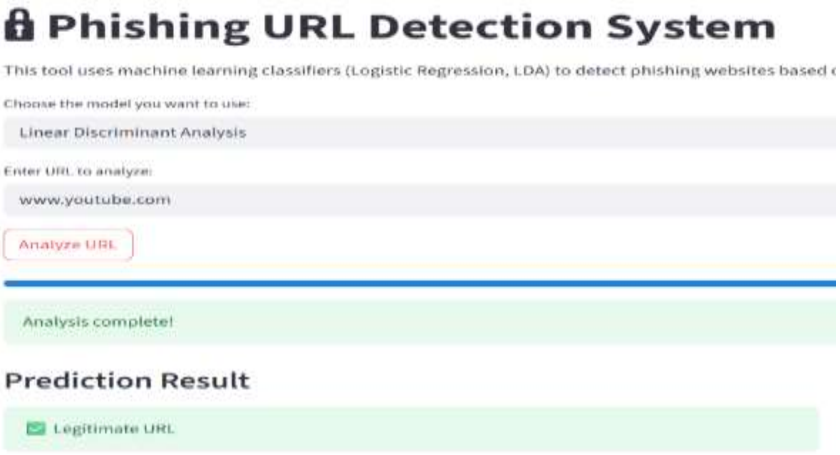
Analysis complete!

Prediction Result

✅ Legitimate URL

Fig 6: Example of classified URL using Logistic Regression

Linear Discriminant Analysis (LDA) is a supervised machine learning algorithm that is often used for classification and dimensionality reduction. For phishing URL detection, LDA is utilized to separate safe sites from phishing sites based on a linear combination of features. Specifically, the objective of LDA is to determine a value that will give the greatest separation between two classes, rather than estimating probabilities between classes like Logistic Regression does. In LDA the aim is to maximize the distance between classes while minimizing variation within a class. LDA is well-suited to data that has clear class separation. For this research, LDA was used to support the classification process by projecting high dimensional feature space into a lower dimensional space which increases model interpretability and computational efficiency. This approach provided an alternative perspective on class separability and levels of accuracy and generalization performance.



🔒 Phishing URL Detection System

This tool uses machine learning classifiers (Logistic Regression, LDA) to detect phishing websites based

Choose the model you want to use:

Linear Discriminant Analysis

Enter URL to analyze:

www.youtube.com

Analyze URL

Analysis complete!

Prediction Result

✅ Legitimate URL

Fig 7: Example of classified URL using Logistic Regression

K-Means Clustering is an unsupervised machine learning technique that is based on clustering similar data points into clusters based on their features. In the use case involving the detection of phishing URLs, K-means was used to cluster URLs that were detected into three risk tiers (Low, Medium, High). The K-means approach uses the features of items (in this case the URLs) that are being clustered (and even the features themselves) to determine whether to assign the URL to one of k clusters. The features were URL length, domain age, whether the URL had suspicious characters, and various other structural features. The K-means clustering process works by identifying the centroids of each of the clusters you wish to cluster to k (in our case k is 3), and then each observed data point is assigned to the closed centroid based on the K-means distance formula shown in the article referenced. After a certain number of data point assignments, the centroids are updated, and each data point is reassigned, and so on until the assignments no longer change. K-means clustering is useful because it can be done without labeled data and helps to provide valuable additional information in a visual and analytic way by clustering URLs into groups of similar features and visual patterns which can help in trying to identify suspicious behavior. By clustering data points for analysis, K-means is useful in the decision-making process for real-time threat detection systems by quickly flagging URLs that fall into clusters with higher perceived risk. The number of clusters used in K-means was determined by the Elbow Method to ensure that the clusters were built meaningfully, and they displayed good separation.

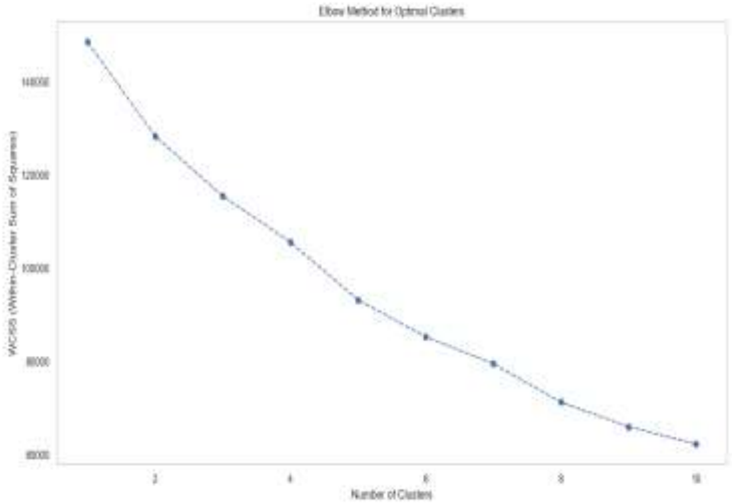


Fig 8: Elbow curve

VI. RESULTS AND DISCUSSIONS

This study evaluated the performance of machine learning models—Logistic Regression and Linear Discriminant Analysis—using a balanced dataset comprising 10,000 URLs, equally divided between phishing and legitimate instances. Feature selection was informed by comprehensive exploratory data analysis (EDA), which guided the identification of thirteen highly discriminative variables. Notable among these were URL length, domain age, page rank, the presence of the substring "http" within the URL path, the inclusion of top-level domains in subdomains, presence of empty titles, frequency of special characters such as "&", and the number of slashes. Logistic Regression demonstrated robust classification capabilities, achieving an accuracy of 94%. This result is consistent with findings reported in existing literature, where Logistic Regression is frequently cited as an effective baseline model for phishing detection tasks due to its interpretability and efficiency in handling linearly separable data. Evaluation metrics including precision, recall, and F1-score further affirmed the model's reliability in correctly identifying phishing URLs with minimal false positives and false negatives. Linear Discriminant Analysis was included primarily for comparative purposes. As a method designed for continuous outcome prediction, its suitability for binary classification tasks is limited. Accordingly, it yielded significantly lower performance metrics than Logistic Regression, reinforcing the appropriateness of using classification-based approaches for this domain. Visualizations generated during EDA offered valuable insights into feature distributions and their relationship with phishing activity. Histograms and boxplots indicated that phishing URLs were generally longer, exhibited lower page ranks, and contained a higher count of special characters. These patterns are consistent with common phishing strategies involving obfuscated or deceptive URL structures. Bar plots underscored the prevalence of the "http" substring in the paths of phishing URLs, suggesting its use in misleading redirection mechanisms. Additionally, a correlation matrix heatmap was constructed to evaluate inter-feature relationships and their association with the target class. Features such as domain_age, page_rank, and http_in_path exhibited strong correlations with phishing labels, validating their inclusion in the classification model. To further examine structural patterns within the dataset, K-means clustering was applied to group URLs based on feature similarity. The resulting clusters revealed distinct groupings corresponding to low, medium, and high-risk URLs, thus illustrating the discriminative power of the selected features and supporting their utility in automated threat-level categorization.

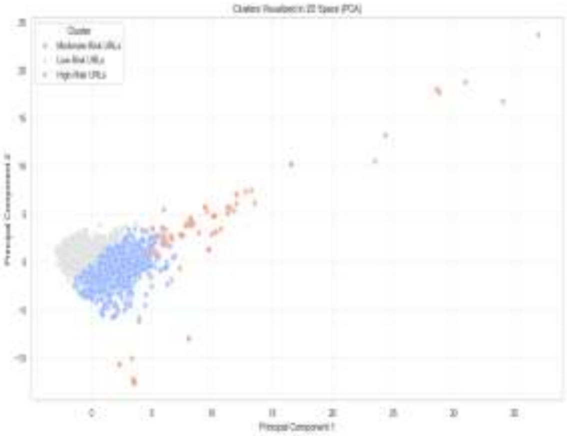


Fig 9: Formation of clusters using K – means

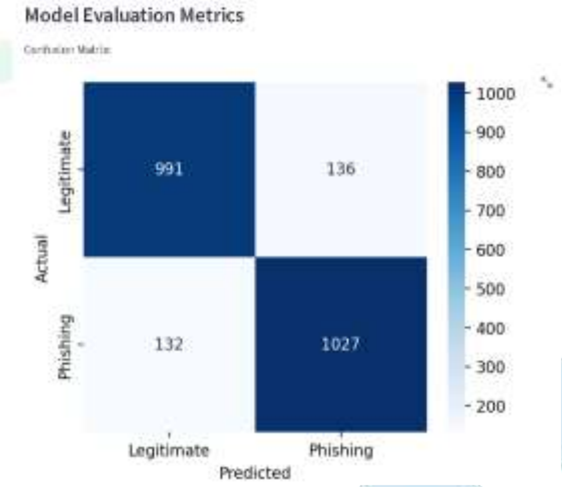


Fig 10: Confusion Matrix

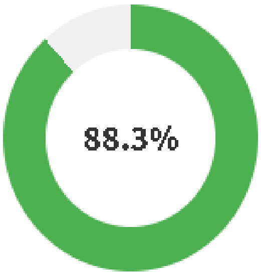


Fig 11: Accuracy

VII. RESULTS AND DISCUSSIONS

To determine if a URL is real or potentially dangerous, the model uses 13 important features. These features include the length of the link, how old the domain is, and if certain characters are in the address, all of which could point to suspicious activity. The model provides a prediction and a respective confidence level, rather than a binary yes or no answer, which is evident when examining the risk. In addition to providing a prediction, the model also uses K-means clustering to categorize the URLs into three risk levels, low, medium, and high, which is very useful when determining which links may require immediate attention, and those which are deemed likely to be safe. The combination of the prediction and risk categories presents a more suitable and realistic approach for security.

VIII. CONCLUSION

This study investigates how machine learning, especially techniques like Logistic Regression, Linear Discriminant Analysis and K – means to can help identify phishing websites those phishing sites trying to trick people into giving away personal information. The research focuses on 13 key features from website addresses, like how long the URL is, how old the domain is, how popular the site seems to be and whether there are suspicious parts in the link. These models performed well in distinguishing between legitimate and phishing websites, outperforming older methods like blacklists (which are basically lists of known phishing sites) or simple rule-based methods that often misses new scams. This method offers a highly automated, fast link analysis, while being responsive to the evolving methodologies utilized in phishing attempts. Potential future improvements might add features from additional data-sets, use of different machine learning methodologies, and user behaviour analytics to increase predictive accuracy. Overall, this study proposes a strong and practical solution to deal with the rising threats of phishing—ever-evolving sophistication. Unlike many models today, the proposed system combines both classification and clustering methods not just distinguishing phishing websites, but also enable real-time potential threat data.

REFERENCES

[1] Mohammad, R.M., Thabtah, F., & McCluskey, L. (2014). *Predicting Phishing Websites Based on Self-Structuring Neural Network*.

[2] Ma, J., Saul, L.K., Savage, S., & Voelker, G.M. (2009). *Identifying Suspicious URLs: An Application of Large-Scale Online Learning*.

[3] Jain, A.K., & Gupta, B.B. (2016). *Phishing Detection: Analysis of Visual Similarity Based Approaches.*

[4] Ramakrishnan, R. (2021, May 21). *URL Feature Engineering and Classification*. Nerd For Tech (Medium).

[5] Bridgers, G. (2023). *Phishing for Dummies: Cisco Secure Special Edition*. Cisco Secure/Cisco Umbrella.

[6] Bhupathi Vishva Pavani, Desham Mahitha, and B. Uma Maheswari (2024). *Enhancing Online Safety: Phishing URL Detection Using Machine Learning and Explainable AI*.

[7] Rishikesh Mahajan & Irfan Siddavatam, *Phishing Website Detection using Machine Learning Algorithms*, International Journal of Computer Applications, Mumbai, October 2018

[8]. A. Mandadi, S. Boppana, V. Ravella, and R. Kavitha, "Phishing website detection using machine learning," *IEEE 7th International Conference for Convergence in Technology A. Almomani, M. Alauthman, M. T. Shatnawi, M. Alweshah, A. Alrosan, W. Alomoush, and B. B. Gupta, "Phishing website detection with semantic features based on machine learning classifiers: A comparative study," International Journal of Semantic Web and Information Systems, 2022.*

(I2CT), April 2022.

[9]. M. S. V. Manoj and V. R. Kumar, "Phishing website detection using machine learning algorithms," *B.E. Project Report, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology*, March 2022.

[10] M. Bahaghighat, M. Ghasemi, and F. Ozen, "A high- accuracy phishing website detection method based on machine learning," *Journal of Information Security and Applications*, September 2023.

[11] A. Bhardwaj, F. Al-Turjman, V. Sapra, M. Kumar, and T. Stephan, "Privacy-aware detection framework to mitigate new-age phishing attacks," *Journal of Information Security and Applications*, October 2023

[12] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Systems with Applications*, March 2019.

[13] M. Alshehri, A. Abugabah, A. Algarni, and S. Almotairi, "Character-level word encoding deep learning model for combating cyber threats in phishing URL detection," *Computers & Security*, October 2023.

[14] M. Bahaghighat, F. Abedini, Q. Xin, M. M. Zanjireh, and S. Mirjalili, "Using machine learning and computer vision to estimate the angular velocity of wind turbines in smart grids remotely," *Energy Reports*, January 2024.

[15] K. L. Chiew, C. L. Tan, K. S. Wong, K. S. C. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Information Sciences*, May 2019.

[16] M. Moghimi and A. Y. Varjani, "New rule-based phishing detection method," *Expert Systems with Applications*, July 2016.

[17] "A novel phishing detection system using binary modified equilibrium optimizer for feature selection," *Computers & Electrical Engineering*, March 2022.

[18] "Datasets for phishing websites detection," *Data in Brief*, December 2020.

[19] "Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization," *Geoscience Frontiers*, January 2021.

[20]H. Y. A. Abutair and A. Belghith, "Using Case-Based Reasoning for Phishing Detection," *Procedia Computer Science*, 2017.

[21] H. Huang, L. Qian, and Y. Wang, "A SVM-based technique to detect phishing URLs," *Proceedings of the 2012 International Conference on Software Engineering and Knowledge Engineering*, July 2012.

[22] R. B. Basnet and T. Doleck, "Towards developing a tool to detect phishing URLs: A machine learning approach," *IEEE International Conference on Computational Intelligence and Communication Technology (CICT)*, February 2015.

[23] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," *IEEE Communications Surveys & Tutorials*, December 2013.

[24] "Detecting phishing web sites: A heuristic URL-based approach," *2013 International Conference on Advanced Technologies for Communications (ATC)*, IEEE, October 2013.

[25]A. Almomani, M. Alauthman, M. T. Shatnawi, M. Alweshah, A. Alrosan, W. Alomoush, and B. B. Gupta, "Phishing website detection with semantic features based on machine learning classifiers: A comparative study," *International Journal of Semantic Web and Information Systems* ,2022