

Detection of Phishing Websites using Machine Learning Algorithm

Richa Yadav¹, Ritika Goyal² and Adhya Mittal³

¹⁻³Department of Electronics and Communication Engg., Indira Gandhi Delhi Technical University For Women Delhi, India.
Email: richayadav@igdtuw.ac.in, ritikagoyal036@gmail.com, Mittal.adhya2001@gmail.com

Abstract—The Internet serves as a global crossroads, enabling users to connect and share information. Unfortunately, this connectivity has become a breeding ground for malicious activities, with phishing attacks emerging as a significant threat. Phishing attacks represent a severe danger to online accounts and data due to their deceptive nature. The average person now finds it increasingly challenging to differentiate between a legitimate email, link, or website and a fraudulent one. It is against this backdrop that the research is initiated. The research paper introduces an effective machine learning algorithm designed for the identification of both phishing and legitimate URLs. By integrating logistic regression and pipeline, the algorithm is adept at processing a sizable dataset comprising 549,347 legitimate and phishing URLs. Notably, the experimental results underscore the algorithm's superior accuracy in distinguishing phishing URLs compared to previous research efforts. This emphasis on leveraging a substantial dataset and achieving enhanced accuracy sets the proposed technique apart in the field.

I. INTRODUCTION

In the vast landscape of the internet, where users connect and exchange information, a looming threat has emerged in the form of phishing attacks. Phishing, characterized by its deceptive tactics, represents a substantial danger to online accounts and data. Malicious actors, particularly phishers, strategically exploit the internet's global reach to conduct extensive attacks. Their modus operandi involves infecting computers with spyware, leading users to deceptive websites designed to extract sensitive information.

The inherent risk lies in the deceptive nature of phishing attacks, where cybercriminals masquerade as trustworthy entities or individuals. Through sophisticated social engineering techniques, victims are manipulated into unwittingly disclosing confidential information, resulting in financial loss or compromised security. The internet, initially conceived as a crossroads for global connectivity, has inadvertently become a breeding ground for these malicious activities.

As phishing attacks evolve into more sophisticated forms, differentiating between legitimate and fraudulent content has become an increasingly challenging task for the average person. Cybercriminals exploit this confusion, leading to the widespread success of phishing schemes. The escalating prevalence of cyber-attacks underscores the urgent need for robust countermeasures.

In response to these challenges, a research project is initiated, specifically addressing the proliferation of fake URLs and domain names associated with phishing attacks. The primary objective is to develop effective strategies for identifying and detecting phishing website links. With a focus on the challenges posed by fake URLs and domain names, the research aims to significantly enhance the overall efficiency of phishing detection.

By concentrating efforts on this critical aspect, the goal is to empower individuals to navigate the online landscape with increased confidence, therefore reducing the likelihood of succumbing to phishing attacks. Acknowledging the escalating complexity of phishing schemes and the hurdles presented by fake URLs and domain names, the research project adopts a proactive approach to enhance detection methods. Ultimately, the aspiration is to contribute targeted solutions that fortify overall cyber security, making the internet a safer space for users amid the evolving threat landscape.

TABLE I: OVERVIEW OF REVIEW PAPERS FOCUSING ON SYSTEMS FOR DETECTING PHISHING INCIDENTS

Applied Approach	Used Algorithm	Used Data Set	Main Findings	Limitations/Challenges
Visual Similarity, Heuristic & Machine Learning(2019) [1]	Decision Tree, K-Nearest Neighbor, AdaBoost Gradient Boost, Logistic Regression, Random Forest, XGBoost	The UCI Machine Learning Repository contains 21,055 individual instances, while Mendeley also offers a collection of data.	The research achieved a 97.51% accuracy employing a UCI dataset and Mendeley's phishing dataset for machine learning, yielding an accuracy of 98.45%.	Dataset 1 (D1) contains a total of 11,055 instances, while Dataset 2 (D2) comprises 10,000 instances.
Heuristic & Machine Learning (2019) [2]	XGBoost, Random Forest , Logistic Regression, K-Nearest Neighbor, Support Vector Machine.	Common Crawl Alexa Phish Tank	The Random Forest yielded an accuracy rate of 94.26% in its predictions.	The proposed model's accuracy score slightly trails behind that of other models documented in the literature. It might struggle in detecting phishing URLs employing shortened links.
Heuristic & Machine Learning(20 21 [3]	Support Vector Machine, Random Forest, K-Nearest Neighbor, Logistic Regression	ISCXURL- 201611,964 instances	Using the Random Forest algorithm, the research achieved a peak Accuracy of 99.57%.	A research endeavor hasn't employed distinct datasets for training and testing to verify the resilience of the suggested methodology. Challenges: It has achieved significant Accuracy using only nine URL features.
Heuristic & Machine Learning(20 20) [4]	Random Forest	Phish Tank 11,055 instances 30 features	This research attained a detection accuracy of 99.36% in identifying phishing websites in real-time.	The study exclusively employed a dataset containing 11,055 instances and tested a single algorithm as part of its experimentation.
Machine Learning (2018) [5]	Support Vector Machine	Used 2,141 phishing pages	With the SVM algorithm, the study obtained the Accuracy of 95.66%.	It produced results for limited dataset
Machine Learning (2022) [6]	Random Forest, XGBoost Algorithm	500 fake and 500 legitimate web pages	The authors achieved a 99.95 accuracy using XGBoost Algorithm and obtained the Accuracy of 99.57% with Random Forest algorithm	The data set used in this work contains only 500 instances or the dataset was limited

II. PROBLEM STATEMENT

Phishing attacks have become more intricate, making it challenging for the average person to distinguish between legitimate and fraudulent email links or websites. These schemes, employed by cybercriminals, are highly prevalent and successful. Therefore, this project aims to tackle fake URLs and domain names by enhancing the identification of phishing website links, seeking a more effective detection method.

The pressing need for a proficient solution has identified the detection of phishing as a prominent area of research. Various approaches, such as visual similarity, blacklisting, white listing, and analyzing website content and URL features, have been explored. However, the visual similarity approach, which extracts website features for phishing identification, faces limitations due to potential misrepresentations in a page's content, leading to inaccurate classification.

This paper aims to overcome the drawbacks observed in prior research by presenting a thorough method for detecting phishing that utilizes machine learning.

III. RELATED WORK

Several papers have delved into understanding and tackling phishing attacks, exploring strategies employed by attackers. Among these, specific studies stand out for their high accuracy in detecting such attacks.

A.V. Ramana et al. introduced a novel approach employing various algorithms like Decision Tree, K-Nearest Neighbor, AdaBoost, Gradient Boost, Logistic Regression, Random Forest, and XGBoost. Their experiment achieved a notable 97.51% accuracy using a dataset comprising 21,055 instances from UCI and Mendeley's phishing dataset for Machine Learning, ultimately reaching an accuracy of 98.45%.

R.S. Rao et al. proposed a classification model for phishing attacks, utilizing algorithms such as XGBoost, Random Forest, Logistic Regression, K-Nearest Neighbor, Support Vector Machine, and others. They accomplished an accuracy of 94.26% specifically with Random Forest. The newly suggested model exhibits a marginally reduced accuracy compared to the other models in the existing literature. It might struggle in identifying phishing URLs that employ shortened links.

B.B. Gupta et al. presented an approach focusing on detecting phishing email attacks, employing Support Vector Machine, Random Forest, K-Nearest Neighbor, and Logistic Regression. Their analysis, conducted on a dataset comprising 11,964 instances, achieved a peak accuracy of 99.57% employing the Random Forest algorithm.

Similarly, M.G. Hr et al. utilized Random Forest exclusively to detect phishing links, achieving a real-time detection accuracy of 99.36% for phishing websites. The utilized dataset comprises 11,055 instances, and the experiment focused exclusively on a single algorithm.

Higashino, M., et al. proposed a model leveraging Support Vector Machine to identify phishing sites using a dataset of 2,141 pages, obtaining an accuracy of 95.66%. However, this study's results were limited due to the smaller dataset size.

R. Alabdan et al. presented an approach focusing on detecting phishing email attacks, employing Random Forest, XGBoost Algorithm. Their study, based on a dataset of 500 fake and 500 legitimate web pages, yielded the highest accuracy of 99.95% using the XGBoost Algorithm. The data set utilized in this research consists of only 500 instances, signifying a limited dataset.

IV. PROPOSED WORK

A. Data Collection

The collection of phishing and legitimate URLs for this study involves utilizing the open-source services Kaggle, Phistank, and Phistrom. These platforms offer datasets in various formats, including CSV and JSON. The process entails combining data from these sources, resulting in a comprehensive dataset. From this amalgamated dataset, a random selection of 549,347 URLs comprising both phishing and legitimate instances is gathered. These URLs are meticulously labeled as 'Good' and 'Bad' to facilitate the training of machine learning models. This approach ensures a diverse and representative set of URLs for model training, enhancing the robustness and generalizability of the developed models. The inclusion of various formats and sources contributes to the richness of the dataset, reflecting the diverse landscape of URLs encountered in real-world scenarios.

B. Preprocessing

Preprocessing is a critical step in the data mining process, involving the manipulation or removal of data to improve performance. In the context of training data for machine learning, basic Natural Language Processing (NLP) techniques were employed. The use of a RegexpTokenizer facilitated the tokenization of words, breaking down the textual data into individual components. Additionally, the application of CountVectorizer transformed the text corpus into a numerical representation, generating a vector of term/token counts. This transformation is pivotal as it converts the textual information into a format suitable for machine learning algorithms, allowing for the effective training of models. The integration of these NLP techniques ensures that the data is appropriately structured and prepared for subsequent stages of model development, contributing to the overall efficiency and accuracy of the machine learning process.

C. Models & Training

In the process of model development and training, an essential initial step includes the division of the dataset into training and testing subsets. This partition is essential for assessing the model's performance accurately. Given the nature of the task at hand, it is evident that this is a supervised machine learning problem. Supervised learning encompasses two primary categories: classification and regression.

In the context of this study, the problem is framed as a regression task. The objective is to classify URLs as either 'good' or 'bad.' The supervised machine learning models selected for training, both falling under the regression category, include Logistic Regression.

The dataset is utilized to train these regression models rigorously. Subsequently, the models undergo evaluation using a separate test dataset. This evaluation phase allows for a comprehensive assessment of the models' ability to generalize to new, unseen data, ensuring their effectiveness in practical applications.

Once the models are successfully trained and evaluated, the next step involves storing the learned parameters and configurations in a .pkl file. This file serves as a repository for the trained data. In practical terms, this stored information can be later utilized to predict the legitimacy of URLs in real-time. By loading the .pkl file, the trained models can quickly assess and categorize URLs as either legitimate or phishing, providing a valuable tool for proactive cyber security measures.

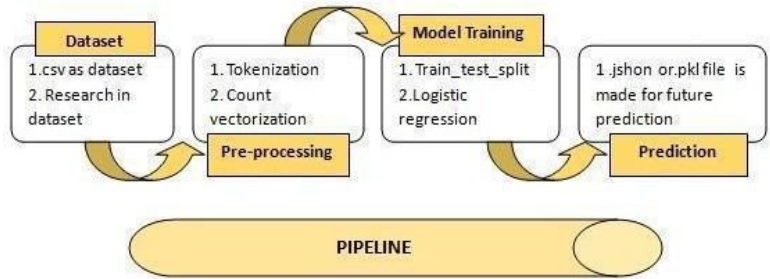


Fig 1. Proposed Technique

V. RESULT

The performance of our applied models is assessed through measures of accuracy and cross-validation scores. Confusion Matrix is used to summarize the performance of the algorithm. A higher Accuracy is achieved using Logistics Regression and Pipeline on a huge dataset of 5,49,347 phishing and legitimate URL to check if a URL is phishing or legitimate.

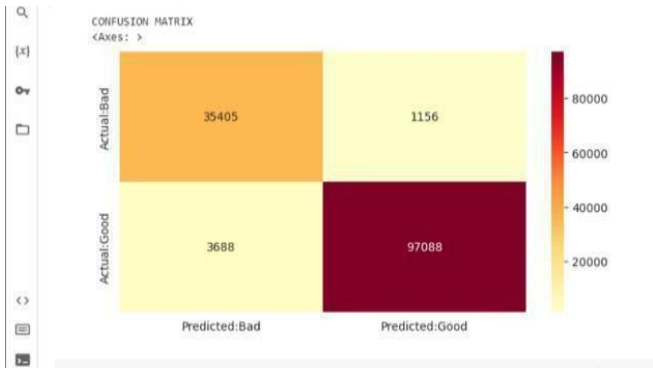


Fig 2. Logistic Regression

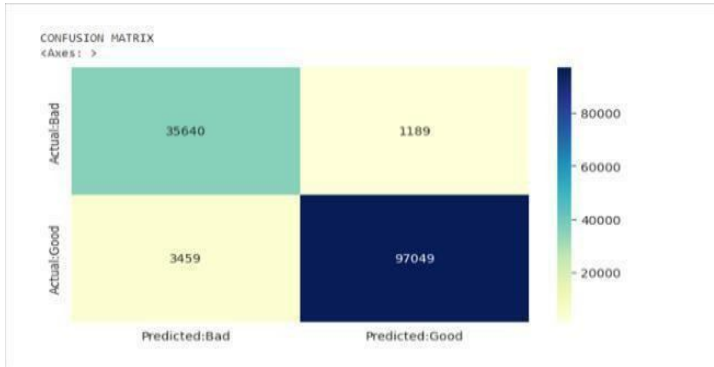


Fig 3. Pipeline

TABLE II: SUMMARY OF RESULTS

Classifier	Dataset Used	Training Accuracy Percentage	Testing Accuracy Percentage
Logistics	549347 legitimate and phishing urls	97.92	96.47
Pipeline	549347 legitimate and phishing urls	98.07	96.61

VI. CONCLUSION AND FUTURE SCOPE

In conclusion, this research significantly contributes to the field of cyber security by providing a valuable resource for both individuals and organizations to identify and comprehend phishing techniques employed by malicious actors. The primary focus is on enhancing the detection of phishing URL attacks through the adoption of machine learning models, representing a notable departure from traditional detection methods.

The aim of this paper is to empower both individuals and organizations to navigate the online landscape with increased confidence. By understanding and being able to identify phishing techniques, users can take proactive measures to safeguard their sensitive information. Moreover, the adoption of machine learning models introduces a dynamic and responsive layer of defense against the ever-changing landscape of phishing attacks.

The Proposed method is currently under developmental phase and more machine learning algorithms will be further used on the same huge dataset to increase its efficiency of detecting phishing links. It can also be further extended to creation of browser extension (for chrome, brave, Firefox etc.) and utilizing neural network and deep learning techniques can aid users in achieving more precise detection of Phishing URLs.

REFERENCES

- [1] A.V. Ramana, K.L. Rao, R.S. Rao Stop-Phish: introduces an advanced approach for identifying phishing attempts by employing feature selection ensemble Social Network Anal. Mining, 11 (1) (2021), pp. 1-9,
- [2] R.S. Rao, T. Vaishnavi, A.R. Pais CatchPhish: detection of phishing websites by inspecting URLs J. Ambient Intell. Humanized Comput., 11 (2) (2019), pp. 813-825, 10.1007/s12652-019-01311-4
- [3] B.B. Gupta, K. Yadav, I. Razzak, K. Psannis, A. Castiglione, X. ChangAn innovative method for real-time detection of phishing URLs using machine learning based on lexical analysis Comput. Commun., 175 (April) (2021), pp. 47-57,
- [4] M.G. Hr, A. Mv, S. Gunesh Prasad, S. Vinay Development of anti-phishing browser based on random forest and rule of extraction framework Cybersecurity, 3 (1) (2020), pp. 1-14,
- [5] Higashino, M., et al. An Anti-phishing Training System for Security Awareness and Education Considering Prevention of Information Leakage. in 2019 5th International Conference on Information Management (ICIM). 2019.
- [6] R. Alabdan, "future internet Phishing Attacks Survey: Types, Vectors, and Technical Approaches", doi: 10.3390/fi12100168
- [7] E. S. Gualberto, R. T. de Sousa, T. P. B. de Vieira, J. P. C. L. da Costa, and C. G. Duque, "From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection," IEEE Access, vol. 8, pp. 76368–76385, 2020, doi:10.1109/ACCESS.2020.2989126.
- [8] L. Gallo, A. Maiello, A. Botta, and G. Ventre, "2 Years in the anti-phishing group of a large company," Computers and Security, vol. 105, Jun. 2021, doi: 10.1016/j.cose.2021.10225.