Journal of Information Technology and Digital World (ISSN: 2582-418X)
www.irojournals.com/itdw/

Masked and Phishing URL Detection using Machine Learning

Sukant S[1], Sujitha R[2], Nithish T[3], Nikitha M[4]

[1,3,4]Information Technology, Bannari Amman Institute of Technology, Sathyamangalam, India

[2]Electronics and Communication, Bannari Amman Institute of Technology, Sathyamangalam, India

[4]Assistant Professor, Bannari Amman Institute of Technology, Sathyamangalam, India

E-mail: [1]sukant.it20@bitsathy.ac.in, [2]sujithar.ec20@bitsathy.ac.in, [3]nithish.it20@bitsathy.ac.in, [4]nikitham@bitsathy.ac.in

Abstract

With the escalating threats in the digital landscape of cybersecurity, the rapid and widespread of masked and phishing URLs poses a significant threat to online users. Detecting these malicious URLs is a paramount concern to safeguard sensitive information and prevent unauthorized access. This study delves into the application of machine learning algorithms for the accurate identification of masked and phishing URLs. Specifically, Decision Tree, Random Forest, and XGBoost algorithms are employed to create predictive models capable of distinguishing between legitimate and malicious URLs. The research involves the collection of a comprehensive dataset comprising both legitimate and various forms of malicious URLs. Feature engineering techniques are applied to extract relevant information from the URLs, transforming them into numerical representations suitable for machine learning. The three selected algorithms are individually trained and finetuned using the dataset, exploiting their unique capabilities to distinguish patterns indicative of phishing attempts and masked URLs. The performance of each model is evaluated using metrics such as accuracy, precision, recall, and web traffic. This study examines the application of machine learning algorithms to identify masked and phishing URLs. By comparing the results of these algorithms, a predictive model capable of distinguishing between legitimate and malicious URLs is created. Experimental results showed promising accuracy rates and potential to contribute to online security efforts.

DOI: https://doi.org/10.36548/jitdw.2023.4.002
Received: 11-10-2023. received in revised form: 03-11-2023, accepted: 17-11-2023, published: 30-11-2023

The implications of this research extend to advanced cybersecurity systems, offering enhanced protection against evolving threats in the digital domain.

Keywords: Masking URL, Phishing, Cybersecurity, Social Engineering, Machine Learning

1. Introduction

In today's interconnected world the internet has completely transformed the business conduct. The are numerous of fraudulent URLs that pretend and function like genuine websites nowadays. These fraudulent URLs are specifically designed to trick the users and gather their private/secret information's. Clicking on these links can expose individuals and organizations to cyber threats like identity theft, financial scams and data breaches. Conventional methods of detecting URLs like blacklisting or relying on signatures have struggled to keep up with the evolving tactics of cybercriminals. Therefore, there is a need for advanced and adaptable solutions that can effectively identify and combat these harmful URLs. Machine learning technology has emerged as an approach in addressing this challenge by improving URL detection capabilities. This presentation gives an outline of the issue presented by URLs and makes way for an exhaustive investigation of how AI can be utilized to distinguish them. It also emphasizes the importance of this research, in today's cybersecurity landscape. Recent studies have shown that machine learning holds potential in enhancing the accuracy and efficiency of URL classification. Machine learning models have the capability to distinguish between malicious links by examining aspects of URLs. This enables them to provide real time protection, against phishing attacks. This paper presents a systematic approach to address this challenge, leveraging the power of artificial intelligence to bolster cybersecurity defenses. In the following sections, we will delve into the methodology, experimental results, and implications of our approach, drawing insights from reference papers that have contributed to the development of advanced techniques for masked and phishing URL detection. These references, highlight the significance of machine learning in countering these threats and emphasize the potential of the proposed approach to mitigate the risks associated with deceptive URLs.

2. Related Work

A mix methodology with new elements to supervise new high level and phishing attacks was presented [1]. To assess the framework, a condition of inventive explanation coordinated decision tree learning models were utilized. This research has taken in a mutt system, for example a mix of static and dynamic strategy for the disclosure of harmful URLs. four sorts of static and dynamic URL details like, URL area name, page source parts and short URLs were eliminated. The short URLs are secluded by checking the space names containing the basic URL shortening associations like bit.ly, goo.gl, tinyurl.com, owl.ly, deck.ly, su.pr and bit.do. The decision tree calculation settles on a decision making of the decision tree for the given dataset by recursive separating information. The choice is made utilizing the first structure. The assessment considers each of the potential tests that can segment the instructive record and picks a test that gives the best data gain. In order to anticipate phishing websites, this study evaluated the outcomes of several machine learning techniques. It discovered that the majority of phishing attacks share certain traits that can be recognised by machine learning techniques. Babu Rao Pawar et al [2] has given a comprehensive analysis of the literature and proposed a novel approach that uses feature extraction and a machine learning algorithm to identify phishing websites. The purpose of the study was to build deep neural networks and machine learning models to detect phishing websites using the dataset that was gathered. Patil, Dharmaraj R., et al [3] offered a novel, efficient hybrid approach with additional features to address this issue by employing cutting-edge models for supervised decision tree learning classifications. The findings of the experiment demonstrate that all decision tree learning classifiers perform well on the labelled dataset when new features are added, with 98–99% detection accuracy and extremely low False Positive Rate (FPR) and False Negative Rate (FNR). Fazal et al [4] using the confusion matrix, has illustrated the accuracy of the decision tree classifier in identifying phishing websites. It scores 95.97% correct and has a high true positive rate and a minimal false positive rate. Selvakumari, M et al [5] presents the review of the most advanced methods for phishing detection utilising various models in both conventional and deep learning algorithms. Additionally, a comparison examination of detecting utilising various approaches has been recognised as the solution for identifying phishing websites. This document highlights the methods that several researchers have suggested. After that, a critique of the techniques' shortcomings is given. Salloum et al [6] has made efforts to thoroughly evaluate and synthesise the literature on the application of natural

language processing (NLP) to the detection of phishing emails, this study discovers that the primary focus of phishing detection research is feature extraction and selection, with techniques for categorising and improving phishing mail detection. M I, Shilpa. et al [7] used the best classifier based on its good performance in classifying URLs as benign or dangerous after several classifiers were applied to detect URLs and various activities were carried out.Khonji et al [8] In order to show how phishing detection strategies fit into the larger mitigation process, a high-level overview of the many types of phishing mitigation approaches is provided. These categories include detection, offensive defence, rectification, and prevention. Aggarwal et al [9] states that PhishAri is far faster than both Twitter's own security mechanism and public blacklists at identifying harmful content. It can identify phishing tweets at zero hour with high accuracy. Singh et al [10] presents a review that enhances readers' awareness of phishing attempts, how to spot them, and promotes them to practise phisher prevention. TRAGHA, et al [11] different machine learning algorithms used to categorise webpages are reviewed, the characteristics of webpage classification are presented, and a literature review is created by compiling and analysing all sources relevant to web page classification that were automatically crawled from the ScienceDirect and Springer websites. Vanhoenshoven, et al [12] in his research examines the binary classification problem of detecting malicious URLs and evaluates the effectiveness of various popular classifiers, including Naïve Bayes, Support Vector Machines, Multi-Layer Perceptrons, Decision Trees, Random Forest, and k-Nearest Neighbours.

3. Proposed Work

This proposed methodology aims to develop an accurate and reliable model for detecting masked URLs as well as phishing URLs using the comparison of the accuracy of three algorithms namely XGBoost, decision tree, and a random forest, contributing to the field of cybersecurity, and providing valuable insights for public and security engineers. The general steps involved in identifying the malicious URL using the machine learning technique is as follows Data Collection: diverse and extensive dataset of URLs are gathered, including both legitimate and malicious examples that covers a wide range of URL characteristics, including variations in structure, content, and domain names Feature Extraction: relevant features are extracted from the collected URLs. These features may include:

- URL structure like length, depth, presence of subdomains, etc. Lexical analysis: Keywords, special characters, and patterns.
- Domain information like domain age, registrar, IP address, SSL certificate status, etc.
- Content-related features like page content analysis, redirection, etc. and finally the
- DNS records like MX records, SPF records, etc.

The dataset is prepared for the machine learning models by applying the appropriate preprocessing and normalization techniques. Labelling: The dataset is annotated by labelling the URLS as legitimate, phishing, or masked. The known ground truth of the data is used for labelling and ensuring a balanced representation of each class in the dataset. Data Splitting: the dataset is partitioned for training, validation, and test sets. Model Selection: the different machine learning algorithms or classification, such as: Random Forest, XG Boosting and Decision Tree models are selected. The selected models are trained to classify URLs as legitimate, phishing, or masked. Finally, the machine learning models are evaluated to assess the accuracy in training and testing. This proposed methodology outlines a systematic approach to detecting masked and phishing URLs using machine learning. It emphasizes the importance of data collection, feature engineering, model selection, and real-time monitoring to create a robust defense against evolving cyber threats. Continuous improvement and adaptation are key to maintaining the effectiveness of such a system in a dynamic cybersecurity landscape. The figure. 1 below show the stages involved in the masked and phishing URL detection using machine learning.

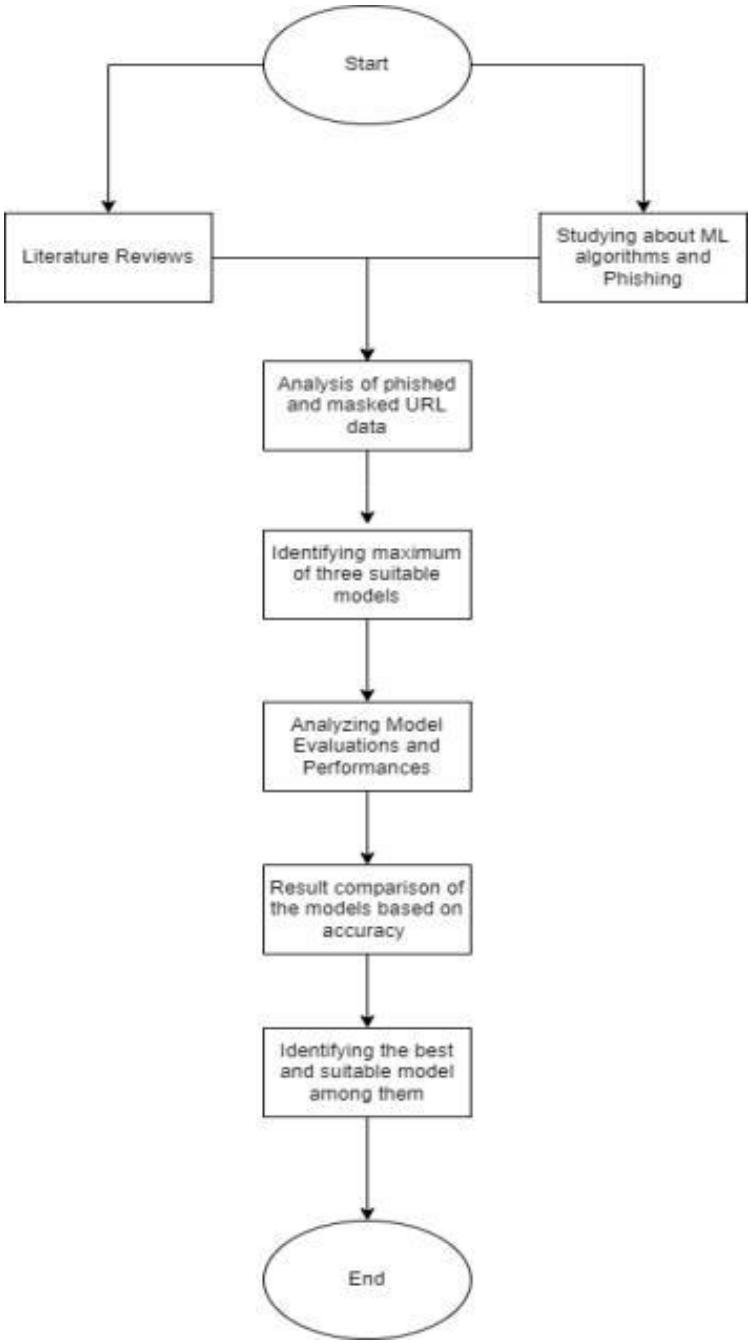

Figure 1. General Flow Diagram of the Masked and Phishing URL Detection

3.1. Implementation

The analysis is carried out in the aim to develop an accurate and reliable model for detecting masked URLs as well as phishing URLs using the comparison of the accuracy of three algorithms XGBoost, Decision Tree, and Random Forest, these machine learning models were selected because of their efficiency in managing classification assignments and their

capacity to identify intricate links within the data, contributing to the field of cybersecurity. The selected machine learning models are implemented using Python and machine learning libraries, such as scikit-learn. The implementation starts by loading the required modules. The Web page Phishing Detection dataset from Kaggle with total of 11430 URLs with 87 extracted features is utilized in the proposed analysis The dataset is intended to serve as industry standards for phishing detection systems that rely on machine learning. Three distinct groups of features are present: twenty-six are derived from the structure and syntax of URLs, twenty-six are derived from the content of the pages that correspond to them, and seven are derived through external service queries. The dataset is evenly distributed, with precisely 50% phishing and 50% genuine URLs.[13]. The dataset is loaded using the pandas library in .csv format and the NumPy array. The collected dataset is pre-processed and normalized applying the MinMaxScaler. The model classifier Random Forest, XGboost and the decision tree is imported using the scikit-learn library. The dataset is split for training and testing purposes as 80% and 20 % respectively. As the dataset contains Boolean data it was easy to work with the decision tree, random forest and XGBoost. The figure. 2 below presents the feature importance observed for Random Forest, XGBoost and Decision Tree. The hyper parameters such as the max_depth and the learning_rate is tunned applying grid search along with K-Fold cross validation, with number of folds = 5, the data are randomised by shuffling the data and the reproducibility is ensured with a fixed random seed.

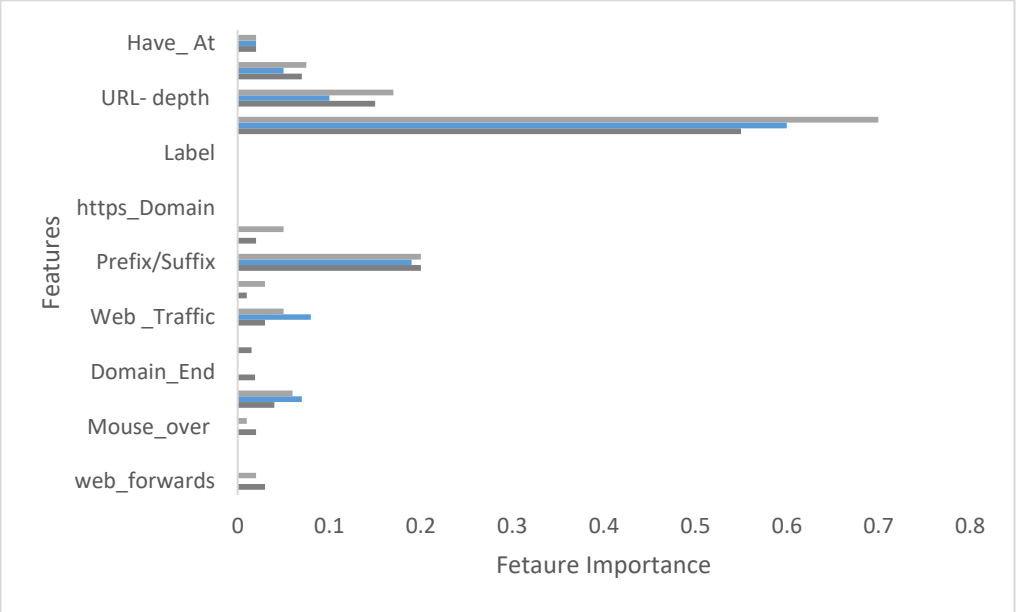

Figure 2. Feature Importance

4. Results and Discussion

The model is evaluated by testing its performance on the test data. The model was trained for 50 Epochs, the testing dataset are used to assess the performance of the models on an unseen dataset. The performance of model is compared using the evaluation metrics like accuracy, precision, recall and f1 score. The table.1 below shows the accuracy observed for the three machine models in the training and the testing.

	ML Model	Train Accuracy	Test Accuracy
2	XGBoost	0.866	0.868
1	Random Forest	0.815	0.821
0	Decision Tree	0.812	0.819

Table 1. Training and Testing Accuracy

The figure.3 below shows the graphical representation of the performance scores observed the three models in terms F1 score, precision and recall.

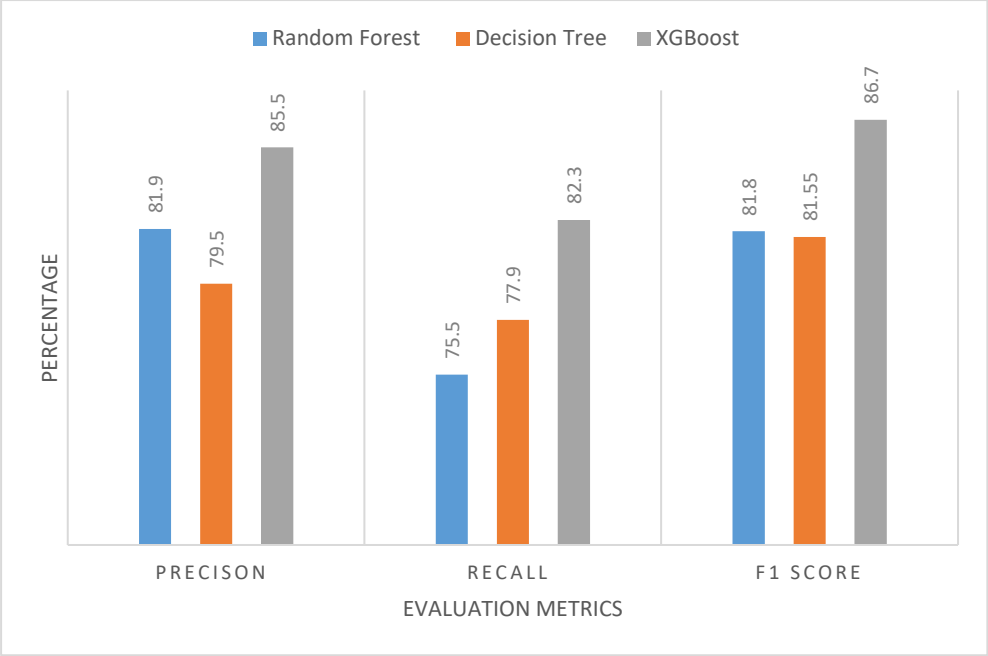

Figure 3. Evaluation Results

5. Conclusion

In this study, the application of the XG Boost, Decision Tree and the Random Forest for detecting masked and phishing URLs is explored. The objective was to develop an accurate and efficient model that can predict the malicious websites based on website data, by leveraging a comprehensive dataset containing diverse and extensive URLs. The three-machine learning model were trained and evaluated to assess its performance. The three models demonstrated strong performance in accurately identifying the malicious websites with high accuracy and precision The XG boost outperformed the other two models with the test accuracy of 86.8 %. Further in future the research aims in developing a user interface that would be helpful for the business organizations as well as individuals to discriminate among the genuine and malicious URLs by integrating the machine learning model identified with high testing accuracy in the comparison that was performed.

References

[1] Shahrivari, Vahid, Mohammad Mahdi Darabi, and Mohammad Izadi. "Phishing detection using machine learning techniques." arXiv preprint arXiv:2009.11116 (2020).

[2] Babu Rao Pawar, Nagasunder Rao Pawar. "Detection of Phishing URL using Machine Learning." PhD diss., Dublin, National College of Ireland, 2021."

[3] Patil, Dharmaraj R., and Jayantro B. Patil. "Malicious URLs detection using decision tree classifiers and majority voting technique." Cybernetics and Information Technologies 18, no. 1 (2018): 11-29.

[4] Fazal, Ashar Ahmed, and Maryam Daud. "Detecting Phishing Websites using Decision Trees: A Machine Learning Approach." *International Journal for Electronic Crime Investigation* 7, no. 2 (2023).

[5] Selvakumari, M., M. Sowjanya, Sneha Das, and S. Padmavathi. "Phishing website detection using machine learning and deep learning techniques." In *Journal of Physics: Conference Series*, vol. 1916, no. 1, p. 012169. IOP Publishing, 2021.

[6] Salloum, Said, Tarek Gaber, Sunil Vadera, and Khaled Shaalan. "A systematic literature review on phishing email detection using natural language processing techniques." *IEEE Access* 10 (2022): 65703-65727.

[7] M I, Shilpa. “Malicious Websites Classification Using Machine Learning Techniques: A Survey Paper.” International Journal for Research in Applied Science and Engineering Technology (2022): n. pag.

[8] Khonji, Mahmoud, Youssef Iraqi, and Andrew Jones. "Phishing detection: a literature survey." *IEEE Communications Surveys & Tutorials* 15, no. 4 (2013): 2091-2121. [9]. Aggarwal, Anupama, Ashwin Rajadesingan and Ponnurangam Kumaraguru. “PhishAri: Automatic realtime phishing detection on twitter.” 2012 eCrime Researchers Summit (2012): 1-12.

[9] Singh, Charu. "Phishing website detection based on machine learning: A survey." In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 398-404. IEEE, 2020.

[10] TRAGHA, Abderrah“m. "Machine learning for web page classification: a survey." International Journal of Information Science and Technology 3, no. 5 (2019): 38-50.

[11] Vanhoenshoven, Frank, Gonzalo Nápoles, Rafael Falcon, Koen Vanhoof, and Mario Köppen. "Detecting malicious URLs using machine learning techniques." In 2016 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1-8. IEEE, 2016.

[12] https://www.kaggle.com/datasets/shashwatwork/web-page-phishing-detection-dataset