iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 Issue: VIII Month of publication: Aug 2023

DOI: https://doi.org/10.22214/ijraset.2023.55506

Improved Phishing Detection using Ensemble Models in Machine Learning

Sri Sai Phani Venkat Dasari[1], Dr. Nitalaksheswara Rao Kolukula[2]
Department of Computer Science, GITAM (Deemed to be University)

***Abstract:** A phishing attack is one of the simplest ways to obtain sensitive information from unaware, innocent users. The main motive of the phishers is to acquire critical information like usernames, passwords, and bank account details using a malicious link that looks genuine. Users with sound technical knowledge might be able to identify these links quickly, but this will cause harm to naive users, leading to a loss of privacy and assets. There are techniques to detect spam, such as content-based and the sender's reputation-based detection. This project aims to present an approach to detect phishing attacks based on the URL and by applying Machine Learning.*

***Keywords:** Phishing, Machine Learning, Classification, Ensemble, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Random Forest, XG Boost, Stacking*

I. INTRODUCTION

Phishing attacks are limitedly defined as stealing personal information from users by a non-trusted source pretending to be a trusted source; however, this is not always true. A link is said to be phishing whenever it acts as a genuine one to confuse naive users and make them perform an action they would do only if they trust. Although phishing is the most straightforward attack to obtain information illegally, it depends on the user's weakness. People call these attacks 'Hacking', but it is a misconception. It is a trick being played by phishers to trap users. The cyber criminals who originate phishing attacks are known as phishers. The phishers usually grab the personal information of people and misuse it. The most common cases are trading information and manipulating accounts. The other possibility of phishing is to inject a worm into the user's system, which gives control to phishers or damages the software/hardware. Recently, the number of phishing cases has been increasing all around the world. The existing spam detection techniques were not able to save people from phishing attacks.

URL means Uniform Resource Locator. URL is an address of a location where specific resources are stored on the internet, and the users get to access them. So, URLs contain a lot of information. The behaviour of the URL, its domain information, and the content on its page describe the nature of a URL. Classification is a Machine Learning concept that can help in detecting phishing attacks. Differentiating benign and phishing links is possible through classification, a supervised machine-learning approach.

II. LITERATURE REVIEW

While social engineering use began to rise worldwide, Phishing has been a simple and convenient way to manipulate naive users and obtain their data maliciously. The first phishing attack was found to happen in the mid-1990s and targeted American online users. The victims unknowingly provided their login details in the phishing links, and the Phishers started using the victim's accounts for spamming and adding likes, said in [1]. In 2000, people received emails titled 'I LOVE YOU,' attached with a love letter. The systems of the users who clicked on the letter got injected by a worm that obtained all the personal image files and sent them to all the contacts in Outlook. The above information is present in [2]. According to [3], Indian citizens are receiving links through SMS pretending to be from official banks or government bodies, saying that the user needs to update his identity details like Aadhar number and PAN immediately. Innocent users tend to click on those links and provide the information, believing it is genuine. The government and network providers can only do something other than warn the users to avoid these links.

The current detection techniques in SMS and Mail applications are based on content-based and sender reputation-based detection, as shown [6] and [7]. The URL detection is based on its features. According to [4], the features are classified into address-bar-based, HTML-JavaScript-based, and Domain-based features. Many studies have evolved around Machine Learning based on different kinds of features. Using different kinds of datasets, different sets of features, and different algorithms resulted in various outcomes. According to Machine Learning, the suitable concept for this problem is Classification, which can be observed in [8]. Generally, a classifier is built to decide which class a particular input belongs to. Based on its features, the URL needs to be classified into a class to which it belongs, either a Phishing URL or a Benign URL. From [5], the proposed algorithm was SVM.

III. PROPOSED FRAMEWORK

A. System Architecture

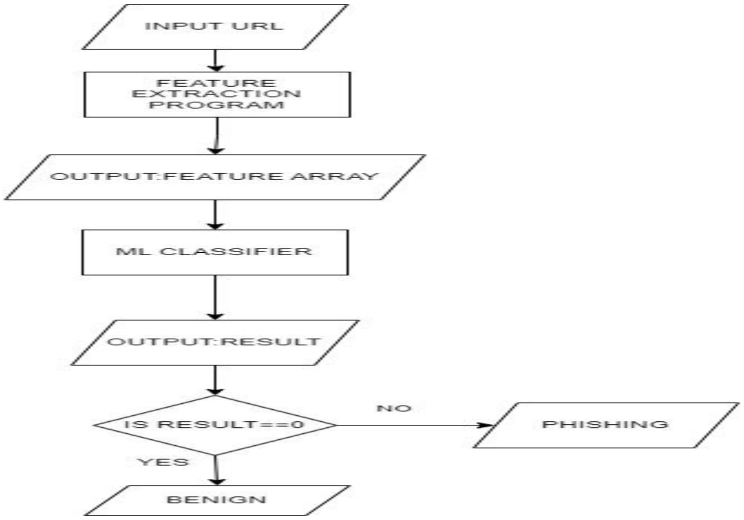


Fig. 1 Proposed System Architecture

As shown in Fig. 1, the input URL is passed through the Feature Extraction Program, and its execution results in an array of feature values. Feature Extraction Program consists of string methods to extract address-based features, web scraping to extract content-based features, and API calls to extract domain-based features. The output array of features is referred to as a Feature Array. The feature array is now passed as an input to the ML Classifier. The classifier, based on the training, makes the detection. If the classifier's output is 0, then it is a benign URL; otherwise, it is a phishing URL.

B. Dataset

A dataset has been taken from Kaggle and re-processed. According to the current scenario, only the required columns(features) have been considered. The dataset consists of 11,430 URLs, of which 5,715 are Phishing, and 5,715 are Benign. The Phishing URLs are labelled '1', and the Benign URLs are labelled '0'.

C. Data Visualization

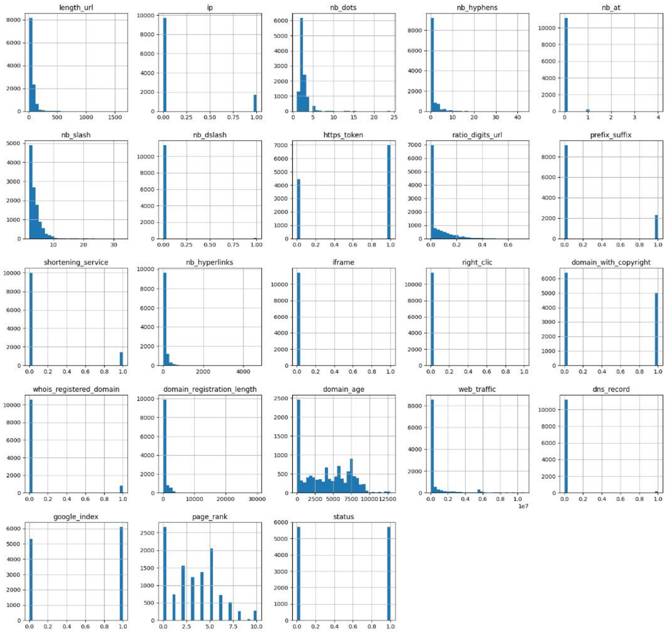


Fig. 2 Bar graphs between features and number of URLs

Fig. 2 shows the bar graphs. Each bar graph in the figure represents the relation between a feature and the number of URLs.

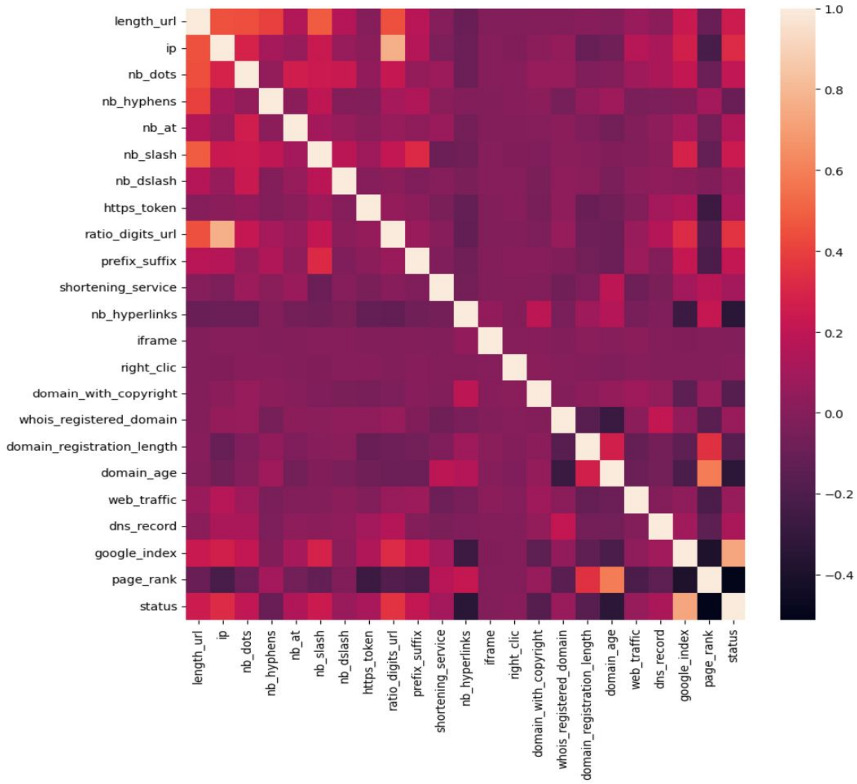


Fig. 3 Correlation Heatmap between features

Fig. 3 shows the Correlation Heatmap between the features. The lighter colour represents higher correlation and the darker colour represents lower correlation.

D. Features

Although many features can be taken from the URL, we have taken the 18 features that play a significant role in the classification. The following are the features considered:

1) Length of the URL
2) Presence of IP address in the URL
3) Number of '.' in the URL
4) Number of '-' in the URL
5) Number of '@' in the URL
6) Number of '/' in the URL
7) Number of '//' in the URL
8) Presence of HTTP/HTTPS token in the URL
9) Prefix-Suffix separation in Domain name
10) URL Shortening Services
11) Number of Hyperlinks in Source code
12) WHOIS Registration of Domain
13) DNS Records
14) Domain Registration Length
15) Domain Age
16) Web Traffic
17) Google Index
18) Page Rank

E. Algorithms

1) Logistic Regression: Logistic Regression is a machine learning algorithm for classification. The classification in Logistic regression is based on the probabilities between the classes. The logistic function curve shows the probability of something to which class it belongs.
2) Support Vector Machine: Support Vector Machine, or SVM, is one of the most popular Supervised Learning algorithms. It is most widely used for classification, although it can perform for Regression problems. The main objective of SVM is to create a hyperplane, which can divide the dimensional space into classes. The algorithm uses extreme data points as support vectors that help create the hyperplane.
3) K-Nearest Neighbor: K-Nearest Neighbor is simply known as KNN algorithm. KNN is the most straightforward algorithm which can be used for both Regression and Classification. In the current scenario, we are using this algorithm to classify URLs. KNN differentiates between the classes based on the distance between the data points. The most common metrics used to calculate the distance between the data points are Euclidean distance and Manhattan’s distance.
4) Random Forest: Random Forest is a supervised machine learning algorithm. It is a type of bagging algorithm in Ensemble learning. It is an algorithm built on the idea of integrating various classifiers to solve complex issues and enhance model performance. In Random Forest, instead of depending on one decision tree, multiple decision trees are used for making a prediction based on the majority value. This algorithm can be used under regression and classification problems.
5) XG Boost: XG Boost Algorithm is a Supervised Machine Learning Algorithm. XG Boost is a boosting type of algorithm in Ensemble Learning. Boosting is a type of Ensemble technique in machine learning where models are built sequentially; a model is built, and then, based on its performance, the following model is built. Here, every model will be a weak learner to its next model, which will stop when we get a model with good performance. The final model will be a strong learner.
6) Stacking: Stacking is a type of ensemble learning technique. Ensemble Learning is a learning technique where the predictive power of multiple models called the base models is taken, and a resultant predictive model is formed. The resultant predictive model formed has much more predictive power than the individual model predictive accuracy. In stacking, a model is used to find the final predictive learning classifier from the other base learners using shared data.

IV. RESULTS AND DISSCUSSIONS

Logistic Regression, SVM, and KNN are the conventional algorithms. Random Forest, XG Boost, and Stacking are algorithms working on the principle of Ensemble Learning. The stacking algorithm is unique and customizable; our choice can take the Level-1 models (weak learners) and meta-model. In this case, the Level-1 models are KNN, Random Forest, and XG Boost, and the meta-model is SVM. Table 1 shows the performance measures of the algorithms when the test size is 20%. The SVM performed the least compared to the other five algorithms. The Logistic Regression is better than SVM but not the best. KNN gave decent results. Random Forest and XG Boost performed well, with above 90% scores. Using better performance models as Level-1 models in Stacking, could give a slight but essential improvement in the performance. There is much distinctness among the KNN, Random Forest, and XG Boost outputs. SVM might be a better choice for the meta-model. The implementation of this Stacking model gave the highest scores, as shown in Table 1. The Stacking classifier performed with the highest accuracy of 97.46%, precision of 97.44%, recall of 97.37%, and F1-score of 97.29%.

TABLE I
PERFORMANCE MEASURES

Model	Accuracy (in %)	Precision (in %)	Recall (in %)	F1-Score (in %)
Logistic Regression	78.83	76.86	78.62	89.69
Support Vector Machine	71.57	65.32	74.94	87.88
K-Nearest Neighbor	83.94	83.33	84.43	83.88
Random Forest	94.09	94.54	93.45	93.99
XG Boost	96.63	96.47	96.72	96.60
Stacking Classifier	97.46	97.44	97.37	97.29

V. CONCLUSION AND FUTURE SCOPE

Phishing attacks have been quite vulnerable to users for many years. In this project, the best possible approaches were attempted for collecting the best set of features and finding the suitable Machine Learning approaches for phishing URL detection. The traditional algorithms, Logistic Regression, SVM, and KNN, must perform better. Random Forest and XG Boost are giving improved results with the same input data. The stacking technique has given much better results than the previous models. The significance of ensemble techniques compared to conventional algorithms has been shown. As per understanding, Stacking can be an excellent technique to make slight improvements for Classification and Regression problems when we already have performing models. In the future, this methodology can be used to detect phishing websites accurately and reduce society's significant social engineering problem. The future scope of this methodology would be automating this model in mobile and computer applications to detect phishing sites.

REFERENCES

[1] History of Phishing. [Online]. Available: https://cofense.com/knowledge-center/history-of-phishing/

[2] Love bug virus creates worldwide chaos. [Online]. Available: https://www.theguardian.com/world/2000/may/05/jamesmeek

[3] A new phishing attack lurking to scam banking customers: Advisory. [Online]. Available: https://timesofindia.indiatimes.com/business/india-business/a-newphishing-attack-lurking-to-scam-banking-customers-advisory/articleshow/85236685.cms

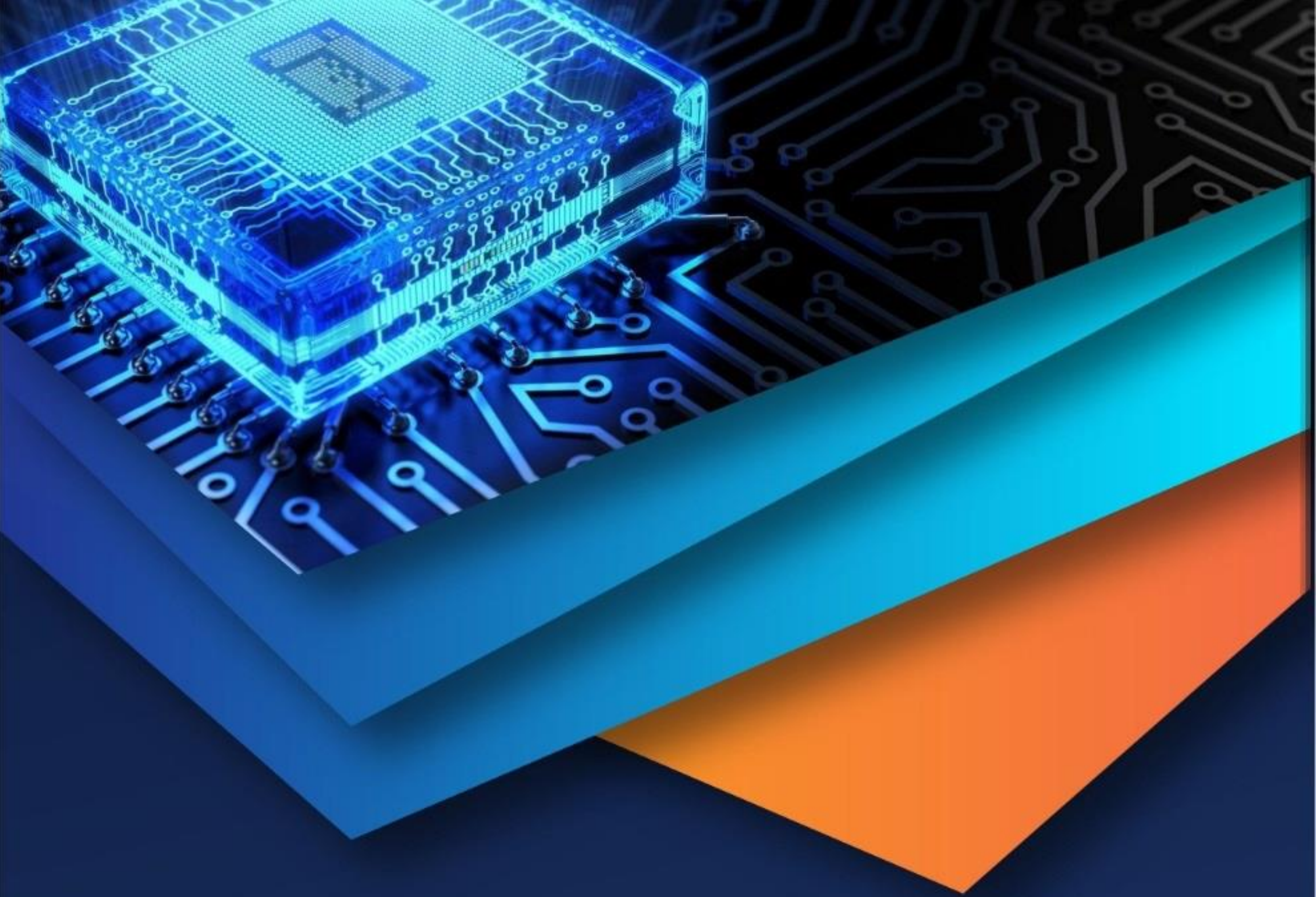
[4] R. M. Mohammad, F. Thabtah and L. McCluskey, "An assessment of features related to phishing websites using an automated technique," 2012 International Conference for Internet Technology and Secured Transactions, London, UK, 2012, pp. 492-497.

[5] J. Rashid, T. Mahmood, M. W. Nisar and T. Nazir, "Phishing Detection Using Machine Learning Technique," 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), Riyadh, Saudi Arabia, 2020, pp. 43-46, doi: 10.1109/SMART-TECH49988.2020.00026.

[6] Uur Ozker, Ozgur Koray Sahingoz, "Content Based Phishing Detection with Machine Learning", 2020 International Conference on Electrical Engineering (ICEE), 25-27 September 2020.

[7] S. Naksomboon, C. Charnsripinyo and N. Wattanapongsakorn, "Considering behavior of sender in spam mail detection," INC2010: 6th International Conference on Networked Computing, Gyeongju, Korea (South), 2010, pp. 1-5.

[8] S. Chowdhury and M. P. Schoen, "Research Paper Classification using Supervised Machine Learning Techniques," 2020 Intermountain Engineering, Technology and Computing (IETC), Orem, UT, USA, 2020, pp. 1-6, doi: 10.1109/IETC47856.2020.9249211.



10.22214/IJRASET

INDEX COPERNICUS

45.98

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY