IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

PHISHING WEBSITE DETECTION USING MACHINE LEARNING BY ANALYZING URL

Ms. Jithina Jose
Assistant Professor, Computer Engineering
Dr. D. Y. Patil Institute of Technology,
Pimpri Pune 18

Vishwesh Patil
Department of Computer Engineering
Dr. D. Y. Patil Institute of Technology,
Pimpri Pune 18

Himanshu Gupta
Department of Computer Engineering
Dr. D. Y. Patil Institute of Technology,
Pimpri Pune 18

Kartik Gandhi
Department of Computer Engineering
Dr. D. Y. Patil Institute of Technology,
Pimpri Pune 18

Gaurav Bhandari
Department of Computer Engineering
Dr. D. Y. Patil Institute of Technology,
Pimpri Pune 18

***Abstract* — Nowadays, everyone is highly dependent on the internet. Our financial work, office-related work, shopping, and any other daily activities have been moved to the internet. This really makes our daily lives easy but at the same time, we are also exposed to greater risks through the internet which are cybercrimes. Phishing is a cybercrime in which attackers often imitate some popular banking and e-commerce sites and tries to steal the user's sensitive information as well as the login credentials and credit card numbers. They target both individuals and organizations, convince them to click on URLs that look legit and secure, and steal the information or inject malware into the system. So, as the internet grows, URL detection becomes very important to provide timely protection to individuals and organizations. In this project, we aim to implement various machine-learning algorithms to analyze the URLs with the dataset and URL features to train the machine-learning models.**

Keywords – Phishing, Cyber Security, Machine Learning, Website Classification

I. INTRODUCTION

Phishing attacks are becoming more sophisticated and prevalent, posing a significant threat to users' security and privacy. According to a recent report, phishing attacks have increased by 220% in the last year alone. Phishing websites are one of the most common methods used by attackers to steal sensitive information from users. These websites mimic legitimate websites, such as online banking or social media, to trick users into revealing their login credentials or other sensitive information.

To address this problem, several approaches have been proposed to detect phishing websites. These approaches typically rely on analyzing various features, such as the website's content, structure, or behavior. However, analyzing these features can be computationally expensive and time-consuming, making them unsuitable for real-time detection.

In this paper, we propose a lightweight approach for detecting phishing websites by analyzing the URL. The URL is a key component of a website, and it contains valuable information that can be used to distinguish between phishing and legitimate websites. By analyzing the URL's various features, we can identify patterns and characteristics that are typical of phishing websites.

II. LITERATURE REVIEW

To address this problem, researchers have developed various techniques for detecting phishing websites. One of the approaches is to analyze the URL of a website to determine whether it is legitimate or fraudulent.

One of the studies in this field was conducted by Arathi Krishna (2021) [1], who proposed a feature-based approach and dimensionality reduction techniques for detecting phishing websites. In this paper, the author has analyzed various phishing detection approaches and discussed the most common machine learning based approaches. By applying feature selection

algorithms and dimensionality reduction techniques, we can reduce the number of features and remove irrelevant data. They have done a study of the process of phishing detection and the phishing detection schemes in the recent research literature which will serve as a guide for new researchers to understand the process and to develop more accurate phishing detection systems.

Several studies proposed using machine learning algorithms to analyze the structure and content of URLs for phishing detection. For example, Seibu Mary Jacob (2020) [2] chose 3 algorithms for classification – Linear Regression, SVM and Random Forest as well as deep learning too. They used URL, DOM structure, page rank and page information related features. Similarly, Bhagwat M.D. [3] proposed an approach to solving the fuzziness in the phishing website assessment and propose an accurate and smart model for detecting phishing websites. His phishing detection technique is based on fuzzy logic and machine learning algorithms in order to distinguish different factors on the phishing website.

In another paper by Sumitra Das Gupta [4], they proposed a data-driven approach for the purpose of detecting phishing websites using various machine learning classifiers, such as Decision Tree, XGBoost, Random Forest, Support Vector Machine, and Naive Bayes by implementing various numbers/types of features such as URL-based features, hyperlink-based features, and hybrid features. Their dataset consists 6000 URLs containing 3000 legitimate URLs and 3000 phishing URLs from Phish tank website.

III. MODEL ARCHITECTURE

Five different algorithms were used in this project. These are: Logistic Regression, K Nearest Neighbors, Gaussian Naïve Bayes, Decision Tree, Random Forest and Gradient Boosting.

A. Logistic Regression

It is a supervised machine learning technique which is mostly used to solve classification problems. The result of this algorithm is the probabilistic value which is between 0 and 1. The logistic regression model estimates the values of the coefficients that best fit the data and can be used to make predictions on new data.

B. K Nearest Neighbors (KNN)

KNN is a supervised learning algorithm which can be used for both classification as well as regression problems. It operates by identifying the K closest training examples to a new input data point and using the class labels or target values of those examples to make a prediction on the new data point. The value of K is a hyper parameter that can be tuned to optimize the performance of the model. A small value of K can lead to overfitting, while a large value of K can lead to under fitting. The algorithm is simple and interpretable, but can be computationally expensive for large datasets and high-dimensional feature spaces.

C. Gaussian Naïve Bayes

It is a probabilistic algorithm for classification problems. It is based on the Bayes theorem. Here, it is assumed that the input features are independent of each other, and they follow a Gaussian distribution. Gaussian Naive Bayes operates by estimating the mean and variance of the input features for each class label during training, and then using these estimates to calculate the afterwards probability of the class label given the input features during inference. The algorithm is simple and computationally efficient, but the assumption of feature independence can limit its performance on more complex datasets where the features are correlated.

D. Random Forest Classifier

It is an ensemble machine learning algorithm used for classification tasks. It consists of multiple decision trees. Each tree is constructed by randomly selecting a subset of the training data and input features, and then combines the predictions of the trees to make a final prediction. The algorithm is powerful, interpretable, and robust to overfitting, and has been used in a variety of applications.

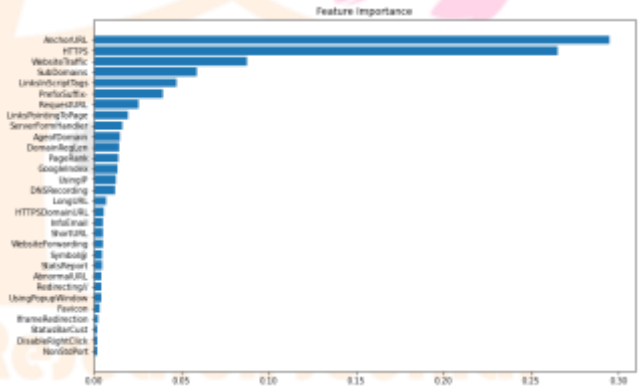
E. Gradient Boosting

It is also an ensemble machine learning algorithm for both regression and classification tasks. It works by combining multiple weak learners to create a strong learner. It creates an ensemble of decision trees incrementally, using the negative gradient of the loss function to correct the errors made by the previous trees. The algorithm can handle both numerical and categorical input features, and can learn complex nonlinear relationships between the input features and the target variable. It is a powerful algorithm that has been used in a variety of applications.

IV. PROPOSED SYSTEM

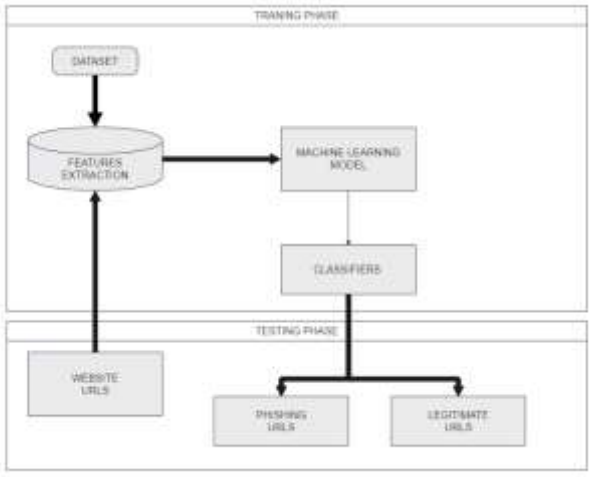
A) *Dataset Description*

This research paper utilizes the dataset from Kaggle, which has a collection of website URLs for 11000+ websites. Each sample has 30 website parameters and a class label identifying it as a phishing website or not (1 or -1). The important features are presented in the bar graph below.



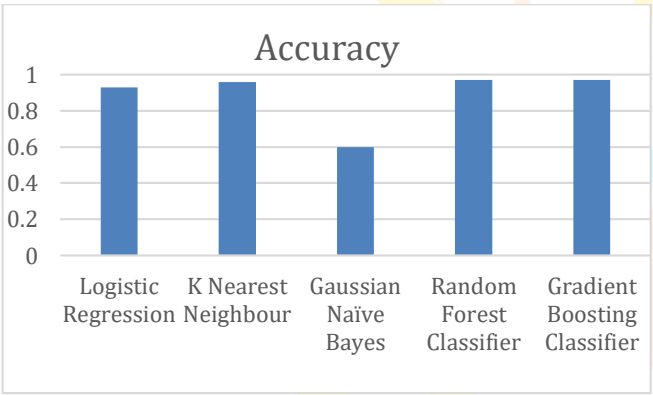
Our proposed method consists of two stages: feature extraction and classification. In the feature extraction stage, we extract various features from the URL, such as the domain name, length, and the presence of certain keywords. These features are used to build a feature vector that represents the URL. We also perform some pre-processing steps, such as removing any special characters or encoding the URL to remove any ambiguity.
In the classification stage, we use machine learning algorithms to classify the URL as either phishing or legitimate. We experimented with several classification algorithms, including decision trees, random forests, and support vector machines. We also used various evaluation metrics, such as accuracy, precision, and recall, to measure the performance of our model

We build the classifier using 5 different techniques by following below steps:
1) Split the data into training and test dataset, which we take 20% for testing and 80% for training.
2) We train and test with all 30 features or the possible combination of 30 features present in the dataset to get the strongest features that arise the accuracy of detection.
3) Then execute our final classifier with the highest classifier.

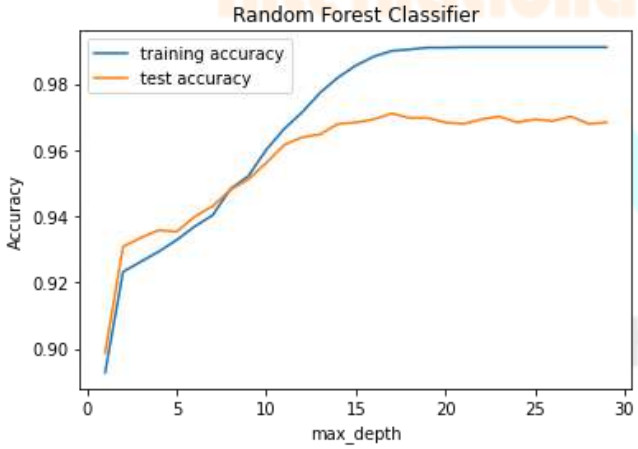


System Architecture

V. RESULT



The above table contains the accuracy rate of different algorithm. According to these data, the highest test classification result was obtained in the model using Random Forest Classifier with an accuracy rate of 97%. At the same time, Gaussian Naïve Bayes had the lowest accuracy with 60% accuracy.



The above fig shows that the training and test accuracy of Random Forest Classifier increases gradually till maximum depth of 15. After that there is rare change in the test accuracy.

In general, when analyzing the results, it is clear that only Gaussian Naïve Bayes has a lower accuracy rate while all other algorithms has more than 90% accuracy.

VI. CONCLUSION AND FUTURE SCOPE

We implemented a phishing detection system by using some machine learning algorithms. The proposed systems are tested with the dataset. The results show that the proposed systems have very good accuracy rates. As for future works, we can create a new and huge dataset for URL based Phishing Detection Systems. Then we plan to improve our system by using some hybrid algorithms as well as try deep learning models in it.

REFERENCES

1. Arathi Krishna, Anusree A, Blessy Jose, Karthika Anilkumar, Ojus Thomas Lee, “Phishing Detection using Machine Learning based URL Analysis: A Survey”, on 2021 International Journal of Engineering Research & Technology (IJERT)
2. Seibu Mary Jacob, Jordan Stobbs and Biju Issac, “Phishing Web Page Detection Using Optimized Machine Learning”, on 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications.
3. Bhagwat M.D. , Dr. Patil P. H, Dr. T. S. Vishawanath, “A Methodical Overview on Detection, Identification and Proactive Prevention of Phishing Websites”
4. Sumitra Das Gupta, Khandaker Tayef Shahriar, Hamed Alqahtani, Dheyaaldin Alsalman, Iqbal H. Sarker, “Modeling Hybrid Feature-Based Phishing Websites Detection Using Machine Learning Techniques”
5. Asst. Prof. Deepa Mary Vargheese, Sreelakshmi N R, “Phishing Website Detection using Machine Learning Techniques and CNN”
6. Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Dr.Aram Alsedrani, “Detecting Phishing Websites Using Machine Learning”
7. Tianrui Peng, Ian G. Harris, Yuki Sawa, “Detecting Phishing Attacks Using Natural Language Processing and Machine Learning”
8. Aniket Garje, Namrata Tanwani, Sammed Kandale, Twinkle Zope, Prof. Sandeep Gore, “Detecting Phishing Websites Using Machine Learning”
9. Mehmet Korkmaz, Ozgur Koray Sahingoz, Banu Diri, “Detection of Phishing Websites by Using Machine Learning-Based URL Analysis
10. Weiheng Bai, “Phishing Website Detection Based on Machine Learning Algorithm
11. Vaibhav Patil, Pritecsh Thakkar, Chirag Shah, Tushar Bhat, Prof. S.P. Godse, “Detection and Prevention of Phishing Websites Using Machine Learning Approach