IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Prediction Of Phishing Url Using Machine Learning Classifier

[1]Bhuvaneshwari S. Patil, [2]Ashvini S. Patil, [3]Devayani R. Mahajan, [4]Khushal G. Patil, [5]Priti R. Sharma

[1,2,3,4]UG Student, [5]Assistant Professor, Department of Computer Engineering, SSBT's College of Engineering and Technology Jalgaon, Maharashtra, India

Abstract: Along with the rise in technology, there is an increase in attacks involving digital platforms. It is very important to stay secure in this digital world. To respond to these attacks, we propose a method that detects phishing websites by categorizing the Internet URL and domain names of websites with the Random Forest classifier algorithm according to seventeen predetermined features. To illustrate the highest accuracy rate, the use of Random Forest algorithm is preferred. In this method, a dataset with 10,000 URLs in which 5000 URLs are non-phishing websites and 5000 URLs are phishing websites to be labelled according to seventeen predetermined features.

Index Terms - Cybersecurity, Phishing detection, Phishing domains, Legitimate domains, Random Forest.

I. INTRODUCTION

Phishing is a cyberattack where the attacker deceives the victim by making him fall into his trap by presenting himself as a legitimate source. By creating urgency and fear in the victim, he acts without thinking and falls into the trap.

For cybersecurity professionals, it is very easy to understand the deceptive tactics of attackers, but for people who do not have knowledge of such attacks, it is difficult for them to understand such attacks.

With the growth of online services, cybercriminals exploit the vast amounts of sensitive data exchanged. Such attacks are mostly committed for money. Attacker uses different ways to trick the victim. In such attacks fear of urgency is created so, that the victims acts, without thinking.

II. MOTIVATION

The major motivation behind this project was to learn about cybersecurity. This project has helped us to gain huge knowledge about phishing and its techniques. The aim is to spread education and awareness about cyberattacks and how to save yourself from them. This digital world is full of attackers who are always in search of new deceptive techniques. It is very important that we stay alert and beware of such attacks.

III. PROBLEM STATEMENT

Create a website to detect phishing domains that imitate look and feel genuine domains using machine learning. This project aims to create a website that can detect a phishing domain and a legitimate domain. The characteristic of a phishing domain is that it looks similar to a legitimate domain and due to this similarity of look, victims are tricked by the attacker.

In this project, we will be using machine learning to detect the domains. The reason behind using machine learning is because of the accuracy it provides. Accuracy plays a very important role in distinguishing between phishing and non-phishing URLs.

IV. LITERATURE SURVEY

Kara [1] used a random forest model for the training of data. It was able to produce acceptable results. On unseen data, it achieved high prediction rates. It was found that the prediction rate was 98%. They used an up-to-date intelligence database.

Abdul Karim [2] observed that the random forest model outperformed all other machine learning models. It showed that ensemble that ensemble tree-based models achieved better results than others. This study also experiments the performance of hybrid (LR+SV+DT) with soft and hard voting.

Majid Alshammari [3] developed phishing detection models using machine learning. They used the UCI dataset for detecting different types of attacks. They employed the Min-Max normalization feature as preprocessing. They found that the RF model has the highest detection rate at 97%.

A. V. Bhagyashree [4] investigated four models, KNN, Kernel SVM, Decision Tree and Random Forest classifier. They used 30 features for prediction. Random forest gave the best accuracy score.

V. RANDOM FOREST ALGORITHM

Machine Learning has many types of algorithms, and the Random Forest Algorithm is one of them. Random forest is a forest of multiple decision trees and also called an ensemble of decision trees [1]. Multiple decision trees are created on training data. Training of each decision tree is done using random data and random features. Each Decision Tree in the ensemble makes predictions independently.

The majority of votes of each decision tree forms the basis of prediction. Even in the case of new, unseen data, each decision tree predicts independently, and the majority of votes is considered the final result [2]. Random forest algorithm can be applied to both classification and regression tasks [4].

It is very efficient to handle missing values during training, eliminating the need for manual imputation. It scales well with large and complex data without any performance degradation. It selects features based on their importance in making predictions, which enhances feature selection.

Each decision tree is unique, as when creating an individual tree, not all features are taken into account. Each decision tree is different from the others. Prevention of the model from overfitting and making the predictions more accurate and reliable is due to the randomness in data samples and feature selection. The result is calculated on the basis of majority votes [3].

VI. METHODOLOGY

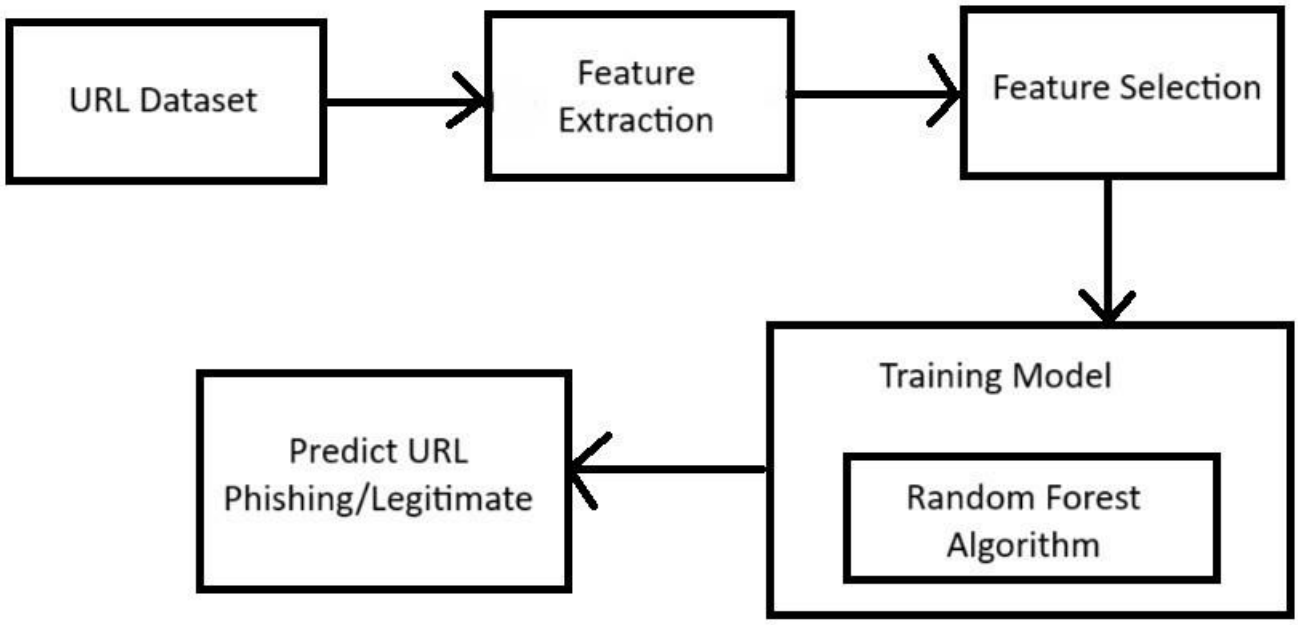

Fig: System Architecture Diagram

A dataset of 10000 URLs has been used. It is important to clean the data and check if there are any missing values. Feature extraction is done on both datasets, and extraction of 17 features is done by using Python modules and libraries. All the extracted features are stored along with the URLs. Here phishing is predicted as 1 and legitimate as 0.

We use this new dataset, which contains feature-extracted columns, to train the model. The dataset is split into training and testing set for training the model. By applying 80-20 split, the training set contains 8000 URLs and the testing set contains 2000 URLs. Using the scikit learn library, we make a random forest classifier.

6.1 Data Collection

The phishing dataset is collected from open source Phish tank and Legitimate URLs are collected from open datasets of the University of New Brunswick.
To download the phishing data: https://www.phishtank.com/developer_info.php.
To download the legitimate data: https://www.unb.ca/cic/datasets/url-2016.html

6.2 Feature Extraction

Overall, 17 features are extracted using python libraries. The extracted features are stored in a list. Feature extraction is very important to understand the characteristics of URL and differentiate between phishing and non-phishing URLs. Python Libraries such as beautiful soup, re, datetime, requests are used for performing the task of feature extraction.

6.3 Feature Selection

Label feature is selected and remaining features that are extracted also containing the domain are selected. Features with greater importance are selected to enhance accuracy. Random forest model naturally selects features by selecting most important feature on the basis of Gini impurity.

6.4 Training the model

Random forest is a supervised machine learning algorithm. As the project is about detecting phishing and non-phishing domains, the random forest algorithm model is chosen because of the accuracy it provides.

Before splitting the dataset, we randomly shuffle all the data for better accuracy. For training the model, we divide the dataset into 80% training and 20% testing dataset.

6.5 Result

Detection of phishing and legitimate URLs is achieved. Output as “This is a Phishing URL” is shown for phishing URL and “This is a Legitimate URL” is shown for legitimate URL.

VII. RESULT

The Random Forest model has achieved an accuracy of 81.9 on the training dataset and 81.5 on the test dataset. It is able to perform real-time detection of websites. When presented with new unseen data, it is efficiently able to display the results with accuracy.

The output is displayed in the form of text such as, for Phishing URL the output is displayed as “This is a phishing URL” and for legitimate URL the output displayed is “This is a legitimate URL”. It is tested on 15 phishing and 15 legitimate domains which is the unseen data. It is able to predict accurately and show the right result.

No.	Accuracy		
	ML Model	Train Accuracy	Test Accuracy
1.	Random Forest Classifier	0.819	0.815

Fig: Accuracy Table

VIII. CONCLUSION

Machine learning method is used for phishing detection, as it provides high accuracy. A Machine learning model based on random forest (RF) techniques is developed. Real-time detection is achieved using this method.

It is able to identify phishing domains. To provide ease of use, a user-friendly interface is created. It provides high accuracy by the majority of votes of multiple decision trees. It chooses the most relevant features for feature selection. It is efficiently able to handle large and complex data.

REFERENCES

[1] Kara, M. Ok, and A. Ozaday, "Characteristics of Understanding URLs, and Domain Names Features: The Detection of Phishing Websites with Machine Learning Methods," *IEEE Access*, pp. 1–1, 2022, Doi: https://doi.org/10.1109/access.2022.3223111

[2] Abdul Karim, and Mobeen Shahroz**,** "Phishing Detection System Through Hybrid Machine Learning Based on URL," *ieeexplore.ieee.org*. https://ieeexplore.ieee.org/abstract/document/10058201

[3] S. Alnemari and M. Alshammari, "Detecting Phishing Domains Using Machine Learning," *Applied sciences*, vol. 13, no. 8, pp. 4649–4649, Apr. 2023, Doi: https://doi.org/10.3390/app13084649.

[4] A. V. Bhagyashree and A. K. Koundinya, "Detection of phishing websites using machine learning techniques", *Int. J. Computer. Sci. Inf. Security.*, vol. 18, no. 7, 2020.

[5] "Phish Tank> Developer Information," *www.phishtank.com*. https://www.phishtank.com/developer_info.php

[6] "URL 2016|Datasets|Research | Canadian Institute for Cybersecurity | UNB," *www.unb.ca*, 2016. https://www.unb.ca/cic/datasets/url-2016.html