# URL PHISHING DETECTION USING MACHINE LEARNING

**[1]Avadhut palange, [2]Rutuja Baravkar, [3]Chaitanya Burande,**
**[4] Saurabh Dashpute, [5]Prof.D.B. Mane**

[1,2,3,4]B. E Students, [5]Professor
Department of Information Technology,
Smt. Kashibai Navale College of Engineering
Pune, Maharashtra, India

***Abstract*—Phishing URL mainly target individuals and/or organization through social engineering attacks by exploiting the humans' weaknesses in information security awareness. These URLs lure online users to access fake websites and harvest their confidential information, such as debit/credit card numbers and other sensitive information. In this work, we introduce a phishing detection technique based on URL lexical analysis and machine learning classifiers. This dataset was processed to generate 22 different features that were reduced further to a smaller set using different features reduction techniques. Random Forest, Gradient Boosting, Neural Network, Xgboost and Support Vector Machine (SVM) classifiers were all evaluated, and results show the superiority of SVMs, which achieved the highest accuracy in detecting the URLs with a rate of 99.89%. Our approach can be incorporated within add-on/middleware features in Internet browsers for alerting online users whenever they try to access a phishing website using only its URL.**

***Index Terms*—Security, Phishing attacks, Machine Learning.**

## I. INTRODUCTION

A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and webpages. The objective of this project is to train machine learning models and deep neural nets on the dataset created to predict phishing websites. Both phishing and benign URLs of websites are gathered to form a dataset and from them required URL and website content-based features are extracted. The performance level of each model is measured and compared. A phisher usually sets up a deceptive website, where the victims are conned into entering credentials and sensitive information. It is therefore important to detect these types of malicious websites before causing any harmful damages to victims. Inspired by the evolving nature of the phishing websites, this paper introduces a novel approach based on deep reinforcement learning to model and detect malicious URLs. The proposed model is capable of adapting to the dynamic behavior of the phishing websites and thus learning the features associated with phishing website detection.

## II. NEED OF STUDY

Phishing has a list of negative effects on a business, including loss of money, loss of intellectual property, damage to reputation, and disruption of operational activities. These effects work together to cause loss of company value, sometimes with irreparable repercussions.To avoid such problems we are introducing our model which uses multiple machine learning techniques such as Decision tree , Random Forest, Multilayer Perceptrons , XGBoost , Autoencoder Neural Network ,SVM etc. for detection of phishing websites.

## III. RESEARCH METHODOLOGY

We are considering multiple techniques for our research such as the following.
1.Decision Tree
2.Random Forest
3.Multilayer Perceptrons
4.XGBoost
6.Autoencoder Neural Network
7.Support Vector Machine
8.Reinforcement learning

### *3.1. Methodology*

**Machine learning**:
It is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.
**Loading Data:**
The features are extracted and stored in the CSV file. The working of this can be seen in the 'Phishing Website Detection Feature Extraction.ipynb' file. The results csv file is uploaded to this notebook and stored in the data frame
**Familiarizing with Data:**
In this step, few data frame methods are used to look into the data and its features.
**Visualizing Data:**

Few plots and graphs are displayed to find how the data is distributed and how features are related to each other.

**Data Preprocessing & EDA:**

Here, we clean the data by applying data preprocessing techniques and transform the data to use it in the models.

**Splitting the Data:**

Here, we split our data in two or more subset, for evaluate, test or train our data. Data splitting is an important aspect of data science especially for creating models.

**Machine Learning Models & Training:**

From the dataset above, it is clear that this is a supervised machine learning task. There are two major types of supervised machine learning problems, called classification and regression.This data set comes under classification problems, as the input URL is classified as phishing (1) or legitimate (0). The supervised machine learning models (classification) considered to train the dataset in this notebook are:

**Decision Tree Classifier:**

Decision trees are widely used models for classification and regression tasks. Essentially, they learn a hierarchy of if/else questions, leading to a decision. Learning a decision tree means learning the sequence of if/else questions that gets us to the true answer most quickly.

**Random Forest Classifier:**

Random forests for regression and classification are currently among the most widely used machine learning methods.A random forest is essentially a collection of decision trees, where each tree is slightly different from the others. The idea behind random forests is that each tree might do a relatively good job of predicting, but will likely overfit on part of the data.

**Multilayer Perceptrons** (MLPs):

Multilayer perceptrons (MLPs) are also known as (vanilla) feed-forward neural networks, or sometimes just neural networks. Multilayer perceptrons can be applied for both classification and regression problems.

### *3.2 Proposed Algorithm*

**XGBoost Classifier**

XGBoost is one of the most popular machine learning algorithms these days. XGBoost stands for eXtreme Gradient Boosting. Regardless of the type of prediction task at hand; regression or classification. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. Of processing to come to a decision.

**Q learning Algorithm on XGBoost**

The reinforcement learning approach has been utilized to gain proficiency for optimal behavior. This adaptive learning paradigm is defined as the problem of an “agent” to perform an action based on a “trial and error” basis through communications with an unknown “environment” which provides feedback in the form of numerical “rewards”

**Agent**: An agent learns the model state St by reading the input Xt, where t denotes the state transitions at time t. In the proposed model, the input to the agent will be the feature vector representation of a given URL.

**Action (U): The** actions influence the updates in the environment.

**State (S)**: At each time step to the state of the environment, the agent is interacting with, changes and affects the action taken by the agent. In this model, a state is determined by the input URL vector xt.

**Policy ($\pi$)**: an action pertaining to that state that maximizes the reward to be performed for that state

**Reward** <R>. The reward describes the immediate feedback from the environment, for an agent, for making the optimum action choice for that particular state.

**Discount factor ($\gamma$)**.It is defined to balance the performance of the agent, in a way, so that agent can make optimum choice of actions for both short term and long term rewards. The value of $\gamma$ ranges between 0 to 1.

## IV. CONCLUSION

Phishing detection is now an area of great interest among the researchers due to its significance in protecting privacy and providing security. There are many methods that perform phishing detection by classification of websites using trained machine learning models. URL based analysis increases the speed of detection. Furthermore, by applying feature selection algorithms and dimensionality reduction techniques, we can reduce the number of features and remove irrelevant data. There are many machine learning algorithms that perform classification with good performance measures. In this paper, we have done a study of the process of phishing detection and the phishing detection schemes in the recent research literature. This will serve as a guide for new researchers to understand the process and to develop more accurate phishing detection systems

## V. ACKNOWLEDGEMENT

We are very grateful and want to express our thanks to **Prof. D.B. Mane** for guiding us in the right manner, correcting our doubts by giving her time whenever we required, and providing her knowledge and experience in making this project work. We are also thankful to the HOD of our Information Technology department **Dr. M. L. Bangare** for his moral support and motivation which has encouraged us in making this project work. We are also thankful to our Principal **Prof. Dr. A.V. Deshpande**, who provided his constant support and motivation that made a significant contribution to the success of this project.

## REFERENCES:

[1]. ABDULGHANI ALI AHMED, NURUL AMIRAH ABDULLAH,”Real Time Detection of Phishing Websites”,IEEE,2021.

[2]” Detection of Phishing Websites by Using Machine Learning-Based URL Analysis”,IEEE, 01-03 July 2020.

[3]. Habiba Bouijij, Amine Berqia," Machine Learning Algorithms Evaluation for Phishing URLs Classification",IEEE, 2021.
[4].Muhammet Baykara, Zahit Ziya Gu¨rel,"Detection of phishing attacks ",IEEE 2020.
[5]. Vyacheslav Lyashenko,Oleg Kobylin,Mykyta Minenko."Tools for Investigating the Phishing Attacks Dynamics",IEEE,2018
[6]. Mahmoud Khonji, Youssef Iraqi," Phishing Detection using Machine learning ",IEEE,2018.
[7]. Arathi Krishna V*, Anusree A, Blessy Jose, Karthika Anilkumar," Phishing Detection using Machine learning",IEEE,2018.
[8]. Mr. B Ravi Raju , S Sai likhitha, N Deepa, S Sushma,"Phishing Websites Detection using Machine Learning ",IEEE, 2021.
[9]. A. Dabrowski, K. Krombholz, J. Ullrich, and E. R. Weippl, "QR inception: Barcode-in-barcode attacks," in Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones Mobile Devices". ACM, 2014
[10].Mahdieh Zabihimayvan and Derek Doran, "Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection",IEEE,2021.