International Journal of Scientific Research and Engineering Development-– Volume 5 Issue 2, Mar-Apr 2022

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Detection of Phishing Websites Using Machine Learning Algorithms

G.Mausam*, K.Siddhant**,S.Soham***,V.Naveen****

*(Student, Department of Computer Engineering,
Watumull Institute Of Electronic Engineering And Computer Technology, Ulhasnagar
Email: mausamgerra@gmail.com)

**(Student, Department of Computer Engineering,
Watumull Institute Of Electronic Engineering And Computer Technology, Ulhasnagar
Email: siddhantkhemlani@gmail.com)

***(Student, Department of Computer Engineering,
Watumull Institute Of Electronic Engineering And Computer Technology, Ulhasnagar
Email: sohamsawant2502@gmail.com)

****(Assistant Professor, Department of Computer Engineering,
Watumull Institute Of Electronic Engineering And Computer Technology, Ulhasnagar
Email: vaswani.naveen@gmail.com)

----------**********----------

Abstract:

Phishing attack is a simple way to obtain sensitive data from the users such as passwords, card credentials. The main aim of the phishers is to obtain such critical information and use them to fulfill their purpose. Cyber Security personnel are now looking for trustworthy and stable detection techniques for phishing website detection. This research paper deals with available machine learning technology for detection of Phishing URLs by extracting and then analyzing features of legitimate and phishing URLs. It includes comparison of various algorithms such as Random Forest Algorithm, KNN Algorithm and XGBoost algorithm. Aim of the paper is to detect phishing URLs as well as suggest the best machine learning algorithm based on the accuracy rate, false positive and false negative rate of each algorithm.

Keywords **—Phishing attack, Detection, Machine Learning, Random Forest, KNN, XGBoost.**

----------**********----------

I. INTRODUCTION

In today's world, Phishing has become a main area of concern for security researches because it is not difficult nowadays to create a fake website looking as good as a legitimate one. Experts can identify such fake websites but not a regular user and these are the ones who become a victim of such attacks. Main aim of the attacker is to acquire sensitive data such as passwords, credit card credentials, account details etc. These attacks are becoming successful because of lack of user awareness.

To overcome drawbacks of previously available methods, researchers are now focused on machine learning algorithms. The ability of algorithms to decide and predict based on learnings from past data will help in analysing various Blacklisted and legitimate URLs and their features to accurately detect phishing websites including zero-hour websites.

II. LITERATURE SURVEY

The tables below show the work done by the others authors which are useful and related to our work.

TABLE I
SUMMARY OF LITERATURE SURVEY FOR PHISHING DETECTION USING MACHINE LEARNING ALGORITHMS

S.No	Title	Year of Publication	Work Done in Brief
1	Detection of Phishing Websites using an Efficient Machine Learning Framework [14]	2020	Proposes a system model having classification and detection phase having multiple machine learning classifiers. The model trains the classifiers and selects the best classifier based on accuracy for future work. The comparison result show Random Forest Algorithm as the best classifier among 4 others.
2	A Machine Learning Approach to Improve the Efficiency of Fake Websites Detection Techniques [15]	2016	Proposes a system model having two parts namely Processing and Extraction part and Evaluation and Training part. The model extracts feature and trains the various machine learning algorithms to choose the one with best accuracy. The comparison results show that K-nearest neighbour algorithm was the best classifier among three others.
3	A Comparative Analysis Of Phishing Website Detection Using Xgboost Algorithm [16]	2019	Proposes a 4-phase model with Problem formulation, Dataset Processing and Feature extraction, Modelling using XGboost algorithm and Result analysis. The Proposed method uses XGBoost algorithm and is compared with other algorithms where it outperforms the other four algorithms.
4	Phishing Detection using Machine Learning based URL Analysis: A Survey [17]	2021	The paper lists various features that are extracted during the phases which further help to classify a website as phishing or legitimate website and mentions some performance evaluation metrics.
5	A Survey of Machine Learning-Based Solutions for Phishing Website Detection [18]	2021	A very detailed paper that provide a list of and explains in brief methodologies and machine learning based solutions for detection of phishing websites. It also clarifies the important steps for anti-phishing and compares various state of the art

			solutions based on their performance with their respective datasets. (Datasets are unique)

III. FEATURE EXTRACTION MODULE

We have implemented a python program to extract features from a URL which will be a user input. Below given are the list of features that we have extracted for detection of phishing URLs.

A. *Presence of IP Address in the URL:*

For suppose an IP address is present in the URL then the feature is set to 1 else it is set to 0 where 1 means true and 0 means false. Most sites don't use IP address in the URL.

B. *Presence of @ symbol in URL:*

For suppose an @ symbol is present in URL then the feature value is set to 1 else it is set to 0. Attackers add @ symbol in the URL which leads the browser to ignore everything preceding the @ symbol and real address is usually followed after it.[3]

C. *Presence of sensitive wordsin URL:*

Phishing websites use sensitive words in their URLs so that the users feel that they are on a legitimate website. Words such as – confirm, account, banking, password, username etc. are examples

D. *URL Redirection:*

Presence of "//" in the URL path sets the URL feature to 1 else sets it to 0. The existence of "//" in the URL means that the user will be redirected to another website.[3]

E. *HTTPS Token in URL:*

If the HTTPS token is present in the URL then the feature value is set to 1 else it is set to 0.Attackers may add the "HTTPS" token to the domain as part of a URL to trick users. For example, http://https-www.paytm-it-mps-home.hard-payemnt.com [3].

F. *URL Shortening or Lengthening services:*

TinyUrl or LongUrl services allow attackers to hide original URL by making it short or long. The goal is to redirect user to malicious websites. If URL is crafted using any servicethen feature value is set to 1 else 0.

G. *Number of Dots in the URL:*

Phishing URLs have many dots representing multiple sub domains for example https://shop.fun.flipkart.pass.com . In legitimate sites the average number of dots are 3. If the number of dots is more than 3 then feature value is set to 1 else its set to 0.

H. *Age of SSL Certificate:*

Presence of HTTPS token is very important and so is having SSL certificate with minimum age of 1 or 2 years. It gives an impression of website legitimacy to the user.

I. *Rank of the Website:*

Rank of the website is another important feature, legitimate websites have rank above 10,0000 so if the condition is satisfied the feature value is set to 1 else 0

J. *Presence of - symbol in URL:*

Phishing make use of – symbol in order to create fake website for example www.onlineamazon.com is a genuine website but www.online-amazon.com seems same but is a fake website hence if the URL has – symbol the feature value is set to 1 else 0

IV. MACHINE LEARNING ALGORITHMS

Three machine learning classification models namely Random Forest, K-Nearest Neighbour and XGBoost algorithms have been selected and compared further to detect phishing websites. The

most accurate algorithm will be used for future implementation.

I. RANDOM FOREST ALGORITHM

Random Forest algorithm is indeed a powerful algorithm in machine learning technology and it is based on concept of decision tree algorithm. Random forest algorithm creates a forest with number of decision trees. Higherthe number of treeshigher is the detection accuracy.

Trees are created based on bootstrap method. In bootstrap method the features and samples of dataset are randomly selected with a replacement to construct single tree. Among randomly selected features, random forest algorithm will choose best splitter for the classification and like decision tree algorithm; random forest algorithm makes use of gini index and information gain methods to find the best splitter. This process will continue until random forest creates n number of trees.

Each tree in the forest predicts the target value after which the algorithm will calculatevotes for each predicted target. Finally random forest algorithm considers the highest voted target as the final splitter.[5][6]

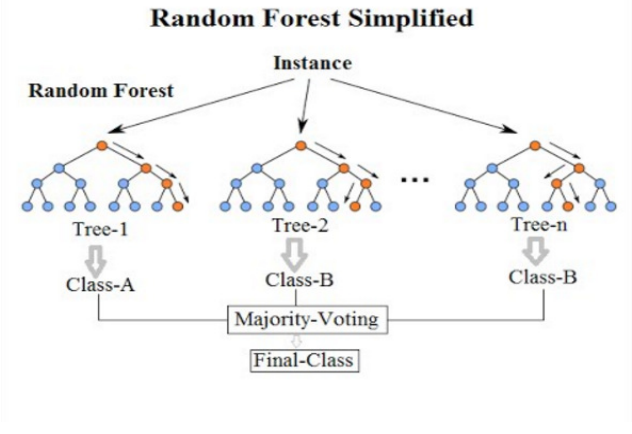

Fig. 1 Simplified pictorial representation of Random Forest Algorithm [11]

II. K-NEAREST NEIGHBOR

K-Nearest Neighbour (KNN) is a supervised Machine Learning model/ classifier. It is a direct and efficient model that can be applied to classification tasks. The K-NN model assumes that similar items can be found nearby. That is, it operates on the basic principle that "similar things are closer to each other." The distance can be calculated using a method known as "Euclidean distance."[7]

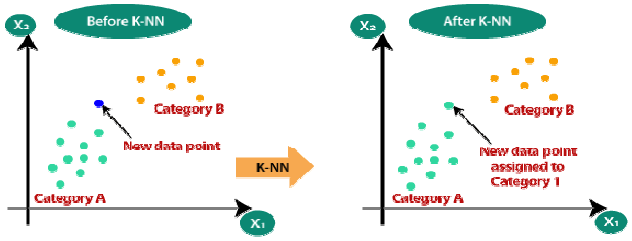

Fig. 2 Simplified pictorial representation of K-Nearest Neighbour Algorithm [12]

III. XGBOOST

XGBoost Extreme Gradient Boosted Tree is an optimized implementation of gradient boosted tree and is a recent supervised learning algorithm that implements process of boosting to improve performance of gradient boosted tree. It has many strengths when compared to traditional or basic gradient boosted trees. Some of its strengths are better regularization ability helping to reduce overfitting, higher speeds and performance because of parallel nature in which tress are built, flexibility, inbuilt routines for handling missing values. These advantages have made it a great tool of choice for many researchers in data science and machine learning.[8][9]

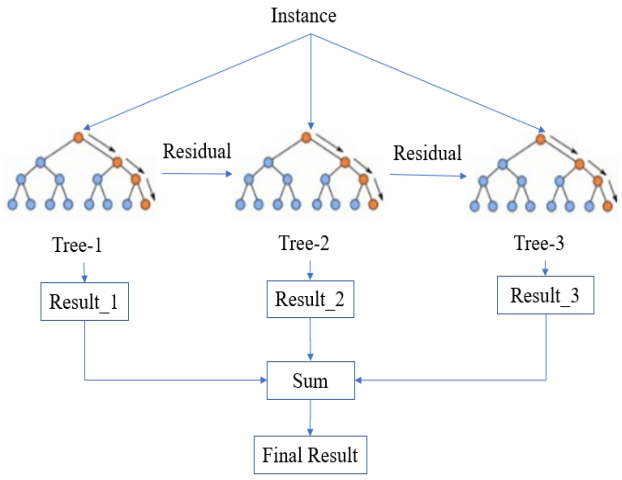

Fig. 3 Simplified pictorial representation of XGBoost Algorithm [13]

IV. DATASET - KAGGLE

Kaggle is a resource platform used by all data scientists, both experienced and aspiring. For anyone wishing to get started, contribute or helpin projects in order to enhance their skills or build up their data science portfolios. They provide 19,000 public datasets and 200,000 public notebooks. They provide Jupyter Notebooks environment which requires no setup and is fully customizable, as well as free GPUs and a massive archive of community-published data and code.[10]

V. IMPLEMENTATION AND RESULTS

The Machine Learning Algorithms were trained using both training and testing datasets which helps to evaluate the classifiers based on their precision and other characteristics.

TABLE II
CLASSIFIER'S PERFORMANCE

Machine Learning Algorithms	Precision (%)
Random Forest Algorithm	**96.759**
K-Nearest Neighbour	93.587
XGBoost	91.448

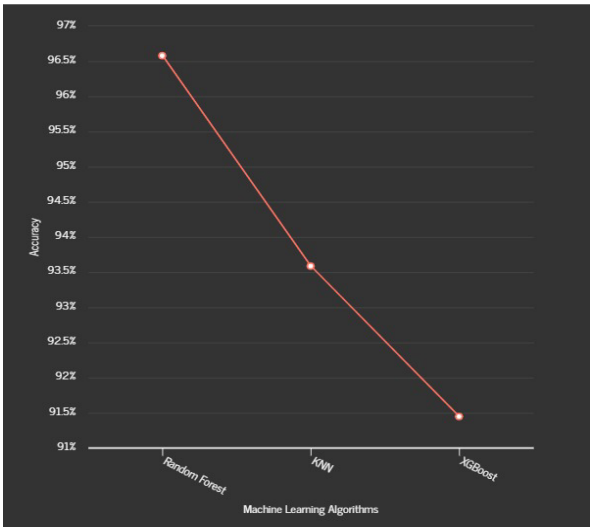

Fig. 4 Accuracy chart representation of various the classifiers used.

Results in Fig. 4 show that Random Forest algorithm gives better detection accuracy which is **96.759** which is better than both K-Nearest Neighbour and XGBoost algorithms.

VI. CONCULSION

This paper aims to enhance the detection method for detecting phishing websites using machine learning technology. We achieved 96.759% detection accuracy using Random Forest algorithm.

We aim to achieve a 97.5% or higher detection accuracy in future by using hybrid technology for which random forest algorithm and other available technology will be used in togetherness.

REFERENCES

[1] Gunter Ollmann, "The Phishing Guide Understanding & Preventing Phishing Attacks", IBMInternet Security Systems, 2007.

[2] https://resources.infosecinstitute.com/category/enterprise /phishing/the-phishing-landscape/phishing-data-attackstatistics/#gref

[3] Mohammad R., Thabtah F. McCluskey L., (2015) Phishing websites dataset. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016

[4] D. P. Yada, P. Paliwal, D. Kumar and R. Tripathi. "A Novel Ensemble Based Identification of Phishing E-Mails", Conference ICMLC, Singapore, February 24– 26, 2017, pp. 2-17.

[5]http://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/

[6] Predrag Radenković. "Random forests" Faculty of Electrical Engineering, University Of Belgrade, 3237/10, 2010.

[7]https://dataaspirant.com/k-nearest-neighbor-classifier-intro/

[8] Jain A. (2016). Complete Guide to Parameter Tuning in XGBOOST (with code in python) Retrieved from https:complete guide to parameter tuning in XGBOOST (with code in Python). 2017/06/13.

[9] A. Gómez-Ríos, J. Luengo, and F. Herrera. "A Study on the Noise Label Influence in Boosting Algorithms: AdaBoost, GBM and XGBOOST". In International Conference on Hybrid Artificial Intelligence Systems, Springer, Cham, June 2017, pp. 268-280.

[10] https://www.kaggle.com/datasets

[11] https://en.wikipedia.org/wiki/Random_forest

[12]https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[13]https://www.researchgate.net/figure/Simplified-structure-of-XGBoost_fig2_348025909

[14] Naresh Kumar D, Nemala Sai Rama Hemanth, Premnath S, Nishanth Kumar V Dept. of IT Panimalar Engineering College Chennai, India.

"Detection of Phishing Websites using an Efficient Machine Learning Framework"

[15] Kruti Lavingia Nirma University, Anuj Dakwala Tech Mahindra. "A Machine learning approach to improve the efficiency of Fake websites detection Techniques"

[16] Musa Hajara, Aishatu Yahaya Umar - Gombe State University, Fatima Umar Zambuk Abubakar Tafawa Balewa University, Jamilu Usman Waziri. "A comparative analysis of phishing website detection using XGBOOST algorithm"

[17] Arathi Krishna V, Anusree A, Blessy Jose, Karthika Anilkumar, Ojus Thomas Lee Department of Computer Science and Engineering, College of Engineering Kidangoor Kottayam, India. "Phishing Detection using Machine Learning based URL Analysis: A Survey"

[18]Lizhen Tang * and Qusay H. Mahmoud Department of Electrical, Computer, and Software Engineering, Ontario Tech University, Oshawa, ON L1G 0C5, Canada. "A Survey of Machine Learning-Based Solutions for Phishing Website Detection"

