ISSN 2347 - 3983
Volume 11. No.1, January 2023
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter031112023.pdf
https://doi.org/10.30534/ijeter/2023/031112023

Phishing Website Detection Using Ensemble Learning

Aditya Soni[1], Jyoti Tiwari[2]
[1] M.E. Computer Engineering Sgsits Indore, India, adityasoni195@gmail.com
[2] Assistant Professor Computer Science & Engineering Department Sgsits Indore, India, jyotimona23@gmail.com

Received Date: December 7, 2022 Accepted Date: December 29, 2022 Published Date : January 07, 2023

ABSTRACT

Phishing is also the most common type of data breach. As a result, it is carried out by sending an email with links that lead to fraudulent websites. This technique is especially targeted to large companies. Usually, the attackers send emails with work-related information. Machine learning is one of the most successful techniques for detecting phishing. This paper analyzed the results of various machine learning techniques for predicting phishing websites. And also describes the various methods that are used to identify phishing websites. Some of these include the SVM classification method, Random Forest method, and AdaBoost method. Ensemble model that combines the SVM, Random Forest, and AdaBoost methods was able to classify a phishing site with an accuracy of 96%.

Key words : Phishing, Phishing attacks, Machine Learning, Random Forest, SVM, Adaboost, Ensemble learning.

1. INTRODUCTION

Securing the financial services offered by the Internet has made people's lives easier. However, maintaining the security of these services is a top priority. One of the most common ways to get access to a website is through phishing. This technique involves tricking specific individuals or businesses into providing sensitive information.

Phishing can result in significant damages to a website. According to a study conducted by Microsoft, an average of USD 5 billion is lost due to phishing attacks[1]. Similarly, the Internal Revenue Service (IRS) has issued a warning about an increase in phishing attacks, claiming a 400 percent increase in reported incidents [2]. Several methods have been proposed to combat phishing, ranging from online user education to enhanced phishing detection systems.

The conventional technique of phishing detection has failed due to the complex and dynamic nature of phishing attacks. The Anti-Phishing Working Group claims that (APWG), 239,910 different phishing reports were reported in 2018 [4]. Over the previous high point in June 2016 was 211,032 [3], the number of reports submitted increased by 12 percent. Despite taking precautions to avoid phishing, this happened. Further investigation revealed that each phishing attempt was unique from the others.

Machine learning algorithms can learn from massive amounts of data and can detect new patterns in phishing attacks. They can also adjust their behavior to prevent getting overwhelmed by the influx of new phishing attacks. The goal is to find the best machine learning algorithm that can detect the most common types of phishing attacks. The report also describes the various techniques that were tried in the past.

2. Literature Survey

"Detecting Phishing Websites Using Machine Learning".This is proposed by Amani Alswailem, BashayrAlabdullah, Norah Alrumayh and Aram Alsedrani [5], they have used dataset from kaggle and the website's features utilising the URL and the Document ObjectModel (DOM) object The URL that was utilised to extract the characteristics of the URL and page rank. While the DOM is used to extract the characteristics of a content page, it is a link between scripts and website pages that have a logical structure of documents and allow programmers to access and manipulate the DOM file.

"Url phishing data analysis and detecting phishing attacks using machine learning in nlpit ". This is proposed by Dr. G. Ravi Kumar, S. Gunasekaran, Nivetha R, Sangeetha Prabha K, Shanthini G and Vignesh A. S [6], they have used dataset from kaggle and the number of variables in the dataset will be reduced to those with the most discriminating information, and the features will be retrieved from the URL. Length of URL, IP Address, Subdomain, HTTPs Symbols, Website traffic, and Dots are some of the factors.

"Phishing Website Detection based on Supervised Machine Learning with Wrappers Features Selection". This is proposed by Waleed Ali [7], they have used dataset from kaggle and wrapper-based assessment and filter-based evaluation are the two primary types of feature evaluation. The significant characteristics are picked based on statistical methods to assess and balance the features without categorization information in filter-based evaluation approaches. The key characteristics are selected using filter-based assessment approaches, which have a strong reliance on the target class and little inter-correlation.

"Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code". This is proposed by Mona Ghotaish Alkhozae and Omar Abdullah Batarfi [8], in this paper the make web based phishing percentage detection system. They not compared with any other methods and classifiers. This proposed method they check website is phishing or not do it physically no automation system are there.

"Detection of Phishing Attacks". This is proposed by MuhammetBaykara and ZahitZiyaG ̈urel [9], they have used dataset from kaggle and in this paper the approach that is taken to solve the problem is through intrusion detection systems. They compared with Bayes classifier pervious result and also compare detection capacity and this proposed method they check website is phishing or not do it physically no automation system are there.

"Phishing Websites Detection through Supervised Learning Networks". This is proposed by Priyanka Singh, Yogendra P.S. Maravi and Sanjeev Sharma[10], they have used dataset URL from kaggle and in this paper the approach that is taken to solve the problem is through Neural Network. They compared with classifier pervious result and also compare detection capacity and this proposed method they check website is phishing or not do it physically.

"Phishing Detection from URLs by using Neural Networks". This is proposed by OzgurKoraySahingoz, SaideI ̧silay Baykal and Deniz Bulut [11], they have used dataset URL from kaggle and in this paper the approach that is taken to solve the problem is through Neural Network. They compared with classifier pervious result and also compare detection capacity and this proposed method they check website is phishing or not do it physically.

"A Phishing Detection System Based on Machine Learning". This is proposed by Che-Yu Wu, Cheng-Chung Kuo and Chu-Sing Yang [12], they have used dataset from kaggle and in this paper the approach that is taken to solve the problem is through SVM .They do not compared to other classifiers.

S.Aarthi et al.[13] To evaluate website URLs, the suggested system employs the URL Mining method. The system is separated into three modules: classifier, feature extraction, and feature analyzer, as shown. The user accesses the webpage in the classifier module, and the system then analyses it. The suspicious URL properties including length, address, and time are extracted in the feature extraction module.

3. DATASET

Kaggle provided the dataset. There are 95910 URLs in all, with 55914 phishing and 39996 safe occurrences. There are 11 characteristics in each instance. The outcome can be either 1 (phishing) or 0 (not phishing). Every column contains a feature and has the values 1 or 0. '1' if the URL is thoroughly scammed, '0' if the URL is not scammed. The following characteristics are included in the dataset: IP address (malicious IP addresses include additional characters or differ in some letters from the original, for example: www.129.B7.fake.html), Long and Short URL (long and short URLs might suggest a phished website), usage of @ symbol, '//' sign can be an indicator of redirection.

4. PROPOSED APPROACH

The three methods used to identify phishing websites were derived from the Kaggle dataset. The dataset consists of various parameters such as IP address, domain registration length, and request URL. The code used in the project was compiled using the libray "sklearn" library and implemented various algorithms. To categorise websites as phishing or authentic, the properties of URL are extracted and Random Forest classification technique, SVM classification method, and Adaboost classification algorithm are used. To increase the detection of phishing scams, we have combined all three methodologies with the ensemble approach. System Architecture is shown in Figure 1.

4.1 Figure

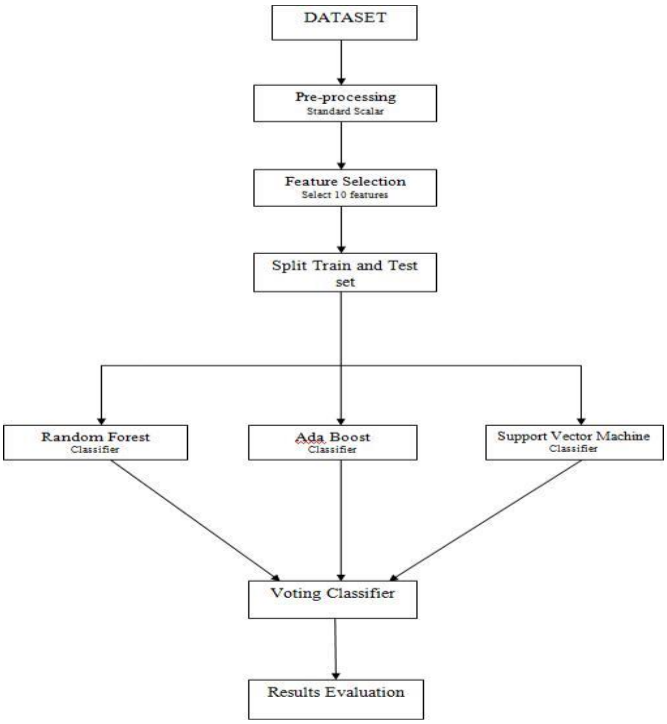

Figure 1: System Architecture

5. MACHINE LEARNING ALGORITHM

Random Forest, SVM, Adaboost and Ensemble learning are the algorithms used. The dataset is first pre-processed and features extracted, after which it is passed to these technique.

5.1 Random Forest

Random forest is a machine learning algorithm that learns and predicts the outputs from various decision trees generated by the training data. It is suitable for various

applications. A bagging technique is used to reduce the overfitting of the model. It achieves better model accuracy than other models. The Random Forest Classifier module is implemented using the sklearn Random Forest module. It is not constant since its results vary from time to time.

5.2 Support vector machine(SVM)

The goal of the SVM is to find the best line or choice cutoff which can separate two dimensional space from classes(here: phished or safe website). The classifier sets safe and phished websites on any of the planes after training with the dataset, thus categorising the websites. The kernel mode was used to create this classifier using sklearns' svm.SVC().

5.3 AdaBoost

Like Random Forest, Adaboost combines weak classification models with strong ones to create a powerful classifier. Doing so can improve overall classification since one model can easily classify an item. After learning about multiple predictions, the trees are generated as weak learners that are corrected by allocating a larger weight to them. The accuracy that was obtained from the previous training sessions is used to determine the weights of the classifiers taught in each iteration.

6. EXPERIMENTAL RESULTS

The goal of this study is to develop a set of advanced machine learning models that can detect phishing attacks. The models have been trained with varying outputs and are expected to detect phishing attacks. Precision, recall, f1 score, and accuracy are all included in the classification report for Ensemble model. Random Forest, SVM, and Adboost algorithms achieve accuracy of 95 %, 88 %, and 94 % respectively shown in Table 1.

6.1 Table 1: Accuracies of Machine Learning Algorithms

Algorithm	Accuracy	Precision	Recall	F-score
Ensemble learning model	96%	94.50%	96%	95%
Random forest	95%	95%	94%	90%
Support vector machine	88%	90%	89%	85%
Ada Boost	94%	93%	95%	92%

7. CONCLUSION

The increasing number of phishing attacks has become worse over the years. This project's goal was to develop an algorithm that would detect phishing. This projects helps in detecting phishing website thus helping avoiding the confidential information. This also helps detecting phishing early so that it can be handle and remove at early stage. The proposed approach is implemented with the help of Python programming language in spyder IDE for detecting Phishing website the system uses a random forest, svm, adaboost algorithm.

REFERENCES

1. Hewage, Chaminda Nawaf, Liqaa Khan, Imtiaz Alkhalil, Zainab. (2021). **Phishing Attacks: A Recent Comprehensive Study and a New Anatomy**. Frontiersin Computer Science. 3. 10.3389/fcomp.2021.563060.

2. Ms. Sophiya Shikalgar , Dr. S. D. Sawarkar , Mrs. Swati Narwane, 2019, **Detection of URL based Phishing Attacks using Machine Learning**, international journal of engineering research technology (IJERT) Volume 08, Issue 11 (November 2019)

3. Anti-Phishing Working Group (2016). **Phishing Activity Trends Report** (4th Quarter 2016). Unifying the Global Response To Cybercrime. [online] APWG.

4. Anti-Phishing Working Group (2018). **Phishing Activity Trends Report** (4th Quarter 2018). Unifying the Global Response To Cybercrime

5. Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh,Aram Alsedrani, **Detecting Phishing Websites Using Machine Learning** ,2nd International Conference on Computer Applications Information Security 2019.

6. Dr. G. Ravi Kumar, Dr. S. Gunasekaran, Nivetha R., Sangeetha Prabha K, Shanthini G., Vignesh A. S., **URL Phishing Data Analysis and Detecting Phishing Attacks using Machine Learning in NLP**, International Journal of Engineering Applied Sciences and Technology(IJEAST) Vol. 3, Issue 8, ISSN No.2455-2143, Pages 70-75, Published: December 2018

7. Waleed Ali **Phishing Website Detection based on Supervised Machine Learning with Wrappers Features Selection**, IJACSA (International Journal of Advanced Computer Science and Applications, Vol. 8 No. 9, Issue:2017

8. Mona Ghotaish Alkhozae,Omar Abdullah Batarfi **Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code**, International Journal of Information and Communication Technology Research, Volume 1 No. 6, October 2011

9. Muhammet baykara, zahit ziya g¨urel **Detection of phishing attacks**, 6th international symposium on digital forensic and security (isdfs), 22-25 march, 2018

10. Priyanka Singh, Yogendra P.S. Maravi , Sanjeev Sharma, **Phishing Websites Detection through Supervised Learning Networks**, International Conference on Computing and Communications Technologies (ICCCT) 2015. 41

11. Ozgur koray sahingoz, saide i silay, baykal and Deniz bulut, **Phishing detection from urls by using Neural networks** ,international conference on Computer science engineering and applications, 2018

12. Che-yu wu, cheng-chung kuo , chu-sing yang, **A phishing detection system based on machine learning,** International conference on intelligent computing and Its emerging applications (icea) 2019

13. S.aarthi , narsepalli vamsi kishan , v.surya teja, n.v.harsha vardhan gupta: **Classification of phishing website based on url features** : international journal Of emerging technologies in engineering research (ijeter) may (2019).