International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

Detecting Phishing Websites: A Machine Learning Approach

Prof. Leena I. Sakri[1], Madhuri Kulkarni[2], Priyanka Kamath[3], Shreedevi Bhat[4], Swati Kamat[5]

Department of Information Science and Engineering, SDM College of Engineering and Technology, Dharwad

---

Abstract - Phishing websites are forged websites that are created by malicious people to seem as a true website. Phishing attacks has to be prevented. One of the ways is by detecting the websites and creating awareness to users to spot the phishing websites. To detect the phishing websites there have been different methods carried out. Machine Learning algorithms have proven to be efficient. Machine Learning algorithms like Ada Boost, Random Forest and Support Vector Machine are proposed for detecting phishing websites in this paper. Graphical User Interface has been created for the algorithm having the very best accuracy.

Key Words: Machine Learning, Ada Boost, Random Forest, Support Vector Machine, accuracy.

1. INTRODUCTION

There are over 1.7 billion websites exist, till date. An internet site may be a collection of sites, images, videos of related data which provides information. There are different types of websites like a personal website, corporate website for company, a government website, an organization website, etc. Corporate, government and organization websites ask the users to enter their personal information for conformation before using their websites. As online technology is evolving there are online advertisements, online shopping, online courses, online etc. Many of which ask for the user for personal information. As online financial activities are raising, online fraudulent is also increasing in which phishing is a major concern. Phishing website is a website which creates the replica of the legitimate website with very few changes so that user cannot identify, because of which the user falls into the trap of the phishing attackers and give away their personal information such as bank account number, social media account details, username, password etc. There are nearly 1.4 million phishing websites created monthly, according to the Webroot Quarterly Threat Trends Report. Developing measures against phishing is a challenging problem because victims help criminals in giving away their data. Besides that users generally have a scarcity of attention to security. The phenomenon of phishing could also be unknown. Users don't see privacy and security as their primary task. Many approaches have been proposed by different authors for detection of phishing websites.

2. EXISTING SYSTEM

Despite growing efforts to educate users and create better detection tools, users are still very susceptible to phishing attacks. There are several promising defending approaches to this problem reported earlier.

2.1 Novel Neural Network: This system [1] divided the dataset into training and testing dataset. From the training dataset is analyzed for the URL and exact features for which the novel neural network will be applied using design risk minimization and Mante Carlo. From the testing dataset the exact features will be collected for which the phishing data model will be built. When the user enters the URL, the URL will be tested and it will be predicted as benign or phishing by the system. The paper utilizes concealed layer and yield layer for the cycle. This paper has an exactness of 97.71 %.

2.2 Extreme Learning Machine: This paper [2] considered 30 attributes from the input dataset. Input dataset can take 0, -1, and 1 as its values. Results of the output dataset obtained during this period may take two different values. An online site is taken under consideration as valid, dubious or spoofing within the obtained rules for input attributes of the dataset. A categorization is finalized and the output is within the type of spoofing or genuine website which is within the dataset. Through the analysis, it is seen that values such as 1, 0, -1 has been utilized for genuine, suspicions and spoofing websites respectively. From the above work, 95.05% was the classification accuracy and hence the absolute best classification exactness was estimated as 95.93

2.3 AI Meta-Learners and Extra-Trees Algorithm: The paper[3] has extended four meta learners models: AdaBoost-Extra Tree (ABET), Bagging – Extra tree (BET), Rotation Forest – Extra Tree (RoFBET) and LogitBoostExtra Tree (LBET) then created using the extra-tree base classifier. These proposed AIput together meta-learners was fitted with respect to the phishing site datasets and their exhibitions were thus assessed. A third approach is to visually differentiate the phishing sites from the spoofed legitimate sites. It proposes to use a randomly generated visual hash to customize the browser window or web form elements to indicate the successfully authenticated sites. All the created models of this examination work accomplished precision above 96%.

3. PROPOSED SYSTEM

The purpose of our project is to improve the efficiency in detection of phishing websites using machine learning techniques. A lot of research work and survey was done to compare various machine learning algorithms to best fit our model. Lots of journals and papers had been read and surveyed to decide upon the effective algorithms with higher accuracy. The idea that we are putting forward here is to improve the efficiency by using Machine Learning algorithms like Ada Boost, Random Forest and Support Vector Machine. The algorithm with higher accuracy is then used in developing the graphical user interface. The user has to input a URL to check whether the entered URL is legitimate or phishing. Once the user enters the URL and enters the check option, the result will be displayed whether the entered URL is phishing website or legitimate website.

4. METHODOLOGY

For the detection of phishing websites, this project has made use of Machine Learning approaches. Below are the different steps involved in the development of the model.

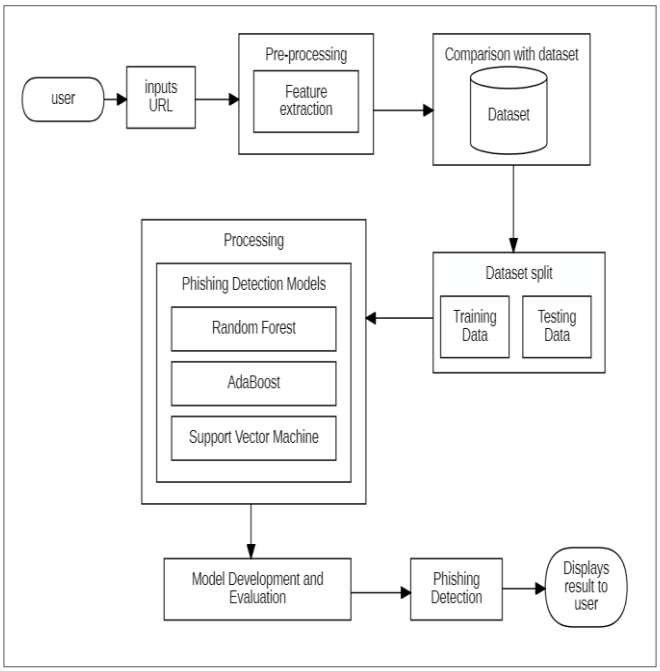

Fig -1: Architectural Design

4.1 Data Collection

The dataset has been collected from the phishtank.com website. The dataset consists of 31 attributes out of which 30 attributes are used for the data analysis and 1 attribute is used for the result prediction. There are 11,065 rows in the dataset. The attributes have 3 values i.e. 0,1 and -1.The attribute having values 0 indicates that the attribute value is suspicious. The attribute having values 1 indicates that the attribute value is legitimate and -1 indicates that it is a phishing attribute.

4.2 Feature Extraction

When a URL is entered, the features present in the URL are extracted. The code is written to extract the features such for calculating the length of URL, presence IP address, symbols like @, positioning of // etc. The length of the URL has to be less than 54 for to be legitimate website. If it is between 54 to 75, it is considered to be suspicious and above 75 is considered to be phishing attribute. Accordingly, values are assigned as 1, 0, -1 for legitimate suspicious and phishing attributes. If there is a presence of @ symbol in the domain name of the URL then it is considered phishing and -1 value is assigned. Presence of IP address in the URL is considered to be phishing attributes so is the value -1 assigned. In this way the features are extracted and values are assigned.
After the feature extraction, it is compared with the dataset.

4.3 Dataset split

Splitting the dataset is very essential for an unbiased evaluation of prediction performance. The dataset is split into training dataset and testing dataset. The dataset is splitted as 60% training and 40% testing. The training set was used to fit the model and validation for the evaluation.

4.4 Phishing Detection Models

1. Random Forest:

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. If we build many trees, all of which work well and overfit in different ways, which can reduce the amount of overfitting by averaging their results. To build a Random Forest model, the number of trees to be built has to be decided. They are very powerful, often work well without heavy tuning of the parameters, and don't require scaling of the data.
The Formula used here is

$$C^B_{rf} = \text{majority vote } \{C_b(x)\}^B_1$$

Where,
B - number of Decision Trees created (B=100)
C^B_{rf} - Random Forest class prediction for B Decision Trees
$C_b(x)$ -the class prediction of the bth Tree.

The number of trees created in this project is 100 by assigning n_estimators=100 in the RandomForestClassifier.

2. Ada Boost:

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique that is used as an Ensemble Method in Machine Learning. AdaBoost, short for Adaptive Boosting, is a Machine Learning meta-algorithm which can be used in conjunction with many other types of learning algorithms to improve performance. It is called Adaptive Boosting as the weights are reassigned to each instance, with higher weights to incorrectly classified instances. Boosting is used to reduce bias as well as the variance for supervised learning. It works on the principle where learners are grown sequentially. Except for the first, each subsequent learner is grown from previously grown learners. In simple words, weak learners are converted into strong ones. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The outputs are combined by using a weighted sum, which represents the combined output of the boosted classifier, i.e.

$$F_T(x) = \sum_{t=1}^{T} f_t(x)$$

Where,

f_t -weak learner that returns the predicted class with respect to input x.

T - Number of interations (t = 1 to 50)

$F_T(x)$ - Strong classifier .

In this project the value of T is taken as 50 by assigning n_estimators=50 in AdaBoostClassifier.

3. Support Vector Machine

Support Vector Machine or SVM is one of the most popular supervised learning algorithms, which is used for classification as well as regression problems. However, primarily, it is used for classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. In Machine Learning, Support-Vector Machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. This classifier uses a nonlinear mapping to transform original training data into a higher dimension and finds hyper planes that partition data samples in the higher dimensional feature space. The separating hyper planes are defined as

$$Wx+b=0$$

Where,

W - coefficient of x

b - Constant

In this project the linear kernel is being used with degree 3.

4.5 Comparison

The 3 algorithms are implemented and the accuracies are compared.

Table-1: Comparison of accuracies

SL no	Algorithm	Accuracy
1.	Random Forest	96.83%
2.	Ada Boost	92.08%
3.	Support Vector Machine	93.08

After the implementation we found that the Random Forest Algorithm has highest accuracy. So, the Random Forest algorithm is used in developing the graphical user interface of phishing websites detection system.

4.6 Graphical User Interface

A graphical user interface is user friendly interface between the user and system. The graphical user interface in phishing websites detection system has been implemented using PyQt. The framework used to develop the GUI is PyQt. PyQt is a GUI widgets toolkit. It is a Python interface for Qt, one of the most powerful, and popular cross-platform GUI library. PyQt5 is implemented as a set of Python modules. It has over 620 classes and 6000 functions and methods.

The user can communicate with the system easily via this interface. The user has to input a URL to check whether the entered URL is legitimate or phishing. Once the user enters the URL and enters the check option, the result will be displayed whether the entered URL is phishing website or legitimate website.

5. RESULTS

5.1 Implementation of Machine Learning Algorithms

1. Random Forest Algorithm

```
array([[1896,   95],
       [  52, 2379]], dtype=int64)
```

Fig -2: Confusion Matrix for Random Forest algorithm

The Confusion Matrix values are:

True Positive (TP)-1896

False Positive (FP)-95

False Negative (FN)-52

True Positive (TP)-2379

```
print('The accuracy of Random Forest Classifier is: ' , 100.0 * accuracy_score(rfc_predict,test_Y))

The accuracy of Random Forest Classifier is:  96.54002713704206
```

Fig -3: Random Forest algorithm accuracy

2. AdaBoost Algorithm:

```
array([[1813,  178],
       [ 130, 2301]], dtype=int64)
```

Fig -4: Confusion Matrix for Random Forest algorithm

The Confusion Matrix values are:
True Positive (TP)-1813
False Positive (FP)-178
False Negative (FN)-130
True Positive (TP)-2301

```
print('The accuracy of Ada Boost Classifier is: ', 100.0 * accuracy_score(adc_predict,test_Y))

The accuracy of Ada Boost Classifier is:  91.42921754862053
```

Fig -5: AdaBoost algorithm accuracy

3. SVM Algorithm:

```
array([[1779,  212],
       [ 138, 2293]], dtype=int64)
```

Fig -6: Confusion Matrix for SVM algorithm

The Confusion Matrix values are:
True Positive (TP)-1779
False Positive (FP)-212
False Negative (FN)-138
True Positive (TP)-2293

```
print("The Accuracy of SVM Classifier is {}".format(100* acc_train_svm))

The Accuracy of SVM Classifier is 92.96698326549074
```

Fig -7: SVM algorithm accuracy

5.2 GUI snapshots

Fig -8: GUI when user enters a URL

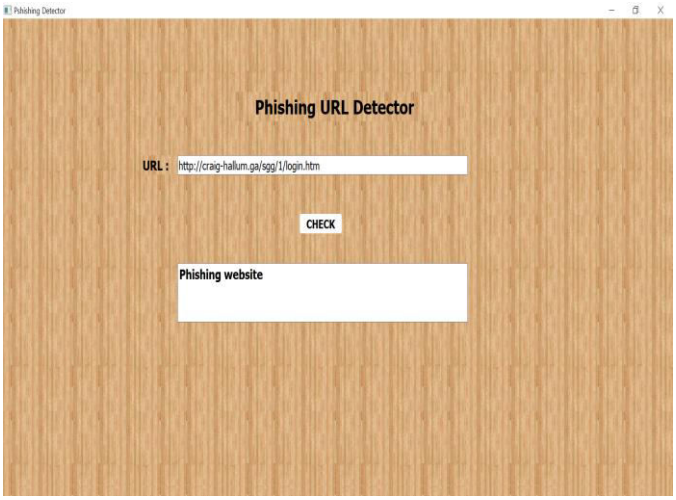

Fig -9: GUI when showing Phishing website

Fig -10: GUI showing legitimate website

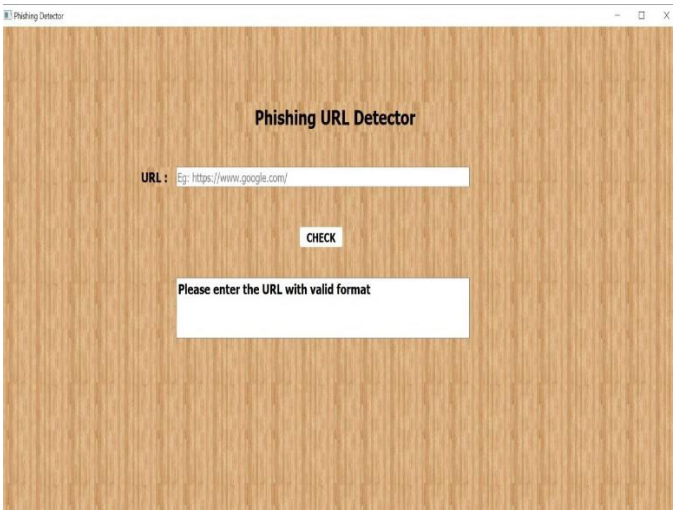

Fig -11: GUI when no URL entered

3. CONCLUSION

The phishing websites detection system has been successfully implemented using the efficient Machine Learning algorithms. The user can communicate with the system easily via the graphical user interface. The phishing websites detection system successfully detects whether the website is phishing or legitimate. This system will present a combined approach to be applied on phishing data to establish a unique and effective solution towards developing smart models to identify phishing websites. This system will offer security for the user that highly helps to achieve better transaction needed for the users, in their own interest. The phishing websites detection system can be further extended to creation of browser extension. A combination of any other two or more classifiers can be used to get maximum accuracy. Various phishing techniques that uses Lexical features, Network based features of web pages can be taken into consideration which can improve the performance of the system. The phishing detection system can be built as a scalable web service which will incorporate online learning so that new phishing attack patterns can easily be learned.

ACKNOWLEDGEMENT

We have been bestowed the privilege of expressing our gratitude to everyone who helped us in completing the project. We sincerely thank, Prof. Leena I Sakri, Department of Information Science Engineering, SDMCET, our mentor for the project who helped us throughout in carrying out the project. Finally, we thank one and all who have directly and indirectly assisted us in the project work.

REFERENCES

1. FangFeng, Qingguo Zhou, Zebang Shen,XuhuiYang,Lihong Han,JinQiang Wang“The application of a novel neural network in the detection of phishingwebsites”, 2018.
2. Mustafa Kaytan, Davut Hanbay “Effective Classification of Phishing Web Pages Based on New Rules by Using Extreme Learning Machines, 2017.
3. Yazan A. Al-Sariera, Victor Elijah Adeyemo, Abdullateef O. Balogunand Ammar K. Alazzawi “AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites”, 2017.
4. Hassan Y.A.Abutair and Abdelfettah Belghith “Use Case Based Reasoning for Phishing Detection”, 2017.
5. Yazan Ahmad Alseriera, Adeyemo Victor Elijah, Abdullateef O.Balogun”Phishing Website Detection: Forest by Penalizing Attributes Algorithm and Its Enhanced Variations”, 2020.
6. Alyssa Anne Ubing, Syukrina Kamilia Binti Jasmi , Azween Abdullah , NZ Jhanjhi4 , Mahadevan Supramaniam “Phishing Website Detection: An Improved Accuracy through Feature Selection and Ensemble Learning”2019
7. Asma Al Sarhan, Riad Jabri, Ahmad Sharieh “Website Phishing Detection Using Dom-Tree Structure and Cant-MinerPB Algorithm “,2017
8. Vaibhav Patil, Pritesh Thakkar, Chirag Shah , Tushar Bhat, Prof. S. P. Godse “Detection and Prevention of Phishing Websites using Machine Learning Approach”
9. Pierrick Robic–Butez, Thu Yein Win. ”Detection of Phishing websites using Generative Adversarial Network”, 2019
10. Shraddha Parekh, Dhwanil Parikh, “A new method for Detection of Phishing Websites: URL Detection”, 2018
11. Muhammad Taseer Suleman and Shahid Mahmood Awan ”Optimization of URL-Based Phishing Websites Detection through Genetic Algorithms”, 2019
12. M Somesha, Alwyn Roshan Pais, Routhu Srinivasa Rao and Vikram Singh Rathour” Efficient deep learning techniques for the detection of phishing websites”, 2020
13. Divya James, “An Innovative Framework for the Detection and Prediction of Phishing Websites”, 2018
14. Kahksha, Sameena Naaz,”Detection of Phishing Websites using Machine Learning Approach”, 2019
15. J. James, L. Sandhya, C. Thomas “Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection”,2019.
16. Tan, C.L., Chiew, K.L., Wong, K., Sze,An “Enhanced Blacklist Method to Detect Phishing Websites” 2016.
17. Zhang W, Jiang Q, Chen L, Li C “A machine learning based approach for phishing detection using hyperlinks information”, 2017.
18. V. S. Lakshmi, M. S. Vijaya,” Detection of Phishing Websites Based on Probabilistic Neural Networks and K-Medoids Clustering”, 2017.
19. W. Zhuang, Y. Ye, T. Li, Q. Jiang “An Intelligent Anti-phishing Strategy Model for Phishing Website Detection”, 2017.
20. M. Aburrous, M. A. Hossain, K. Dahal, F. Thabtah, “A new method for Detection Of Phishing Websites URL Detection”, 2016.

21. Waleed Ali , Sharaf Malebary “Particle Swarm Optimization-Based Feature Weighting for Improving Intelligent Phishing Website Detection”, 2020
22. Charu Singh, Smt.Meenu “Phishing Website Detection Based on Machine Learning”,2020.
23. Rahul Anoop Kelkar, Vijayalakshmi A “ML Based Model for Phishing Website Detection”, 2020.
24. Ankit Kumar Jain , B. B. Gupta “Phishing Detection: Analysis of Visual Similarity Based Approaches”,2017.
25. Robat Das, Md. Mukhter Hossain, Shariful Islam, Abujarr Siddiki “Learning a Deep Neural Network for Predicting Phishing Website”, 2019
26. Hemali Sampat, Manisha Saharkar, Ajay Pandey , Hezal Lopes “ Detection of Phishing Website Using Machine Learning”, 2018.
27. Andrew J. Park, Ruhi Naaz Quadari, Herbert H. Tsang “Phishing Website Detection Framework Through Web Scraping and Data Mining”, 2017.
28. Gururaj Harinahalli Lokesh, Goutham BoreGowda “Phishing website detection based on effective machine learning approach”, 2020.
29. Mahajan Mayuri Vilas, Kakade Prachi Ghansham, Sawant Purva Jaypralash, Pawar Shila “Detection of Phishing Website Using Machine Learning Approach”,2019.
30. Peng Yang, Guangzhen Zhao, Peng Zeng “ Phishing Website Detection based on Multidimensional Features driven by Deep Learning ”,2019.
31. Kuldeep Randhawa, Chu Kiong Loo, (Senior Member, IEEE), Manjeevan Seera , (Senior Member, IEEE), Chee Peng Lim , and Asoke K. Nandi, (Fellow, IEEE) “ Credit Card Fraud Detection Using AdaBoost and Majority Voting” March 28, 2018
32. John Arthur Jupin , Tole Sutikno , Mohd Arfian Ismail, Mohd Saberi Mohamad , Shahreen Kasim , Deris Stiawan,” John Arthur Jupin1 , Tole Sutikno2 , Mohd Arfian Ismail3 , Mohd Saberi Mohamad4 , Shahreen Kasim5 , Deris Stiawan6”
33. Sharifi, M. and Siadati “Phishing Websites Detection through Supervised Learning Networks”, 2018.
34. C. Konradt, A. Schilling and B. Werners, "Phishing: An economic analysis of cybercrime perpetrators", Computers & Security, 58, pp.39-46, 2016.
35. R. Jabri and B. Ibrahim, "Phishing Websites Detection Using Data Minin Classification," Transactions on Machine Learning and Artificial Intelligence, vol. 3, no. 4, 2015.
36. Waleed Ali , Sharaf Malebary “Particle Swarm Optimization-Based Feature Weighting for Improving Intelligent Phishing Website Detection”, 2020
37. Charu Singh, Smt.Meenu “Phishing Website Detection Based on Machine Learning”,2020
38. Rahul Anoop Kelkar, Vijayalakshmi A “ML Based Model for Phishing Website Detection”, 2020
39. Ankit Kumar Jain, B. B. Gupta “Phishing Detection: Analysis of Visual Similarity Based Approaches”,2017
40. Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., &NunamakerJr, J. F, "Detecting fake websites: the contribution of statistical learning theory," Mis Quarterly, pp. 435-461, 2010.
41. Robat Das, Md. Mukhter Hossain, Shariful Islam, Abujarr Siddiki “Learning a Deep Neural Network for Predicting Phishing Website”, 2019
42. Hemali Sampat, Manisha Saharkar, Ajay Pandey , Hezal Lopes “ Detection of Phishing Website Using Machine Learning”, 2018