INTERNATIONAL JOURNAL OF RESEARCH AND DEVELOPMENT IN ENGINEERING SCIENCES
e-ISSN: 2582 - 4201 Volume: 2 | Issue: 3 | Year: 2020

Phishing URL Detection using Machine Learning

YEGIREDDI RAMESH [1]

[1] Associate Professor,
Department of Computer Science and Engineering,
Aditya Institute of Technology and Management (A), Tekkali
Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India

L Praveen Kumar [2], K Harini [3], M Tejeswara Rao [4], M Ravi Kiran [5]

[2,3,4,5] IV B.Tech Students,
Department of Computer Science and Engineering,
Aditya Institute of Technology and Management (A), Tekkali
Jawaharlal Nehru Technological University Kakinada

Abstract — *Phishing sites which expects to take the victims confidential data by diverting them to surf a fake website page that resembles an honest to goodness one is another type of criminal acts through the internet and its one of the especially concerns toward numerous areas including e-managing an account and retailing. Phishing site detection is truly an unpredictable and element issue including numerous components and criteria that are not stable.*

On account of the last and in addition ambiguities in arranging sites because of the intelligent procedures programmers are utilizing, some keen proactive strategies can be helpful and powerful tools can be utilized, for example, fuzzy, neural system and data mining methods can be a successful mechanism in distinguishing phishing sites. We applied Random Forest (RF), one of the different types of machine learning based algorithms used for detection of Phishing websites, we extract 25 features of an URL and predict the whether website is Phishing or not. Finally, we measured and compared the performance of the classifier in terms of accuracy.

Keywords: *Phishing, legitimate, URL, feature extraction, Machine learning, applications, classification, approach, Algorithm.*

1. Introduction

Now days, As there are so many people are being aware of using internet to perform various activities like online shopping, online bill payment ,online mobile recharge, banking transaction .Due to wide use of this customer face various security threats like cybercrime .There are many cybercrime that are widely performed for example spam , fraud ,cyber terrorisms and phishing. Among this phishing is new cybercrime and very popular nowadays.

Phishing is fraud attempt, which performed to obtain sensitive information of user. Phisher design website which looks same as any legitimate site and spoof user for obtaining private information of user such as username, password, banking details for miscellaneous reasons. Phishing is the most unsafe criminal exercises in cyber space and it is a criminal mechanism employing both social engineering and technical tricks to steal consumers' personal identity data and financial account credentials. Social engineering schemes use spoofed e-mails, purporting to be from legitimate businesses and agencies, designed to lead consumers to counterfeit websites that trick recipients into divulging financial data such as usernames and passwords.

Technical subterfuge schemes install malicious software onto computers, to steal credentials directly, often using systems to intercept consumers' online account user names and passwords [1].

The Procedure of Phishing Attack

Phishing attack imitates the characteristics and features of emails and makes it look the same as the original one. It appears similar to that of the legitimate source. The user thinks that this email has come from a genuine company or an organization. This makes the user to forcefully visit the phishing website through the links given in the phishing email.

These phishing websites are made to mock the appearance of an original organization website. The phishers force user to fill up the personal information by giving alarming messages or validate account messages etc. so that they fill up the required information which can be used by them to misuse it.

They make the situation such that the user is not left with any other option but to visit their spoofed website. [2]

Figure. 1 represents the webpage of the popular website www.facebook.com. Figure. 2 represents a webpage similar to that of Facebook, but is the webpage of a site which spreads phishing activities.

A user may misunderstand the second site as genuine Facebook site and provide his personal identity details. The Phisher can thus steal that information and he may use it for vicious purposes.

Figure. 1: Original Face book webpage

Figure.2: Phishing webpage

2. Literature Survey

Chunlin et al. [3] proposed approach that primarily focus on character frequency features. In this they have combined statistical analysis of URL with machine learning technique to get result that is more accurate for classification of malicious URLs. Also they have compared six machine-learning algorithms to verify the effectiveness of proposed algorithm which gives 99.7% precision with false positive rate less than 0.4%.

Bhagyashree et al.[4] proposed a feature-based approach to classify URLs as phishing and non-phishing. Various features this approach uses are lexical features, WHOIS features, Page Rank and Alexa rank and Phish Tank-based features for disguising phishing and non-phishing website. In this paper, web-mining classification is used.

Pradeepthi et al.[5] In this paper, Author studied different classification algorithm and concluded that tree-based classifiers are best and gives better accuracy for phishing URL detection. Also Author uses various features such as lexical features, URL based feature, network based features and domain based feature.

Authors in this paper [6] also proposed reduced feature selection model to detect phishing websites. They used Logistic Regression and Support Vector Machine (SVM) [8] as classification methods to validate the feature selection method. 19 features reduced from 30 site features have been selected and used for phishing detection. The LR and SVM calculations performance was surveyed dependent on precision, recall, f-measure and accuracy. Study shows that SVM algorithm achieved best performance over LR algorithm.

This paper [7] proposes a efficient way to detect phishing URL websites by using c4.5 decision tree approach. This technique extracts features from the sites and calculates heuristic values. These values were given to the c4.5 decision tree algorithm [7] to determine whether the site is phishing or not. Dataset is collected from PhishTank and Google. This process includes two phases namely pre-processing phase and detection phase [7]. In which features are extracted based on rules in pre-processing phase and the features and their respected values were inputted to the c4.5 algorithm and obtained 89.40% accuracy.

The work by Garera et al. [8] uses logistic regression over hand-selected features to classify phishing URLs. The features include the presence of red flag keywords in the URL, features based on Google's Page Rank, and Google's Web page quality guidelines. It is difficult to make a direct comparison with our approach without access to the same URLs and features.

Using [9] the parameters of the HTML, JavaScript, URL, Host Based Features can also help in determining the URLs. We will see the Novel method [10] of using

the DNS at the higher-level hierarchy where the model will access the domains in the DNS.

Authors [11] in this paper created an extension to Google Chrome to detect phishing websites content with the help of machine learning algorithms. Dataset UCI-Machine Learning Repository used and 22 features were extracted for this dataset. Algorithms kNN, SVM and Random Forest were chosen for precision, recall, f1-score and accuracy comparison. Random Forest obtained a best score and HTML, JavaScript, CSS [11] used for implementing chrome extension along with python.

Mustafa et al. [12] developed safer framework for detecting phishing website. They have extracted URL features of website and using subset-based selection technique to obtain better accuracy. In this paper, author evaluated CFS subset based and content-based subset selection methods And Machine learning algorithms are used for classification purpose.

Ahmad et al. [13] proposed three new features to improve accuracy rate for phishing website detection. In this paper, Author used both type of features as commonly known and new features for classification of phishing and non-phishing site. At the end author has concluded this work can be enhanced by using these novel features with decision tree machine learning classifiers.

Apart from the work of Choi et al. in [14] where authors evaluate the performance of a classifier for discriminating three types of malicious URLs (Spam, Phishing, Malware), most of the related works focus on either malicious URLs in general or any specific type of URLs (e.g. Phishing). Regarding feature selections, most of the works depend on various kind of features: lexical, content, obfuscation, DNS etc.

Luong et al. [15] proposed new technique to detect phishing website. In proposed method, Author used six heuristics that are primary domain, sub domain, path domain, page rank, and Alexa rank, Alexa reputation whose weight and values are evaluated.

This approach gives 97 % accuracy but still improvement can be done by enhancing more heuristics.

3. METHODOLOGY

In this section we will learn about the various machine learning classifiers to predict phishing site. We will also explain our proposed methodology to detect phishing URL. In section A we shall explain various classifiers models and methods which can be used to check the URL weather phishing or legitimate URL. In section B we shall explain our Proposed Method.

A. Various machine learning classifiers and their methodlogies to detect phishing URL:

Detecting and identifying Phishing Websites is really a complex and dynamic problem. The phishing attacks can be carried out in many ways such as email, website, malware, SMS and voice. In this work, we concentrate on detecting website phishing (URL), which is achieved by making use of the Hybrid Algorithm Approach.one of the Hybrid Algorithm like:

Stacking is mixture of different classifiers working together to gives a good prediction rate and results. Each of classifiers have its own way of working and classification. Let us discuss each of them in details [16].

- **K-Nearest Neighbours (KNN):** KNN is one of the simple machine learning algorithm classification. KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function).

  Classification is done by a majority vote to its neighbors. The data is assigned to the class which has the nearest neighbors. As you increase the number of nearest neighbors, the value of k, accuracy might increase [17].

- **Support vector machine (SVM)**: This is also one of the classification algorithms which is supervised and is easy to use. It can used for both classification and regression applications, but it is more famous to be used in classification applications.

  In this algorithm each point which is a data item is plotted in a dimensional space, this space is also known as n dimensional plane, where the 'n' represents the number of features of the data. The classification is done based on the differentiation

in the classes, these classes are data set points present in different planes.

- **Logistic Regression**: Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable.

  Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used.

  Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

- **Decision Tree:** A decision tree is a flowchart-like tree structure where an internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value.

  It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.[19]

B. Proposed Methodlogy

The dataset of phishing and legitimate URL's contains the features have around 30 characteristics of phishing websites which is used to differentiate it from legitimate ones. Each category has its own characteristics of phishing attributes and values are defined. The phishing attributes values are represented with binary numbers 0 and 1 which indicates the attribute is present or not.

After this the data is trained we shall apply a relevant machine learning algorithm to the dataset. The machine learning algorithms are already explained in previous section. After this we use a hybrid classification Algorithm like Stacking in which we combine four of the classifiers namely K-Nearest Neighbours, Logistic Regression, SVM and Decision Tree classifier. To predict the accuracy of the detection of the phishing URL, hence we get our desired result.

This is also called a hybrid approach to test the data, in this method we propose to use the combination of four classifiers, as mentioned above. We shall then test the data and evaluate the prediction accuracy which shall be more than the existing system. We shall now see the different classifiers and discuss the hybrid combination used for our proposed system.

In the training phase, we should use the labelled data in which there are samples such as phish area and legitimate area. If we do this then classification will not be a problem for detecting the phishing domain. To do a working detection model it is very crucial to use data set in the training phase. We should use samples whose classes are known to us, which means the samples whom we label as phishing should be detected only as phishing URL (-1). Similarly the samples which are labelled as legitimate will be detected as legitimate URL (1).

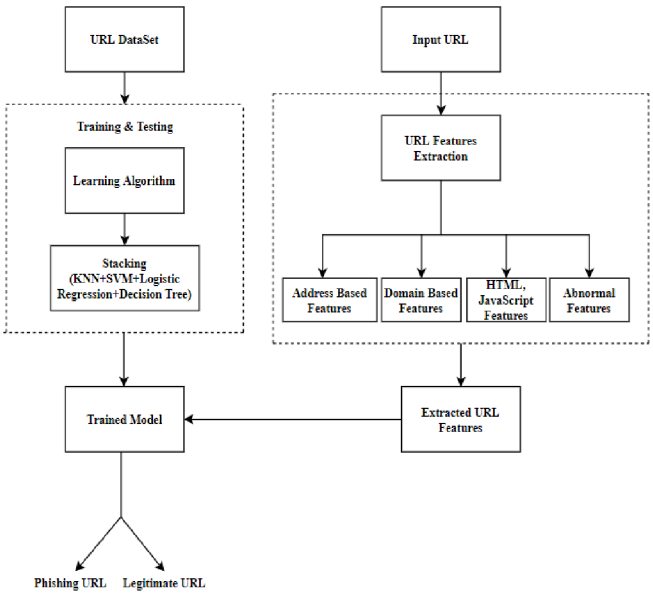

Figure. 3: Proposed System Block Diagram

The dataset to be used for machine learning must actually consist these features. There so many machine learning algorithms and each algorithm has its own working mechanism which we have already seen in the previous chapter. The existing system uses any one of the suitable machine learning algorithms for the detection of phishing URL and predicts its accuracy.

Each of the algorithms which explain in the earlier section has some disadvantages hence it is not recommended to use one machine learning algorithm to detect the phishing website. [20]

- **Dataset:** In our work, we shed light on the important features that have proved to be sound and effective in predicting phishing websites. The data of URLs is obtained from Phishtank website, where Phishtank is an anti-phishing site. It contains 440022 URLs which is in structured form. Our main objective is to detect whether the URL is phishing or legitimate based on the features extracted.

  Each row represents the 22 features of URL in format 1, -1 and 0. Where 1 indicates Legitimate, 1 represents Phishing and 0 represents Suspicious URL. Sample Dataset is as follow:

having_IP_Ad	URL_Length	Shortining	having_At_	double_sla	Prefix_Suf	having_Sul	SSLfinal_St	Domain_re
-1	1	1	1	-1	-1	-1	-1	-1
1	1	1	1	1	-1	0	1	-1
1	0	1	1	1	-1	-1	-1	-1
1	0	1	1	1	-1	-1	-1	1
1	0	-1	1	1	-1	1	1	-1
-1	0	-1	1	-1	-1	1	1	-1
1	0	-1	1	1	-1	-1	-1	1
1	0	1	1	1	-1	-1	-1	1
1	0	-1	1	1	-1	1	1	-1
1	1	-1	1	1	-1	-1	1	-1
1	1	1	1	1	-1	0	1	1
1	1	-1	1	1	-1	1	-1	-1
-1	1	-1	1	-1	-1	0	0	1

Figure. 4: Data Set

Feature name	Description
Using the IP Address	If Domain Part of URL has an IP Address → Phishing Otherwise → Legitimate
Domain Registration Length	If Domains Expires on < 1 years → Phishing Otherwise → Legitimate
Using Non-Standard Port	If Port No is of the Preferred Status → Phishing Otherwise → Legitimate
Server Form Handler (SFH)	If SFH is "about: blank" Or Is Empty → Phishing SFH Refers to A Different Domain → Suspicious Otherwise → Legitimate
Disabling Right Click	If Right Click Disabled → Suspicious Otherwise → Legitimate
Age of Domain	If Age of Domain ≥ 6 months → Legitimate Otherwise → Phishing
Google Index	If Webpage Indexed by Google → Legitimate Otherwise → Phishing
Website Traffic	If Website Rank < 100,000 → Legitimate Website Rank > 100,000 → *Suspicious* Otherwise → Phish
DNS Record	If no DNS Record for The Domain → Phishing Otherwise → Legitimate
Statistical-Reports Based Feature	If Host Belongs to Top Phishing IPs or Top Phishing Domains → Phishing Otherwise → Legitimate

Table1.1: URL Features

- **URL Features:** Referring Table 1. above, features are associated with Phishing/Legitimate/suspicious URL patterns and characters. Feature 1: If an IP address is used as an alternative of the domain name in the URL. Currently, to prevent a user from identifying that a site is not legitimate, phishing sites typically hide the primary domain; the URLs of these phishing sites have unusually long subdomains.

  Feature 2 and 6 are another new feature that reflects current phishing trends. This feature includes referred as domain-based features. Google Index feature examines whether a website is in Google's index or not. When a site is indexed by Google, it is displayed on search results. many phishing webpages may not be found on the Google index.

  The number of links pointing to the webpage indicates its legitimacy level, even if some links are of the same domain, legitimate websites have at least 2 external links pointing to them. Several parties such as PhishTank (PhishTank Stats, 2010-2012), etc. formulate numerous statistical reports on phishing websites at every given period of time.

  In our research, we used 2 forms of the top ten statistics from PhishTank: "Top 10 Domains" and "Top 10 IPs" according to statistical-reports published in the last three years.

4. IMPLEMENTATION

This section provides knowledge about the implementation environment and throws light on the actual steps for the implementation of dataset to get better accuracy to predict phishing by using different classifier and their accuracies

Implementation Steps

In this section we shall discuss about the actual steps which were implemented while doing the experiment. We shall explain the stepwise procedure used to analyse the data and to predict the phishing.

The program flow for the classifier performance is shown in Figure 5.

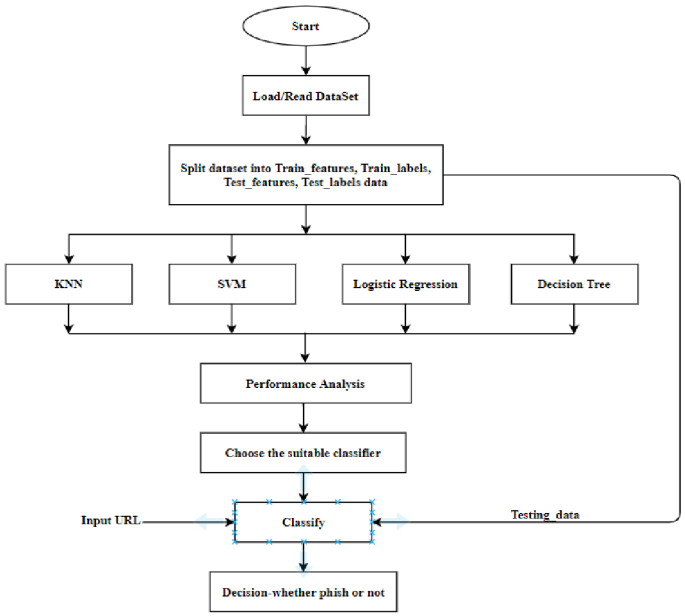

Figure. 5. Program Flow

The system consists of the following main steps:

1. We have used structured data which consists 30 features of URL which obtained from Phishtank website which consists of both phishing and legitimate URL where most of URLs obtained are phishing.

2. These features are Using the IP Address, Long URL, URL Shortening Services, URL's having "@" Symbol, Redirecting using "//", Adding Prefix or Suffix Separated by (-) to the Domain, Sub Domain and Multi Sub Domains, Domain Registration Length, Favicon, Using Non-Standard Port, HTTPS Token in the Domain Part of the URL, Request URL, URL of Anchor, Links in <Meta>, <Script> and <Link> tags, Server Form Handler (SFH), Submitting Information to Email, Abnormal URL, Website Forwarding, Status Bar Customization, Disabling Right Click, IFrame Redirection, Age of Domain, DNS Record, Website Traffic, Google Index, and Statistical-Reports Based Feature.

3. Next, we train the four different classifiers and compare their performance on the basis of accuracy four classifiers used are XG Boost, SVM, Logistic regression, and RandomForest.

4. In our work, data set is splitted into 80-20% i.e. 80% of data used to train the model and 20% of data is used for test the model.

5. Then classifier detects the given URL based on the training data that is if the site is phishing it shows a This is Phishing URL or if legitimate it shows this is Legitimate URL on User Interface (UI).

6. We compare the accuracy of different classifiers and found XG Boost and RandomForest are the best classifiers which gives the maximum accuracy.

7. Below are the screen shots for the implementation process.

Project Input User Interface (UI) Screens

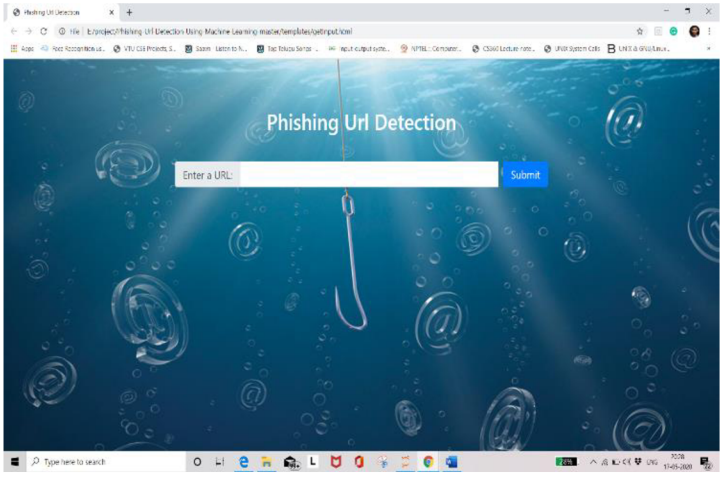

Figure. 6: UI for User Input URL

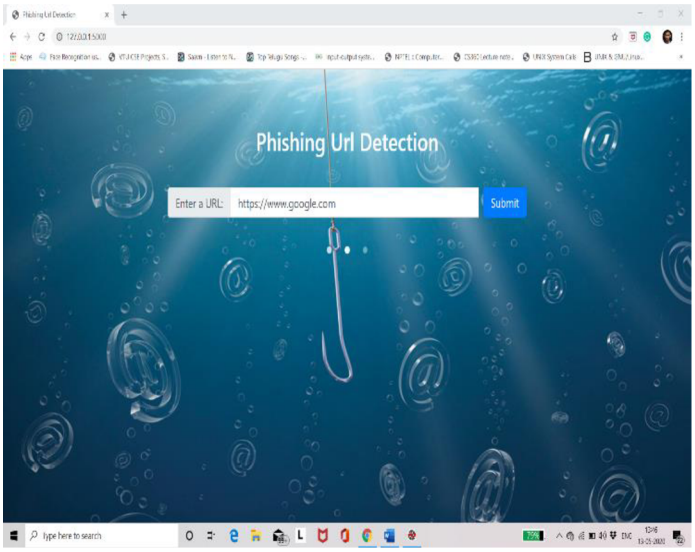

Figure. 7: Testing for Legitimate

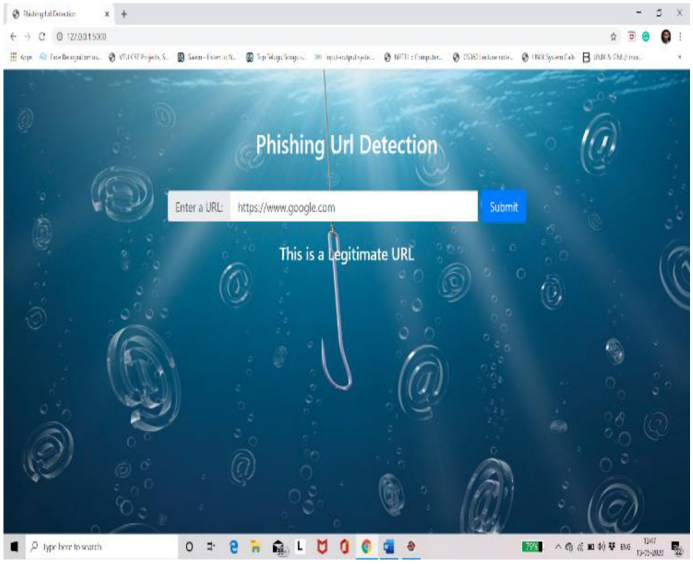

Figure. 8. Showing Result

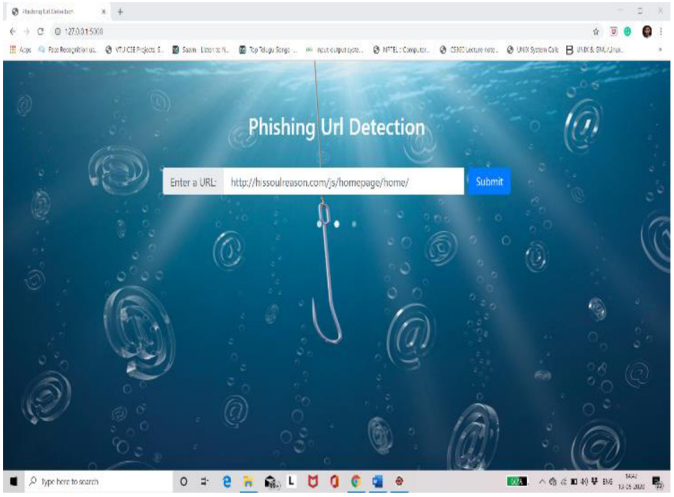

Figure. 9: Testing for Phishing

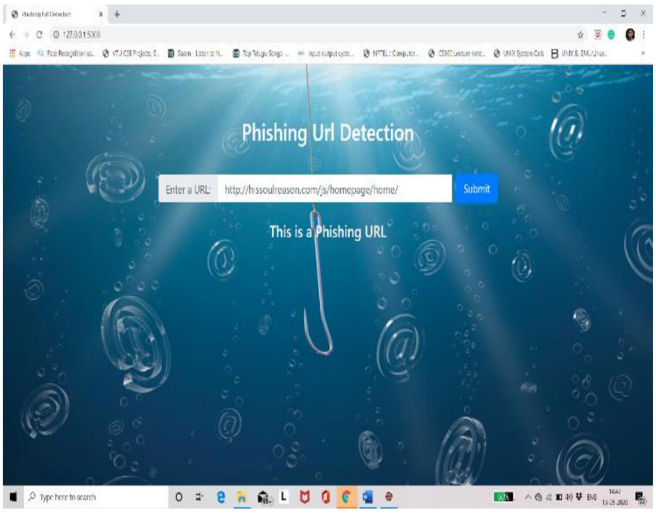

Figure. 10: Showing Result

We do this testing by using 4 different techniques of classification. We shall show the screenshot of our Stacking classifier of confusion Matrix and ROC curve for the other 4 classifier models

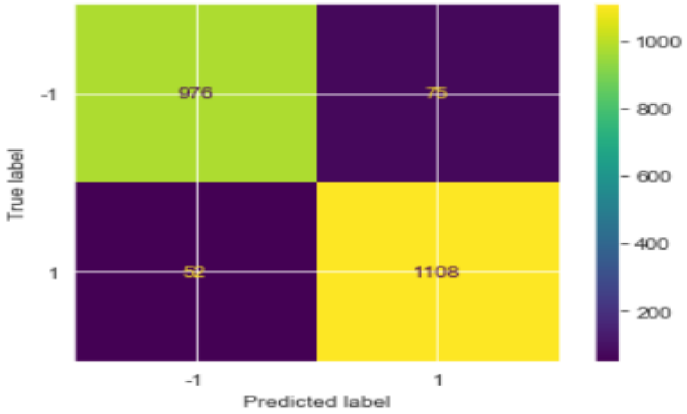

Figure 11: Confusion Matrix of Stacking Classifier

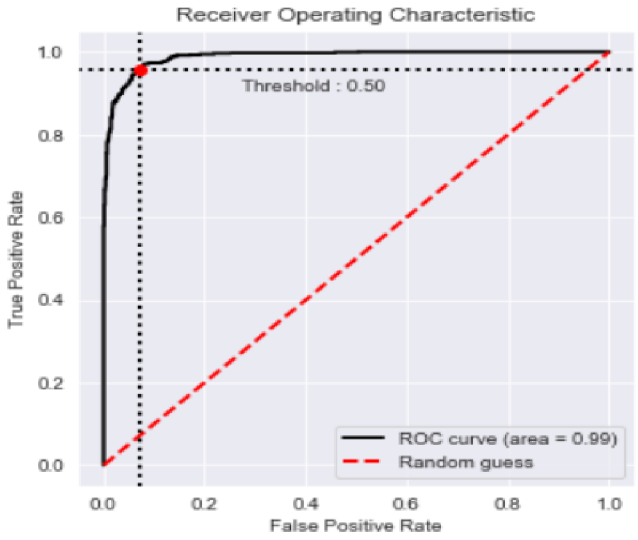

Figure. 12: ROC curve of stacking classifier

Next, we follow the same procedure to plot the ROC curves for remaining models KNN, SVM, Logistic Regression, and Decision Tree classifiers as same as for Stacking classifier model.

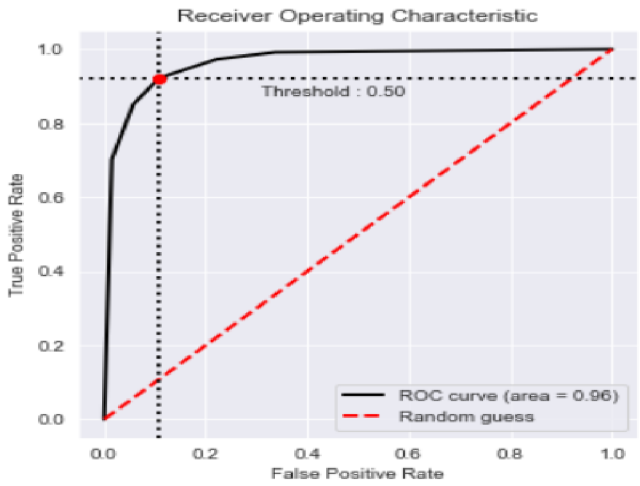

Figure.13: ROC curve of KNN model

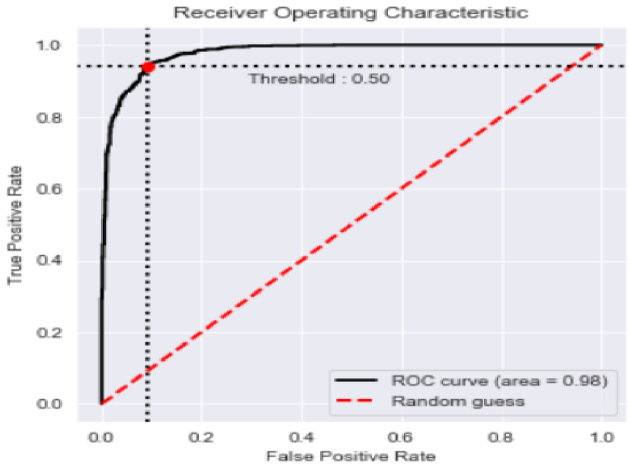

Figure. 14: ROC curve of SVM model

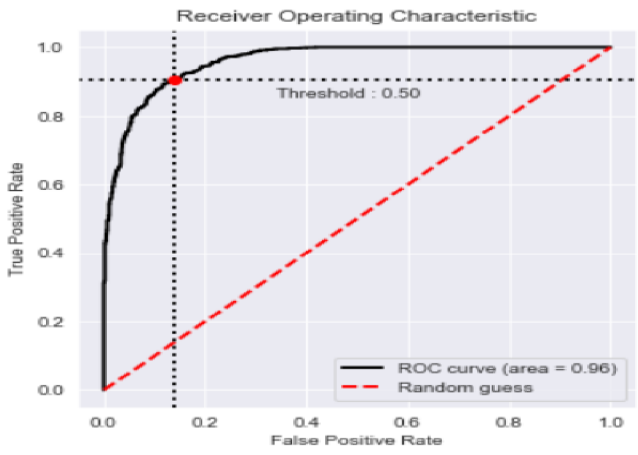

Figure.15: ROC curve of Logistic Regression model

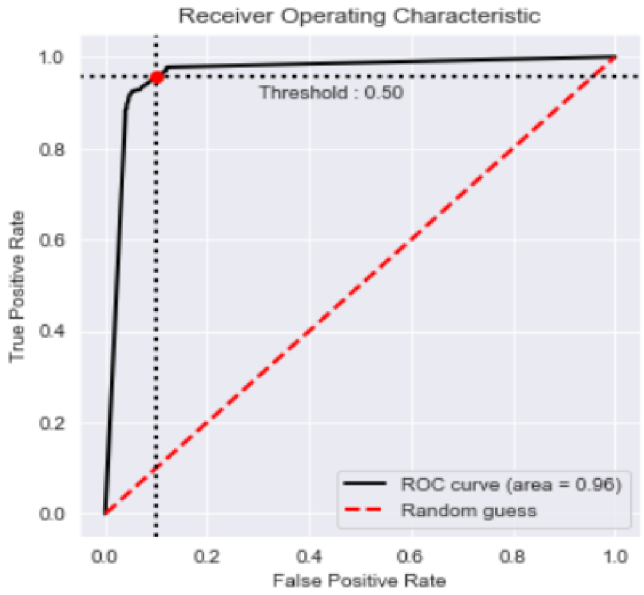

Figure. 16. ROC curve of Decision Tree model

5. OBSERVATIONS AND RESULT

a. Observation:

As we discussed in the previous section, we have used four Machine Learning classifiers to predict and detect whether given URL have Phishing Features or not. To increase the prediction accuracy, we combine all four model using Stacking technique. Comparisons of these classifiers have been shown below in the accuracy table.

Classifiers / Metrics	KNN	SVM	Logistic Regression	Decision Tree	Stacking
Confusion matrix	937 114 90 1070	943 108 57 1103	906 145 112 1048	959 92 64 1096	967 84 56 1104
Precision	0.9037	0.9108	0.8784	0.92255	0.92929
Recall	0.9224	0.9508	0.9034	0.94482	0.95172
F1 score	0.9129	0.9304	0.8907	0.93356	0.94037
ROC AUC	0.9069	0.9240	0.8827	0.92864	0.93590
Accuracy	90.773	92.237	88.376	92.144	93.660

Table 2: Comparison on various model and its metrics

b. Result:

We have got desired results of detecting the given URL is Phishing or Legitimate URL by using the above four classifiers together. Fig 17 shows the individual accuracies of the model and last bar shows the combined result of stacking classifier. Testing is done on data by splitting it into 40% and 20% test split.

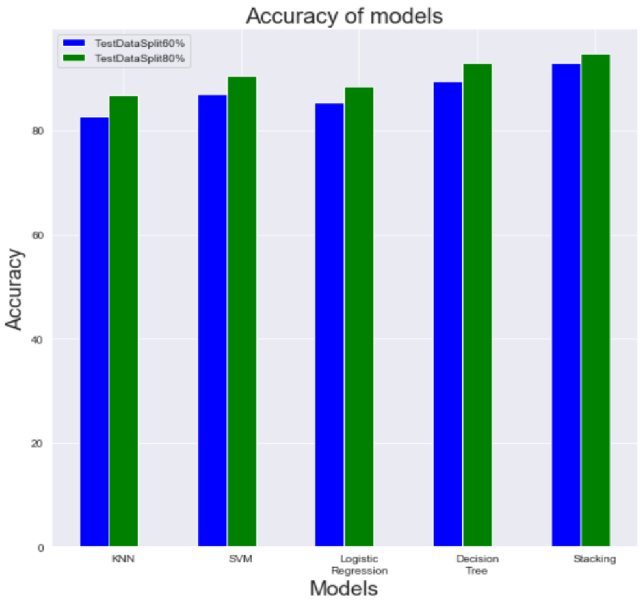

Figure.17: Accuracy Results

6. CONCLUSION

This project is to predict whether a given URL is phishing website or not. It turns out in the given experiment that by combining above stated classifiers (Stacking) and its performances as a best classifier with great classification accuracy of 93.66% for the given dataset of phishing site. As a future work we might use this model to other Phishing dataset with larger size then now and then testing the performance of those classification algorithm's in terms of classification accuracy for more better accuracies.

REFERENCES

[1] Phishing Trends Report for Q3 2012, Anti Phishing Working Group. http://antiphishing.org.

[2] Pujara, P., & Chaudhari, M. B. (2018). Phishing Website Detection using Machine Learning: A Review.

[3] Chunlin Liu, Bo Lang : Finding effective classifier for malicious URL detection : In ACM,2018

[4] Bhagyashree E. Sananse, Tanuja K. Sarode : Phishing URL Detection: A Machine Learning and Web Mining-based Approach : In International Journal of Computer Applications,2015

[5] Pradeepthi. K V and Kannan. A: Performance Study of Classification Techniques for Phishing URL Detection: In 2014 Sixth International Conference on Advanced Computing(ICoAC) IEEE,2014

[6] W. Fadheel, M. Abusharkh, and I. Abdel-Qader, "On Feature Selection for the Prediction of Phishing Websites," 2017 IEEE 15th Intl Conf Dependable, Auton. Secur. Comput. 15th Intl Conf Pervasive Intell. Comput. 3rd Intl Conf Big Data Intell. Comput. Cyber Sci. Technol. Congr., pp. 871–876, 2017.

[7] L. MacHado and J. Gadge, "Phishing Sites Detection Based on C4.5 Decision Tree Algorithm," in 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017, 2018, pp. 1–5.

[8] Garera S., Provos N., Chew M., Rubin A. D., "A Framework for Detection and measurement of phishing attacks", In Proceedings of the ACM Wo rkshop on Rapid Malcode (WORM), Alexandria, VA.

[9] Davide Canali, Marco Cova, Giovanni Vigna, Christopher Kruegel, "Prophiler: A Fast Filter for the Large-Scale Detection" International World Wide Web Conference Committee (IW3C2), www, 2011 Hyderabad, India ACM 978-1-4503-0632-4/11/03.

[10] Manos Antonakakis, Roberto Perdisci, Wenke Lee, Nikolaos vasiloglou II, and David Dagon," Detecting Malware Domains at the Upper DNS Hierarchy" Conference: Proceedings of the 20th USENIX conference on Security in August 2011 https://www.usenix.org/legacy/event/sec11/tech/full_papers/A ntonakakis.pdf

[11] A. Desai, J. Jatakia, R. Naik, and N. Raul, "Malicious web content detection using machine leaning," RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc., vol. 2018–Janua, pp. 1432–1436, 2018.

[12] Mustafa AYDIN, Nazife BAYKAL : Feature Extraction and Classification Phishing Websites Based on URL : IEEE,2015

[13] Ahmad Abunadi, Anazida Zainal ,Oluwatobi Akanb: Feature Extraction Process: A Phishing Detection Approach :In IEEE,2013.

[14] Choi, H., Zhu, B.B., Lee, H.: Detecting malicious web links and identifying their attack types. In: Proceedings WebApps (2011)

[15] Luong Anh Tuan Nguyen † , Ba Lam To † ,Huu Khuong Nguyen † and Minh Hoang Nguyen : Detecting Phishing Websites: A Heuristic URL-Basd Approach: In The 2013 International Conference on Advanced Technologies for Communications (ATC'13)

[16] Shekokar, N. M., Shah, C., Mahajan, M., & Rachh, S. (2015). An ideal approach for detection and prevention of phishing attacks. Procedia Computer Science, 49, 82-91.

[17] https://blog.quantinsti.com/machine-learning-k-nearest-neighbors-knn-algorithm-python/

[18] Mohammad, R. M., Thabtah, F., & McCluskey, L. (2012, December). An assessment of features related to phishing websites using an automated technique. In 2012 International Conference for Internet Technology and Secured Transactions (pp. 492-497). IEEE.

[19] https://www.datacamp.com/community/tutorials/decision-tree-classification-python

[20] Jain, A. K., & Gupta, B. B. (2018). PHISH-SAFE: URL features-based phishing detection system using machine learning. In Cyber Security (pp. 467-474). Springer, Singapore