IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Phish Catch: Machine Learning Way of Detecting Phishing Websites

Sahil Hussain,

Department of computer science and it,
University of Jammu,Jammu,India

Abstract: With the advent of 4G technology, the internet became available to masses. Everyone started to use internet services in different spheres of their life, making them vulnerable to diverse threats. One of the primary risks for internet users is Phishing Websites. Instead of breaching the security of systems phishing websites try to fool the users and make them give away the credentials which they are not supposed to share with anyone. In this study, we took 21 features and tried to predict their class i.e
legitimate or phish using a supervised learning algorithm

Index Terms - **Phishing, Machine Learning, SVM, Decision Tree, Random Forest, Internet, Security**

1. INTRODUCTION

Phishing is the type of attack where the target is a user, not system. The term phishing was introduced in 1987 [1] . Phishing is a crime employing both social engineering and technical subterfuge to steal consumers' identity data and financial account credentials. Social engineering schemes prey on unwary victims by fooling them into believing they are dealing with a trusted, legitimate party, such as by using deceptive email addresses and email messages. These are designed to lead consumers to counterfeit Web sites that trick recipients into divulging financial data such as usernames and passwords. Technical subterfuge schemes plant malware onto computers to steal credentials directly, often using systems that intercept consumers' account user names and passwords or misdirect consumers to counterfeit Web sites [2].
The phishing attacks are detected based on various approaches such as BlackList /whiteList approach, machine learning-based techniques, Fuzzy logic-based techniques , image-based techniques and Heuristic techniques.
Types of phishing attacks:

1.1.1 Spear phishing

Phishing attempts directed at specific individuals or companies is known as *spear phishing*. In contrast to bulk phishing, spear phishing attackers often gather and use personal information about their target to increase their probability of success.

1.1.2 Whaling

The term **whaling** refers to spear phishing attacks directed specifically at senior executives and other high-profile targets. In these cases, the content will be crafted to target an upper manager and the person's role in the company. The content of a whaling attack email may be an executive issue such as a subpoena or customer complaint.

1.1.3 Clone phishing

Clone phishing is a type of phishing attack whereby a legitimate, and previously delivered, email containing an attachment or link has had its content and recipient address(es) taken and used to create an almost identical or cloned email. The attachment or link within the email is replaced with a malicious version and then sent from an email address spoofed to appear to come from the original sender. It may claim to be a resend of the original or an updated version to the original. Typically this requires either the sender or recipient to have been previously hacked for the malicious third party to obtain the legitimate email.

1.1.4 Link manipulation

Most methods of phishing use some form of technical deception designed to make a link in an email (and the spoofed website it leads to) appear to belong to the spoofed organization. Misspelled URLs or the use of subdomains are common tricks used by phishers.

1.1.5 Filter evasion

Phishers have sometimes used images instead of text to make it harder for anti-phishing filters to detect the text commonly used in phishing emails. In response, more sophisticated anti-phishing filters are able to recover hidden text in images using OCR (optical character recognition).

1.6 Website forgery

Some phishing scams use JavaScript commands in order to alter the address bar of the website they lead to. This is done either by placing a picture of a legitimate URL over the address bar, or by closing the original bar and opening up a new one with the legitimate URL.

1.1.7 Covert redirect

Covert redirect is a subtle method to perform phishing attacks that makes links appear legitimate, but actually redirect a victim to an attacker's website. The flaw is usually masqueraded under a log-in popup based on an affected site's domain.

1.1.8 Social engineering

Users can be encouraged to click on various kinds of unexpected content for a variety of technical and social reasons. For example, a malicious attachment might masquerade as a benign linked Google Doc.

Alternatively users might be outraged by a fake news story, click a link and become infected.

1.1.9Voice phishing

Not all phishing attacks require a fake website. Messages that claimed to be from a bank told users to dial a phone number regarding problems with their bank accounts. Once the phone number (owned by the phisher, and provided by a voice over IP service) was dialed, prompts told users to enter their account numbers and PIN. Vishing (voice phishing) sometimes uses fake caller-ID data to give the appearance that calls come from a trusted organization.[3]

2. LITERATURE SURVEY

Many researchers have worked for the detection of phishing websites. Some of the notable work is mention below

Amani Alswailem et al [4] suggested 36 features of a website based on URL, page content and rank. Random forest classifier is applied to these features. From the experiment, authors have reported the best 26 features which produce maximum accuracy of 98.8% for detection of phishing websites.

Arun Kulkarni et al [5] selected 10 features of the website which are used to detect phishing website. Authors have used decision tree classifier, support vector machine (SVM), naïve Baye's classifier and neural networks for an experiment in MatLab. After experiment authors reported best results of 91.5% using a decision tree, followed by SVM producing 86.69%, naïve Baye's 86.14% and neural networks producing an accuracy of 84.87%.

Vaibhav Patil et al [6] devised a hybrid solution using all three basic approaches i.e. blacklist-whitelist, heuristic approach and visual similarity. Authors employed logistic regression, decision tree & random forest classifier on python using scikit-learn package. After experiment best accuracy of 96.58% was found using random forest classifier.

Neda Abdelhamid et al [7] identified 30 features of the website which are used to determine the authenticity of the website. They implemented eight machine learning algorithms namely Bayes Net, C4.5, SVM, AdaBoost, eDRI, OneRule, conjunctive rule & RIDOR on weka tool and concluded decision tree and eDRI have the best performance and outperformed others.

Hossein Shirazi et al[8] used 30 features of websites which are based on URL, DNS, HTML, JavaScript, External statistics. They performed an experiment using python using scikit-learn for SVM, also built a deep neural network using TensorFlow and tf contrib. the result of the study showed tensor flow implementations took larger time to train while being marginally accurate than SVM. The highest accuracy of 90% is reported using Gaussian-SVM.

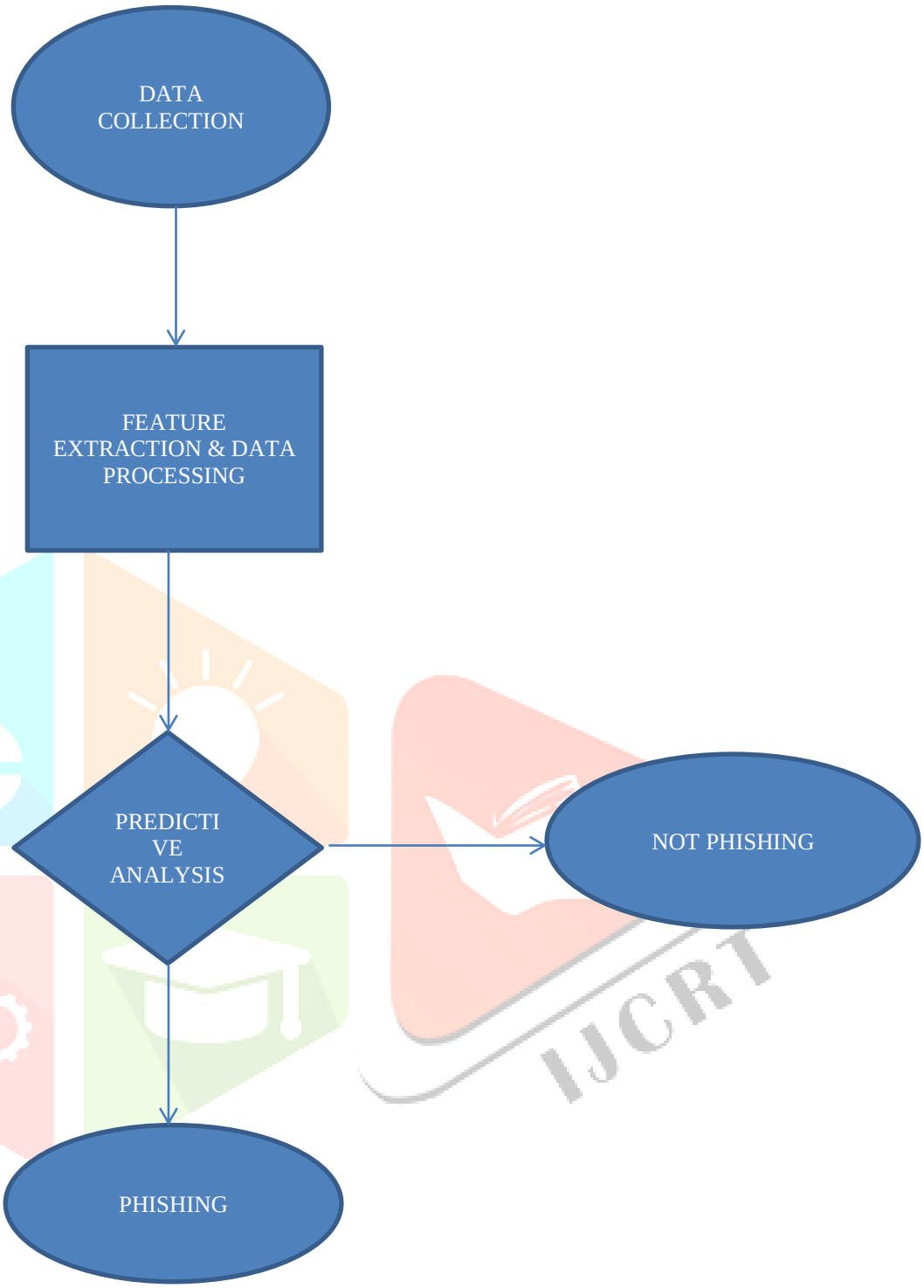
Wiena Niu et al [9] constructed hybrid classifier called Cuckoo Search Simple Vector Machine (CS-SVM) in which they used 23 features of the website which include body-based, URL based and Header based features to detect phishing email. In the result of the experiment accuracy of 92.8% was reported using CS-SVM and accuracy of 89% using simple SVM.

Sadeh et al. [10] proposed a system called PILFER for classifying phishing URLs. They extracted a set of ten features that are specifically designed to highlight deceptive methods used to fool users. The data set consists of approximately 860 phishing e-mails and 6950 non-phishing emails. They used a Support Vector Machine (SVM) as a classifier in the implementation. They trained and tested the classifier using 10-fold cross validation and obtained 92 percent accuracy.

Miyamoto et al. [11] provide an overview of several different machine learning techniques, including SVM, Random Forests, Neural Networks, Naive Bayes and Bayesian Additive Regression Trees. They analyze how accurate each one is on a dataset developed by Z. Hong et al., called CANTINA[12]. Miyomoto et al. achieved a maximum accuracy of 91.34%.

In [13], Guang Xiang, Jason Hong, Carolyn P. Rose, Lorrie Cranor proposed CANTINA+, a comprehensive feature-based approach in the literature including eight novel features, which exploits the HTML Document Object Model (DOM), search engines and third party services with machine learning techniques to detect phish. Also two other filters are designed in it to help reduce FP and achieve good runtime speedup. The first is a near-duplicate phish detector that uses hashing to catch highly similar

3 Experiment



Flow chart 3.1 showing the work involved for study

3.1 Dataset:

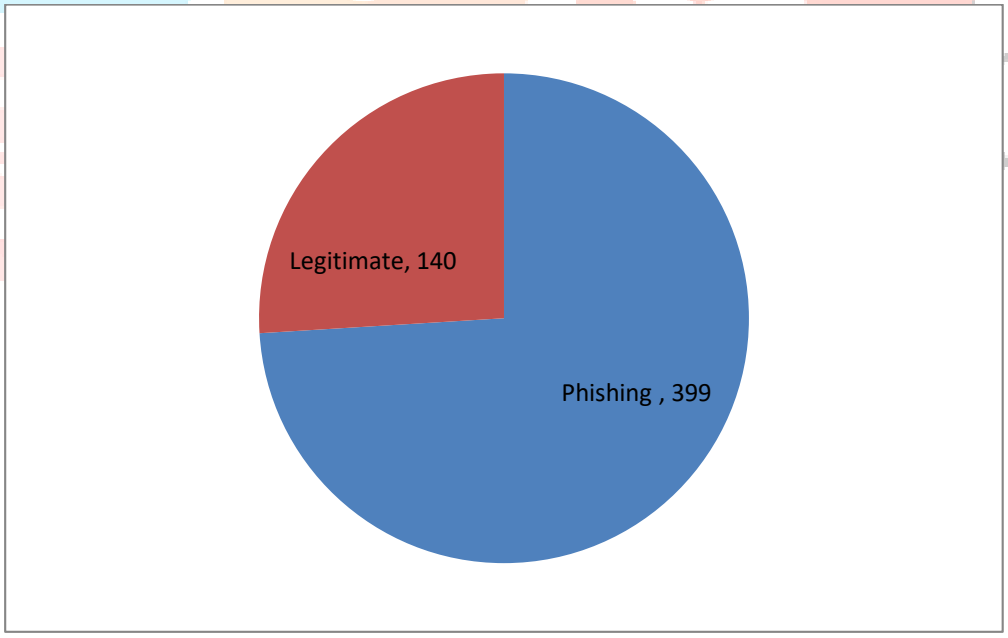
The dataset used for the study has two types of tuples Phishing and legitimate. For phishing tuples, the dataset was downloaded from phishtank.com which contain almost 15000 phish URLs and for a legitimate dataset, URLs are collected in a survey using Google Forms in which 52 people from different professional background participated and copied few URLs from their web browser which they believe are legitimate. From the survey total, 438 legitimate URLs are collected.

After the collection of URLs, the features are extracted from the URLs using specially designed tool coded in python on jupyter notebook.

After feature extraction total of 539 tuples is randomly chosen for the experiment out of which 140 tuples are legitimate and 399 are phishing. For each tuple total, 21 features are extracted. The 21 features are selected on the basis of study done by **Amani Alwailem et al [4]** they have suggested 26 features among which 5 features based on Rank of website are omitted thus we are left with 21 features .

Features	Features
URL Length	Total number of forms
Number of AT(@) in URL	Number of Buttons
Number of backslash (\) in URL	Number of Submit
Number of special characters in URL	Number of GET
Presence of IP in URL	Number of POST
Presence of WWW in URL	Number of Password
Presence of transport security layer	Number of OUTERSCRIPT
Length of Host	Number of iframe
Underscore in Host	Number of Links
Dot(.) in Host	Digits in Host
Hyphen in Host	

Table 3.1 showing Features used



Pie chart 3.1 showing the number of phishing and legitimate tuple in final dataset

3.2 Classifiers:

For the study total 4 classifiers are used which are :

3.2.1 Logistic Regression: Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analysis, it is predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio level independent variables. [14]

3.2.2 Decision Tree: [15] Decision trees are non-parametric classifiers. As its name indicates, a decision tree is a tree structure .where each non-terminal node denotes a test on an attribute, each branch represents an outcome of the test, and the leaf node denotes classes. The basic algorithm for decision tree induction is a greedy algorithm that constructs the decision tree in the top-down recursive divide – and – conquer manner.[16] At each non-terminal node, one of the attributes is chosen for the split. The attribute that gives the maximum information gain is chosen for the split.

3.2.3 Random Forest: The random forest is a supervised learning algorithm that randomly creates and merges multiple decision trees into one "forest." The goal is not to rely on a single learning model, but rather a collection of decision models to improve accuracy. The

primary difference between this approach and the standard decision tree algorithms is that the root nodes feature splitting nodes are generated randomly. [17]

3.2.4 Simple Vector Machines:[18] 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

$Wx+b=0$ (3)

Where W is a weight matrix, and b is a constant. The SVM algorithms find the weight matrix such that it maximizes the distance between the hyper planes separating two classes. Tuples that fall on the hyper planes are called as support vectors [19]

3.3 Tools Used:

3.3.1Softwares used:

1. Python
2. Jupyter notebook
3. Microsoft Excel
4. Firefox internet browser

3.3.2 Hardware configuration:

1. Intel core i5 10th gen 2.4 GHz
2. RAM:8GB

3.3.3 Operating System:

Windows 10

3.4 Result:

The data set is applied on the above said classifiers while the ratio set for training and testing is set to 50-50. The performance is measured on the basis of accuracy ,precision and recall for each the figures are shown below.

Figure 3.2 Showing precision of classifiers used in Study

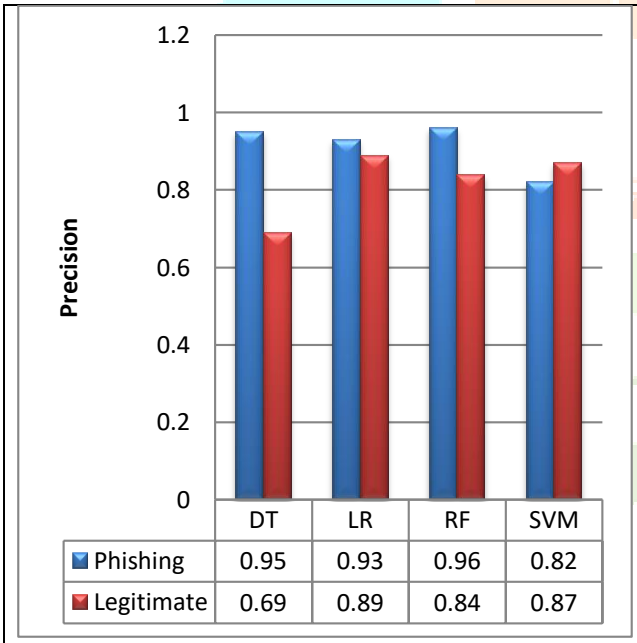
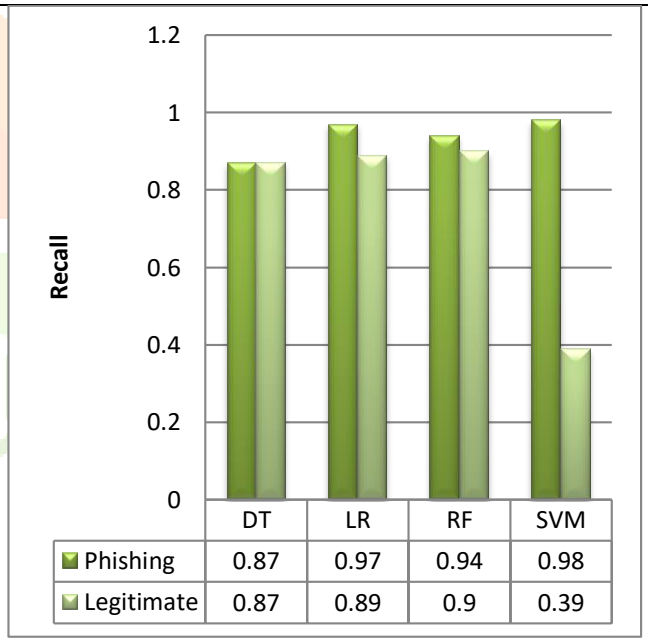


Figure 3.3 showing recall values of Classifiers used



In The table 3.2 below is the summary of Result

	Precision		Recall		Accuracy
	Phishing	**Legitimate**	**Phishing**	**Legitimate**	
Decision Tree	0.95	0.69	0.87	0.87	0.87
Logistic Regression	0.93	0.89	0.97	0.80	0.92
Random Forest Classifier	0.96	0.84	0.94	0.90	0.93
Simple Vector Machine	0.82	0.87	0.98	0.39	0.83

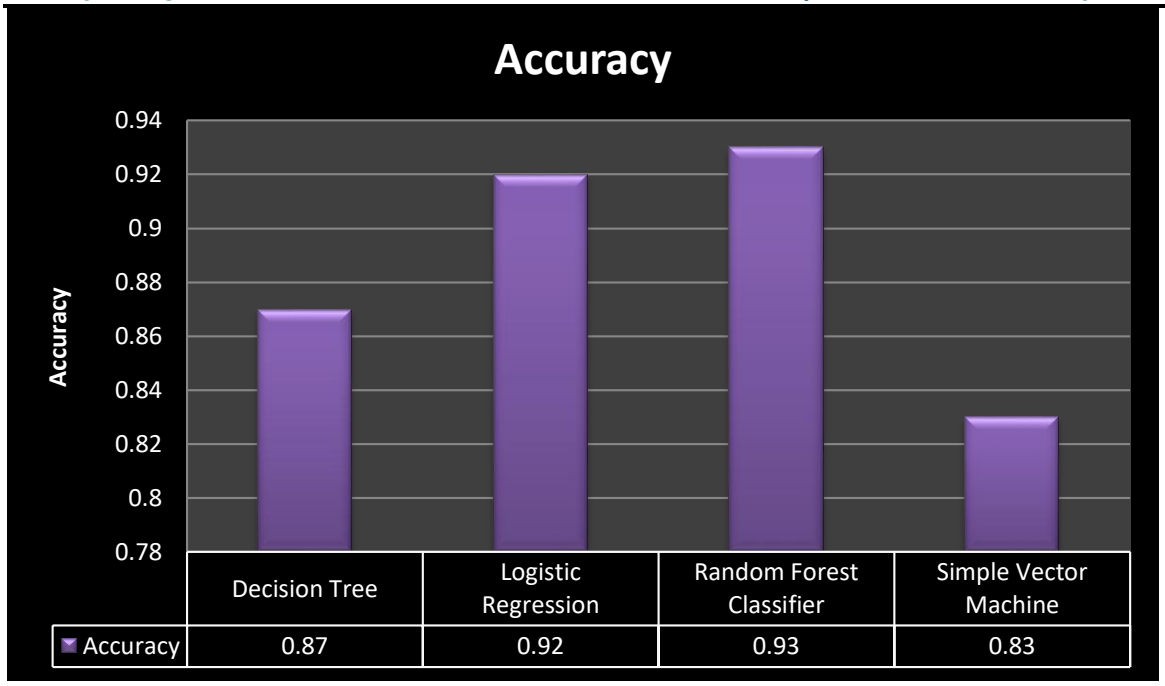


Figure 3.4 showing the accuracy of different classifiers

3.5 Conclusion:

From the study conducted on the Decision tree, Random Forest classifier, Logistic Regression and Simple Vector Machine , The study involved two steps in first step features are extracted from dataset using specially designed Feature Extraction tool and the data cleaning is performed by removing missing data and removing duplicate data, In second step the data set has been applied to classifiers after the experiment it has been observed that Random Forest classifier outperformed other classifiers and performed better.

3.6 Future Work:

The work presented in this paper has certain limitations which can be removed in future are as under:

- The size of data set can be increased.
- The features used for classification can be increased or decreased as per the latest studies.
- More classifiers can be applied to check their accuracy .

Reference:

[1]Ms. Israni, N R. Mr Jaiswal, A N. "A Survey on Various Phishing and Anti Phishing Measure", International Journal of Engineering and Research and Technology, Volume 4 (Issue 01), ISSN: 2278-0181, January 2015.

[2]https://docs.apwg.org/reports/apwg_trends_report_q4_2019.pdf

[3] WIKIPEDIA **PHISHING**
https://en.wikipedia.org/wiki/Phishing

[4]Alswailem, A. Alabdullah, B. Alrumayh, N. Dr.Alsedrani,A*," Detecting Phishing Websites Using Machine Learning," 2nd* International conference on computer applications & information , May 01 - 03, 2019 - Riyadh, Kingdom of Saudi Arabia added to IEEE Explorer on Jul 2019.

[5] Kulkarni, A. Brown, L.L.*"Phishing Websites Detection Using Machine Learning,"* International Journal of Advanced Computer Science and Applications, Vol. 10, No. 7, 2019 page no 8-13.

[6] Patil,V. Thakkar,P. Shah,C. Bhat,T. Prof Godse,S.P.*"Detection and Prevention of Phishing Websites using Machine Learning Approach,"* 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).

[7]Abdelhamid,N. Thabtah,F. Jaber,*H.A."Phishing Detection: A Recent Intelligent Machine Learning Comparision Based on Models Content and Features."*2019 IEEE International Conference on Intelligence and Security Informatics (ISI).

[8] Shirazi,H. Haefner,K. Ray,I. *"Fresh-Phish: A Framework for Auto-Detection of Phishing Websites,"* 2017 IEEE International Confrence on Information Reuse and Integration.

[9]Niu,W. Zhang, X. Yang, G. ma, Z. Zuho, Z."*Phishing Email Detection Using CS-SVM."* 2017 IEEE international Symposium on Parallel and Distributed Processing with Application and 2017 IEEE International Conference on Ubiquitous Computing and Communications(ISPA/IUCC).

[10] Sadeh,N. Tomasic,A. and Fette,I. "Learning to detect phishing emails", Proceedings of the16th international conference on world wide web, pp.649–656, 2007.

[11] Miyamoto,D. Hazeyama,H. and Kadobayashi, Y. "An evaluation of machine learning-based methods for detection of phishing sites," in International Conference on Neural Information Processing, pp. 539–546, Springer, 2008.

[12] Zhang, Y. Hong, J. I. and Cranor, L. F. "Cantina: a contentbased approach to detecting phishing web sites," in Proceedings of the 16th international conference on World Wide Web, pp. 639–648, ACM, 2007.

[13] Xiang,G. Hong, J. Rose, CP. Cranor,L. "CANTINA+: A Feature- rich Machine Learning Framework for Detecting Phishing Web Sites", School of Computer Science Carnegie Mellon University, ACM Society of computing Journal, 2015.

[14] https://www.statisticssolutions.com/what-is-logistic-regression/#:~:text=Logistic%20regression%20is%20the%20appropriate,variable%20is%20dichotomous%20(binary).&text=Logistic%20regression%20is%20used%20to,or%20ratio-level%20independent%20variables

[15] Kulkarni, A. Brown, L.L.*"Phishing Websites Detection Using Machine Learning,"* International Journal of Advanced Computer Science and Applications, Vol. 10, No. 7, 2019 page no 8-13.
[16] Quinlan,J.R. "*Induction of Decision Trees*", Machine Learning, vol. 1, no. 1, pp. 81–106, 1986
[17]https://deepai.org/machine-learning-glossary-and-terms/random-forest
[18] Kulkarni, A. Brown, L.L.*"Phishing Websites Detection Using Machine Learning,"* International Journal of Advanced Computer Science and Applications, Vol. 10, No. 7, 2019 page no 8-13.
[19] Vapnik, V. N. "*Support-vector networks*", Machine Learning, vol. 20 no. 3, pp 273–297, 1995.