Dr. Phish : Phishing Website Detector

Harish Kumar, Anshal Prasad, Ninad Rane, Nilay Tamane , Dr. Anjali Yeole VESIT , Mumbai

Abstract **:** Phishing is a common attack on credulous people by making them disclose their unique information . It is a type of cyber-crime where false sites allure exploited people to give delicate data. This paper deals with methods for detecting phishing websites by analyzing various features of URLs by Machine learning techniques.This experimentation discusses the methods used for detection of phishing websites based on lexical features, host properties and page importance properties. We consider various data mining algorithms for evaluation of the features in order to get a better understanding of the structure of URLs that spread phishing. To protect end users from visiting these sites, we can try to identify the phishing URLs by analyzing their lexical and host-based features.A particular challenge in this domain is that criminals are constantly making new strategies to counter our defense measures. To succeed in this contest, we need Machine Learning algorithms that continually adapt to new examples and features of phishing URLs.

Keywords **: phishing , anti-phishing , machine learning , cyber-crime ,**

cyber-attack

I . INTRODUCTION

Phishing is an act of fraud, attempting to target sensitive information or data, such as usernames, passwords, credit card details or any other details behind the veil of a trustworthy party in this digital world. It was seen that around 76% of businesses (that includes transactions worth billions of dollars) were subjected to phishing attacks in the past year alone. This is because people opt for an antivirus software which doesn't necessarily address phishing. Hence we have decided to propose a project by building a website using our anti-phishing software (Dr.Phish) which works by scanning any nefarious links or possible malware downloads. These programs warn against phishing urls with a high accuracy.

In our proposed project we are using machine learning since multiple datasets with a range of attributes are trained in order to achieve a high accuracy rate. The first phishing lawsuit was filed in 2004 against a Californian teenager for creating the imitation of a website named "America Online". With the help of this fake website, the hacker was able to gain access to sensitive information of users, access their credit card details to withdraw money from their accounts. Besides email and website phishing, there are other kinds of phishing, like vishing (Voice phishing), smishing (SMS phishing) and various other phishing techniques that the hackers are coming up with.

Phishing attacks target vulnerabilities that exist in systems due to the human factor. A cyber-attack costs a small business on average $53,987. Spear-phishing (the type of phishing utilized to target data) is aimed specifically at stealing sensitive information such as account credentials or financial information to use for nefarious purposes.Henceforth user should be aware about the environment against such intentional loss and here comes Dr.phish tool in the picture. Dr.Phish aims to create a secure environment against phishing attacks so that a warning is given to the user to make him/her alert whenever a malware is detected, protecting the privacy and information of the user. To create a secure environment against phishing attacks so that a warning is given whenever a malware is detected, protecting the privacy and information of the user.The main objective of our project is to warn the user from any tentative phishing attack via a website. The datasets of URLs are trained by ML algorithms. The website shows a warning message if the URL contains malicious content letting safety to the user from phished websites. To implement the logic we have used various ML algorithms like Random Forest, Logistic Regression and Decision Tree.With the help of our website along with continuing the process of notifying safety measures, this process will always be running by which the user is protected from phishing attacks.

II . LITERARY REVIEW

Many previous researchers have studied the detection of phishing URLs in general. Many of these works used machine learning approaches to detect these . The performance of their work depends primarily on the feature set, the dataset and the particular algorithm used.

Andronicus et al.[1] used random forest machine learning classifiers for classification of phished emails. They have aimed to maximize the accuracy and minimize the number of features required for classification. A content-based phishing detection approach which has high accuracy is presented. Here the authors proposed a model based on extracted features which appear in the header and HTML body of email which are classified using feed forward neural networks. The results indicate 98.72% accuracy of classification.

Gilchan Park et al.[2] aimed to extract robust features in order to discriminate between legitimate and phished emails. A comparison of sentence syntactic similarity and the difference in subjects and objects of target verbs between phishing emails and legitimate emails is done.

Tan et al.[3] further tested many more classifiers including decision tree (DT), gradient boosted trees (GDB), perceptrons (PE), K-nearest neighbor (KNN), and random forest (RF). However, the drawback of their work is they worked with a very large dataset with very limited number of features (24). Such training often suffers from over fitting. In other words their algorithm is comparable to a traditional blacklist.

Ma et al.[4] compared three classifiers, namely Naive Bayes (NB), support

vector machine (SVM) and logistic regression (LR) on a very good dataset. They used features like bag of words (BOW), IP address, WHOIS information, domain characteristics and geolocation. They cross verified the results using two datasets created from 4 sources. However, the simple classifiers used in their work are not suitable for deployment. Furthermore, the dataset had a ratio of 1 : 3 for malicious and benign URLs. In reality any such a system will see many more benign URLs than malicious URLs.

III. PROPOSED SYSTEM AND ARCHITECTURE

A] System Architecture :

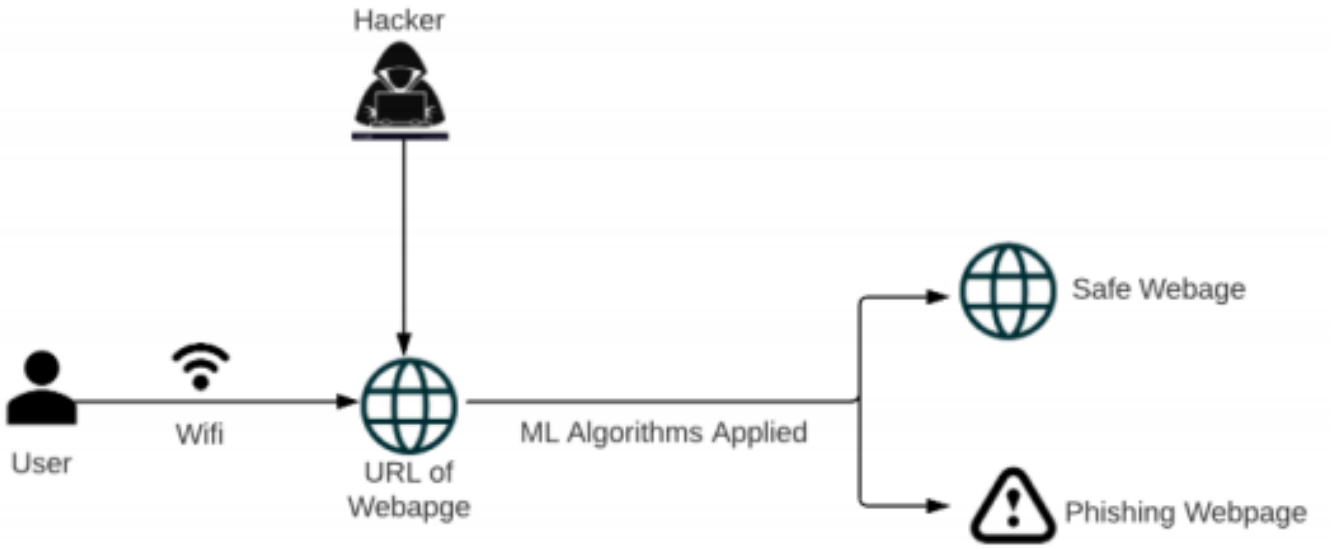

Fig. 1. Basic system architecture

The diagram in Fig. 1 shows the basic system architecture of our proposed system . When a user is connected over a network , the webpage is checked for spam by our website. If the website is identified as spam, a warning message pops up indicating that it is unsafe to use.

B] Dataset :

To evaluate our machine learning techniques, we have used the 'Phishing Websites Datasets' from UCI Machine learning repository.We used 11055 URL entries and then cleaned the data and splitted the dataset into training and testing sets. Out of which are legitimate 4898 URLs and 6157 are phishing URLs. Each instance

contains 31 features(Table 1) and 1 target variable to distinguish between legitimate and phishing URLs. Each feature is associated with a rule. If the rule satisfies, it is termed as phishing. If the rule doesn't satisfy then it is termed as legitimate. The features take three discrete values. '1' if the rule is satisfied, '0' if the rule is partially satisfied and '-1' if the rule is not satisfied.

Sr. No.	Feature	Description
1	id	Number to identify the website
2	having_IP_Address	If an IP address is used instead of the domain name in the URL
3	URL_Length	Phishers can use a long URL to hide the doubtful part in the address bar
4	Shortening_Service	Links to the webpage that has a long URL.
5	having_At_Symbol	Using the @ symbol in the URL leads the browser to ignore everything preceding the @ symbol
6	double_slash_redirecting	The existence of // within the URL which means that the user will be redirected to another website
7	Prefix_Suffix	Phishers tend to add prefixes or suffixes separated by (-) to the domain name
8	having_Sub_Domain	Having subdomain in URL
9	SSLfinal_State	Shows that website use SSL
10	Domain_registration_length	Based on the fact that a phishing website lives for a short period
11	Favicon	If the favicon(icon) is loaded from a domain other than that shown in the address bar , its a phishing URL
12	port	To control intrusions, it is much better to merely open ports that you need
13	HTTPS_token	Having deceiving https token in URL
14	Request_URL	Request URL examines whether the external objects contained within a web page such as images, videos, and sounds are loaded from another domain
15	URL_of_Anchor	An anchor is an element defined by the < a > tag. This feature is treated exactly as a Request URL.
16	Links_in_tags	It is common for legitimate websites to use ¡Meta¿ tags to offer metadata about the HTML document; ¡Script¿ tags to create a client side script; and ¡Link¿ tags to retrieve other web resources.
17	SFH	If the domain name in SFHs(Server Form Handler) is different from the domain name of the webpage
18	Submitting_to_email	A phisher might redirect the users information to his email
19	Abnormal_URL	It is extracted from the WHOIS database. For a legitimate website, identity is typically part of its URL
20	Redirect	If the redirection is more than four-time , its a phishing URL
21	on_mouseover	Used for hiding link
22	RightClick	It is treated exactly as Using onMouseOver to hide the Link
23	popUpWindow	Showing pop-up windows on the web page
24	Iframe	IFrame is an HTML tag used to display an additional webpage into one that is currently shown
25	age_of_domain	If the age of the domain is less than a month
26	DNSRecord	Having the DNS record
27	web_traffic	This feature measures the popularity of the website by determining the number of visitors
28	Page_Rank	Page rank is a value ranging from 0 to 1. PageRank aims to measure how important a webpage is on the Internet
29	Google_Index	This feature examines whether a website is in Google's index or not
30	Links_pointing_to_page	The number of links pointing to the web page
31	Statistical report	If the IP belongs to top phishing IPs or not

Table 1: Features of URL

C] Classifiers :

This section will give a detailed description of the classifiers used . We have used Random Forest, Logistic Regression, Decision Tree for the detection of phishing URLs

● ***Random Forest*** **:** Random Forest[7] is a very popular machine learning formula that belongs to the supervised learning technique. It is used for each Classification and Regression problems in Machine Learning.It supports the thought of ensemble learning, that is a method of combining multiple classifiers to reach the solution of a complex problem and to enhance the performance of the model. Random Forest is a classifier that contains variety of decision trees on numerous subsets of the given dataset and takes the typical to enhance the predictive accuracy of that dataset.Rather than hoping on one decision tree, the random forest takes the prediction from every tree and supported the maximum votes of predictions, and it predicts the ultimate output.The bigger variety of trees within the forest ends up in higher accuracy and prevents the matter of overfitting.

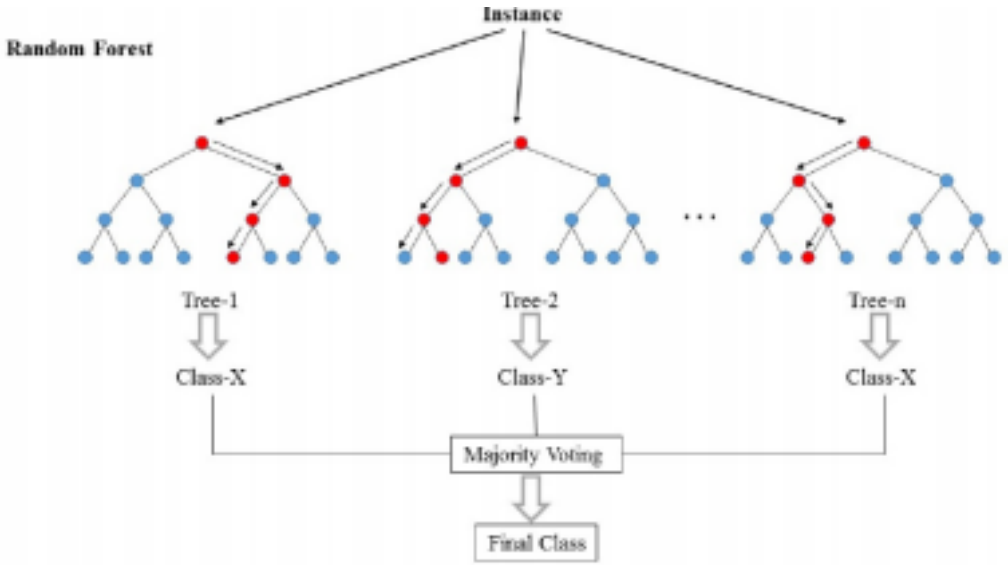

Fig 2. Random Forest Working

Random Forest works in two-phase: 1st is to form the random forest by combining N decision trees, and second is to form predictions for every tree created within the 1st part. The Working method is explained within the below steps and diagram: Step-1: choose random K data points from the training set.

Step-2: Build the decision trees related to the chosen data points
Step-3: Select the amount N for call trees that you need to make.
Step-4: Repeat Step one & two.
Step-5: For brand new data points, notice the predictions of every decision tree and assign the new data points to the class that wins the maximum votes.

- ***Logistic Regression*** **:** Logistic Regression[8] is the correct regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistical regression could be a prognostic analysis. logistic regression is employed to explain the relationship between one dependent binary variable and one or additional nominal, ordinal, interval or ratio-level independent variables. Logistic regression is known as for the function used at the core, the logistic function additionally referred to as the sigmoid function, was developed by statisticians to explain properties of increase in ecology, rising quickly and maxing out at the carrying capability of the surroundings.It's an S-shaped curve that can take any number and map it into a worth between zero and one, but never exactly at those limits.

  1 / (1 + e^-value)

  Where e is the base of the natural logarithms (Euler's number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform.

- ***Decision Tree*** **:** Decision Trees[9] are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be seen in two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes and the decision nodes are where the data is split. In the Decision Tree the major challenge is to identify the attribute for the root node in each level. This process is known as attribute selection. Here we have two popular attribute selection measures:
  1. Information Gain :When we use a node in a decision tree to partition the training instances into smaller subsets the entropy changes. Information gain is a measure of this change in entropy.
  Definition: Suppose S is a set of instances, A is an attribute, Sv is the subset of S with A = v,and Values (A) is the set of all possible values of A, then

$$G(S,A) = I(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v)$$

  2. *Gini Index* : Gini Index is a measure of how often a randomly chosen element would be incorrectly identified.It means that an attribute with lower

Gini index should be preferred.Sklearn
supports "Gini" criteria for Gini Index and by default,takes "gini" value.The formula for the calculation of the Gini Index is given below

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

D] Proposed Approach :

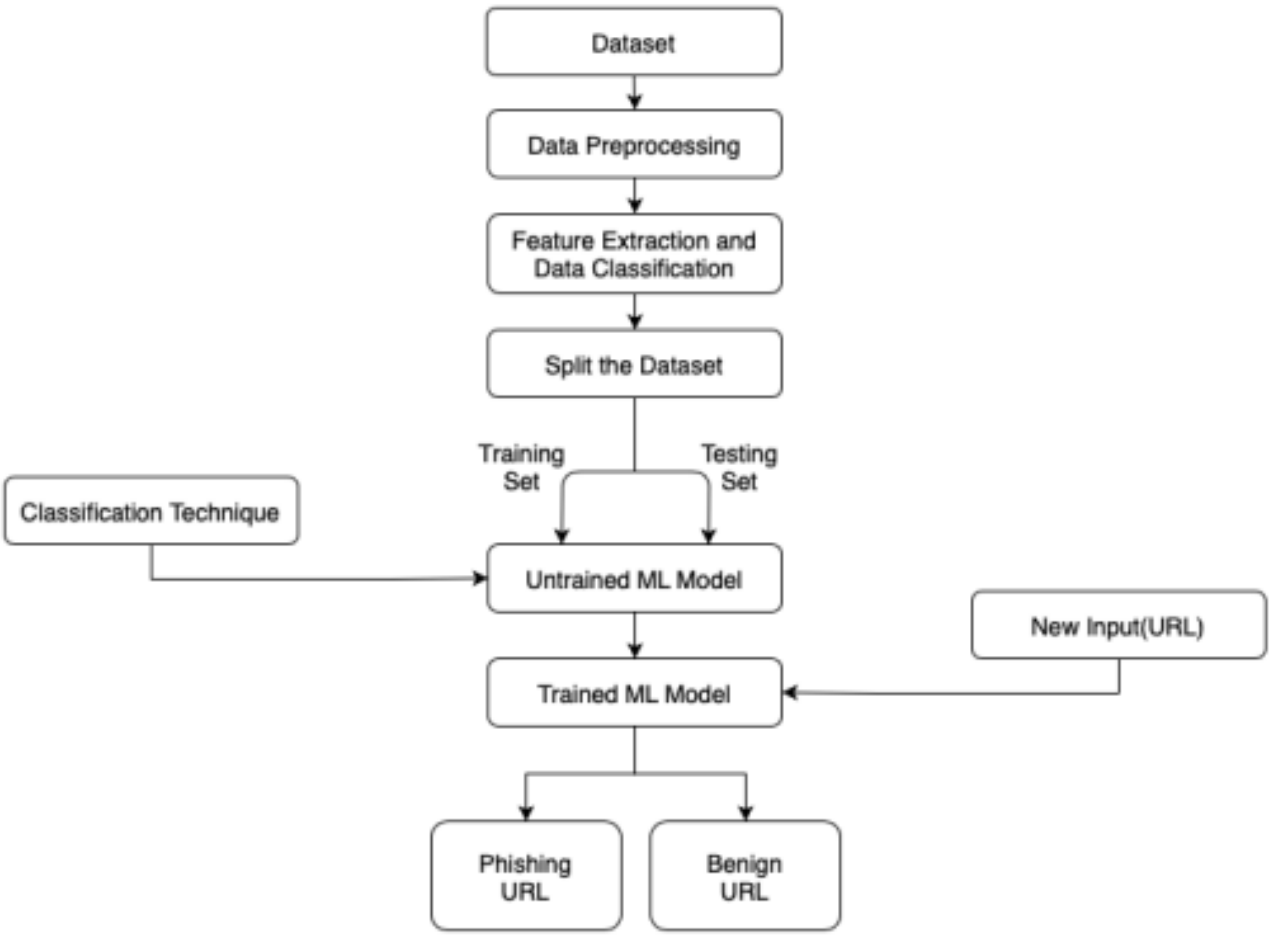

Fig 3. Proposed Approach

The first stage is data gathering and collection at a centralized location after which comes data preprocessing which does the job of removing unnecessary data which makes no sense in running meaningful analysis . Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. The three steps in data preprocessing are data cleaning, data transformation and data reduction. After this the features of URLs are extracted and then they are classified as either -1, 0, 1.

Next the dataset is splitted into two subsets i.e. a training set and a testing set which are fed to an untrained ML Model which uses any one of the classifiers such

as Random Forest , Decision Tree , Logistic Regression to detect phishing URLs. Now a new data input(URL) is fed to the trained model which further predicts whether it is a phishing URL or a benign URL.

IV. EXPERIMENTAL RESULTS

This section demonstrates the experimental studies to investigate the predictive accuracy of various Machine Learning Classification Algorithms on the same dataset and also compares it to the existing Machine Learning Techniques.

- The dataset used comprises 11055 URLs out of which 6157 are malicious and 4898 are legitimate websites.
- Each instance had 31 features which were extracted and then fed to the untrained ML classifiers.
- The training set and the testing set consists of 8844 URLs(80%) and 2211(20%) URLs respectively.
- The Machine Learning Classification Technique used for the detection of phishing URLs are Random Forest(RF), Logistic Regression(LR) and Decision Tree(DT) .

The following Table compares the obtained predictive accuracy of our model with the predictive accuracy of the existing models on an average :

Classification Technique	**Obtained Accuracy(%)**	**Existing Technique Accuracy(%)**
Random Forest(RF)	95.10	95.50
Logistic Regression(LR)	92.25	94.10
Decision Tree(DT)	89.23	93.90

Table 2 : Accuracy Comparison

From these results, it is evident that Random Forest gives better performance in terms of classification accuracy as compared to others in terms of accuracy.

The following Graph compares the predictive accuracies of the above mentioned three ML classifiers :

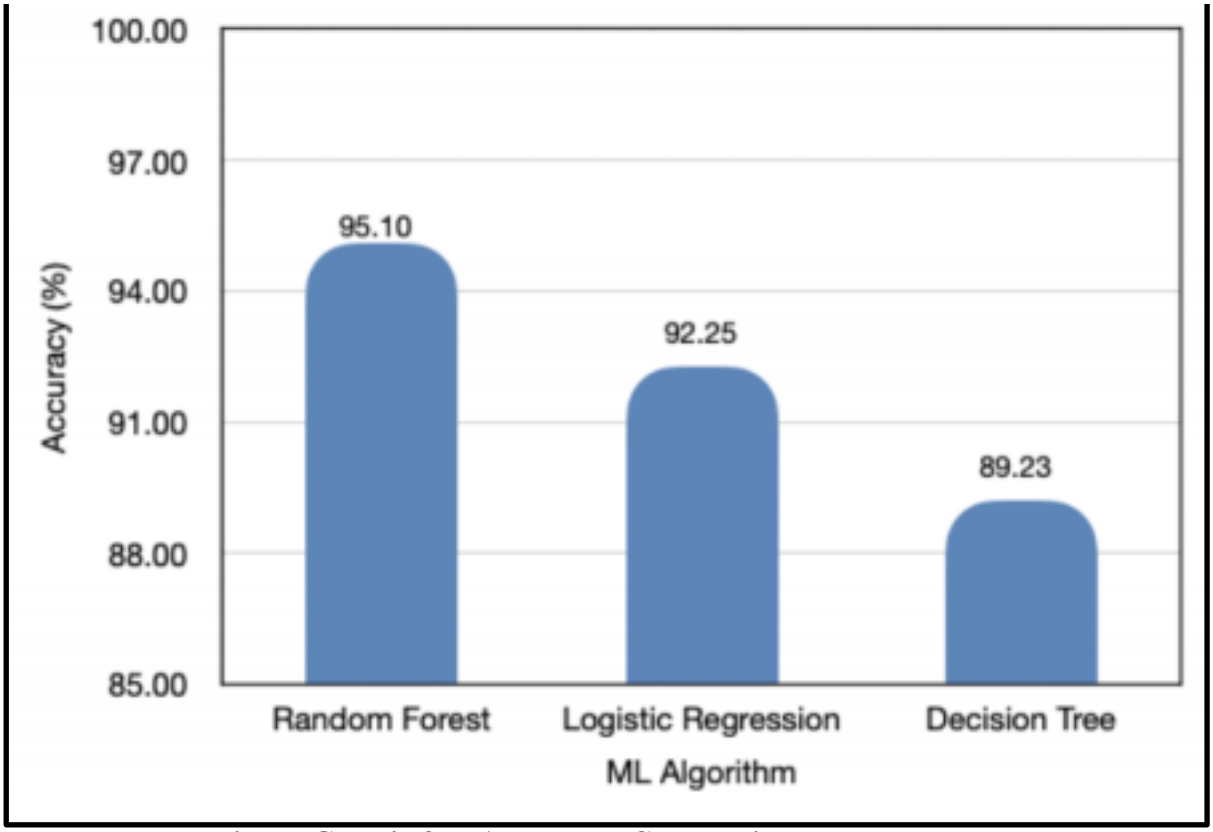

Fig 4. Graph for Accuracy Comparison

V. CONCLUSION

This paper aims to enhance detection methods to detect phishing websites with the help of machine learning algorithms.The attempts made by different researchers to solve this problem through the use of machine learning classifiers was discussed. The features of phishing URLs were extracted with the help of python programming language. The model was fed into Random Forest, Logistic regression, and Decision tree to detect phishing URLs. The classification results were encouraging as the highest accuracy of 95.10% was achieved using random forest algorithms for phishing URLs. Also results show that classifiers give better performance when we use more data as training data. This work has produced encouraging results, however, the dataset used may not necessarily replicate a real life scenario . In future works, the proposed system can be improved by increasing the dataset and creating a browser extension. By adding a variety of URLs of both type phished and legitimate, the system would be closer to the real life scenario

where fraudsters are day by day improving their techniques. Using real life samples would enable us to deploy a formal system that can be used across organizations and privately to prevent users from being victims to phishing attacks.

VI. REFERENCES

[1] Andronicus A. Akinyelu and Aderemi O. Adewumi. Classification of Phishing Email using Random forest Machine Learning Technique 2014.

[2] Gilchan Park, Julia M. Taylor, Using Syntactic Features for Phishing Detection 2015, https://arxiv.org/ftp/arxiv/papers/1506/1506.00037.pdf

[3] G. Tan, P. Zhang, Q. Liu, X. Liu, C. Zhu, and L.Guo,"Malfilter: A lightweight real-time malicious url filtering system in large-scale networks"2018 IEEE ISPA/IUCC/BDCloud /SocialCom/SustainCom.IEEE, 2018,pp. 565-571.

[4] J. Ma, L. K. Saul,S. Savage, and G. M. Voelker, "Beyond blacklists : learning to detect malicious web sites from suspicious urls" in Proceedings of the 15th ACM SIGKDD in international conference on Knowledge discovery and data mining . ACM , 2009 , pp. 1245-1254 .

[5] R. Verma and A. Das, "What's in a url: Fast feature extraction and malicious url detection," in Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics. ACM, 2017, pp. 55–63.

[6] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from urls," Expert Systems with Applications, vol. 117, pp. 345–357, 2019.

[7] Random Forest :
- https://www.javatpoint.com/machine-learning-random-forest-algorithm
- https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d

[8]Logistic Regression :
- https://machinelearningmastery.com/logistic-regression-for-machine-learning/

[9]Decision Tree :
- https://www.geeksforgeeks.org/decision-tree/
- https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html