IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PHISHING WEBSITE PREDICTION USING SVM AND NAÏVE BAYES TECHNIQUES

[1]S. R. Jagadeesh, [2]Dr. N. Syed Siraj Ahmed,
[1]PG Scholar, [2]Assistant Professor
Master of Computer Applications
Madanapalle Institute of Technology and Science
Angallu, Madanapalle.

Abstract: In recent years, the online phishing attack has emerged as one of the most serious web security issues, with phishers gaining access to sensitive financial information about web users in order to commit financial fraud. Phishing is a type of cyber-attack that uses a spoof email as a weapon, and it has been on the rise in recent years. Innocent internet users who accidentally click on a malicious link risk revealing personal information such as their Mastercard PIN, login credentials, financial information, and other sensitive information. Attackers might use a variety of techniques to persuade victims to give personal information. During this project, we will choose key phishing URL attributes that an attacker will utilise to deceive internet users into taking the attacker's intended action. For accurate phishing detection and to reduce false positive detection, a combination of support vector machine (SVM) and naive Bayes algorithms is used.

Index Terms **- Phishing attack; data sets; URL classification; phishing URL; attackers; machine learning; classifiers; Internet.**

I. INTRODUCTION

In the once decades, the operation of internet has been increased extensively and makes our live simple, easy and transforms our lives. It plays a major part in areas of communication, education, business conditioning and commerce. A lot of useful data, information and data can be attained from the internet for particular, organizational, profitable and social development. The internet makes it easy to give numerous services through online and enables us to pierce colorful information at any time, from anywhere around the world.

Phishing is the act of transferring a indistinguishable dispatch, dispatches or vicious websites to trick the philanthropist / internet druggies into discovering delicate particular information similar as personal identification number (PIN) and word of bank account, credit card information, date of birth or social security figures. Phishing assaults affect hundreds of thousands of internet druggies across the globe. Individualities and associations have lost a huge sum of plutocrat and private information through Phishing attacks.

II. STATE OF THE ART

Rashmi Karnik et al., proposed a model of classification method, kernel-based approach. In this we categories phishing . This method produces estimated accuracy of 95% in detecting the phishing and malware sites.

Andrei Butnaru et al., used a supervised Machine Learning algorithm to block phishing attacks, based on novel mixture phishing attacks and compare with Google Safe browsers.

Vahid Shahrivari et al., proposed a one of the most successful techniques for identifying these malicious works is Machine Learning. It is because of most Phishing attacks have same features which can be noticed by Machine learning techniques. In this many machine learning-based classifiers are used for forecasting the phishing websites. The main advantage of machine learning is the ability to create flexible models for

specific tasks like phishing detection. Since phishing is a classification problem, Machine learning models can be used as a forceful tool.

Ammara Zamir et al., proposed a framework for identifying phishing websites using heaping model. Information gain, gain ratio, Relief-F, and recursive feature elimination (RFE) are some of the feature selection algorithms that can be used to analyse Phishing characteristics. The greatest and weakest traits are combined to create two features. Bagging is used in principal component analysis using several Machine learning algorithms, including random forest [RF] and neural network [NN]. Two heaping representations heaping1 (RF + NN + Bagging) and heaping2 (kNN + RF + Bagging) are applied by merging highest scoring classifiers to progress classification accuracy.

Nguyet Quang Do, Ali Selamat et al., conducted a study on phishing detection and proposed a four different deep learning technique, includes deep neural network (DNN), convolution neural networks (CNN), Long Short-term memory (LSTM), and gated recurrent unit (GRU). To analyse behaviour of these deep learning architectures, extensive experiments were carried out to examine the impact of parameter tuning on the performance accuracy of the deep learning models. In which each model shows different accuracies from different models.

Ashit Kumar Dutta proposed a URL detection procedure based on Machine Learning methods. An RNN is used for identifying the phishing URL. It is evaluated with 7900 malicious and 5800 genuine sites, respectively. The outcome of this method shows a good concert compare to recent tactics.

Atharva Deshpande et al., proposed a combination of machine learning algorithms and natural language processing methods to detect the phishing domain appearances, the feature that distinguish them from real domains.

Ms. Sophiya Shikalgar et al., proposed a machine learning classifiers and methods to detect phishing website using Hybrid machine learning approach is a combination of different classifiers working together which gives a good prediction result. Each of classifiers have its own way of working and classification. Uses a data of URLs which contains 2905 URLs which is in unstructured form.

Nureni Ayofe Azeez et al., tried to handle this challenge, attempts have been made to address two major problems. The first is how can the suspicious URL's be recognized on social networks and how can internet users can be protected from unreliable and fake URLs on the social network. It adapts six machine learning methods – AdaBoost, Gradient Boost, random forest, Linear SVM, decision tree and Naïve Bayes classifier for training using features obtained from the social network and for additional processing. A total of 532,403 posts were analysed. At last 87,083 posts were considered suitable for training the models. AdaBoost performs well among all with an accuracy of 95% and a precision of 97%.

Ademola Philip Abidoye and Boniface Kabaso proposed a machine learning technique to accurately classify the dataset to identify the phishing URLs features that can be used by the attackers.

R. Kiruthiga and D. Akila explained a novel way of detecting phishing websites using machine learning methods and proposes a classification model in order to classify the phishing attacks. Also presents a way to detect phishing email attacks using natural language processing and machine learning produces a good accuracy.

Arun Kulkarni and Leonard L. Brown has established a system that uses machine learning methods. The main aim of this is to develop these methods of defence applying quite a few methods to federations websites. For that used four classifiers: decision tree, naïve Bayesian classifier, support vector machine (SVM) and neural network. These classifiers were tested with a dataset containing 1,353 real world URLs where each could be categorized as a legitimate site. This classifier shows the overall accuracy of 90% of the time.

Orunsolu et al., proposed a model of advanced machine learning based forecasting model is used to improve the efficiency of anti-phishing schemes. The predictive model consists of feature selection model used for construction of an effective feature vector. It is based on datasets containing of 2541 phishing instances and 2500 benign instances. Using 10 – fold cross- validation, it results a amazing presentation of 0.04% False Positive.

III. METHODOLOGY

In this segment we going to learn about the classifiers used in machine learning to envisage phishing. Here we intend to explain our proposed methodology to detect phishing website. In this we divided into 2 parts one for classifiers and another to explain our proposed system.

A. *Machine learning classifiers and methods to perceive the phishing website*

Distinguishing and recognizing phishing websites is really an intricate and energetic problem. Machine learning has been extensively used in numerous areas to produce automated results. Phishing attacks can

take numerous forms, including dispatch, website, malware, and voice. This paper focuses on detecting website phishing (URL) using the Hybrid Algorithm Approach. It is a mix of different classifiers that work together to improve the system's accuracy and estimate rate.

Depending on the application and the nature of the dataset used we can use any classification algorithms. As there are various applications, we cannot discriminate which of the algorithms are superior or not.

- **Naïve Bayes Classifier:** This classifier can also be known as a Propagative Learning Model. The bracket is erected on Bayes Theorem. In simple words this classifier will assume that the actuality of specific features in a class isn't related to the presence of any other point. Naïve Bayes classifier is a probabilistic machine learning model that's used for classification task. This classification algorithm is veritably important useful to large datasets and is veritably easy to use.
- **Support Vector Machine (SVM):** This is also one of the supervised and simple to use classification algorithms. It can be used in both classification and regression applications; however, classification applications are preferred. SVMs differ from other classification algorithms in that they employ the distance between the nearest data points of all classes to determine the decision boundary. The maximum margin classifier or maximum margin hyper plane is the decision boundary created by SVMs. The classification is based on the differences between the classes, which are data set points in various planes.

B. Proposed System

The phishing dataset and genuine URL's is given to the classification which is then pre-processed so that the data is in the usable arrangement for scrutiny. The dataset has around 11056 records. The dataset has 30 features like having_Ihaving_IP_Address, URLURL_Length, Shortining_Service, having_At_Symbol, etc., For each the input values ranges from 0 to 100, The output range is 0 to 100, whereas the input range is 0 to 100. The phishing attributes are represented by binary integers 0 and 1, with 0 indicating that the attribute is present and 1 indicating that it is absent.

After the data has been taught, we will use a machine learning algorithm to analyse the dataset. The machine learning algorithms have previously been discussed in the previous section. Next, we'll use hybrid classification, in which we combine two classifiers, Support Vector Machine (SVM) and Nave Bayes, to estimate the accuracy of the phishing URL detection, resulting in the expected outcome.. It's also known as a hybrid strategy to data testing, and we propose using the above-mentioned combination of two classifiers for this. The data will then be tested, and the forecast accuracy will be evaluated, which will be higher than the current method.

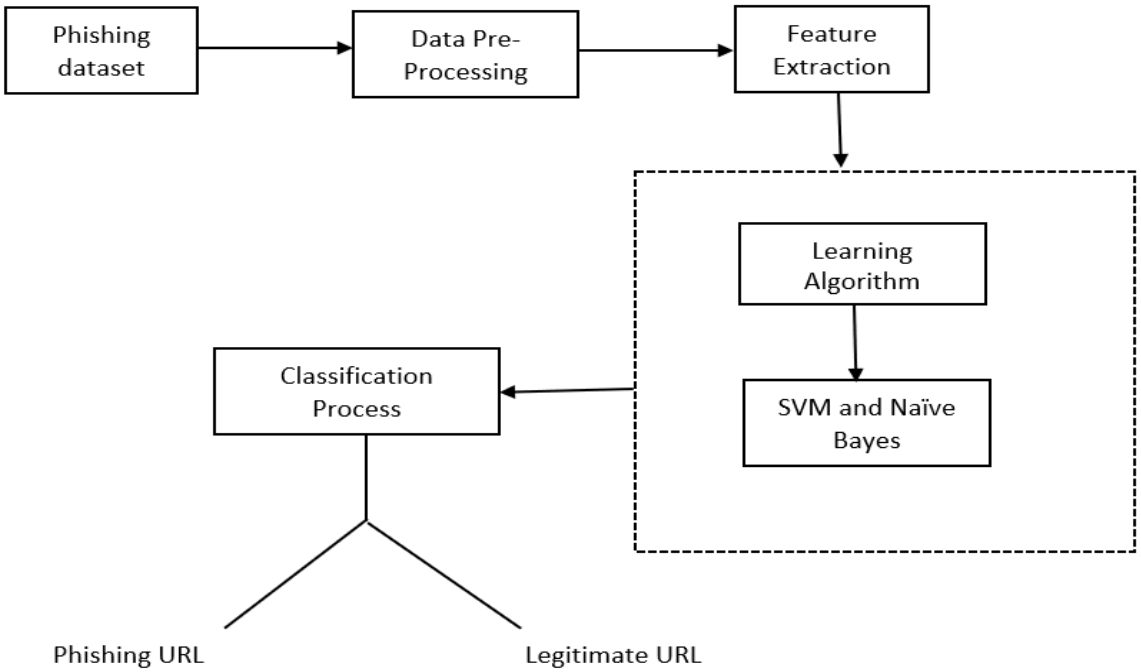

Fig. 1: Proposed system Block Diagram

In the training phase, We should employ samples from known classes, which means that samples labelled as phishing should only be detected as phishing. Similarly, valid URLs will be discovered in samples that have been labelled as such. These characteristics must be present in the dataset used for machine learning. There are numerous machine learning algorithms, each with its own operating mechanism, as we saw in the previous

part. The existing system uses any one of the appropriate machine learning algorithms for the recognition of phishing URL and forecasts its accuracy.

IV. SYSTEM OVERVIEW

System strategy is used to empathetic the building of system. We have clarified the flow of our system and the software used in the system in this section.

A. System Flow

The Fig. 2 explains the flow chart of the system design, we will describe each of the components of the flow chart in each section below. To get structured data we do feature generation or feature extraction of the data at the pre-processing stage. The techniques like SVM and Naïve Bayes classifiers to notice the phishing and genuine websites.

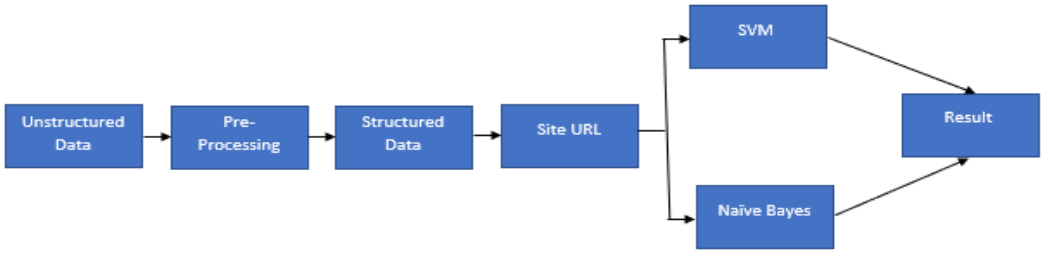

Fig. 2 System flow

- **Data set:** The data of URLs is obtained from Phishtank website, where Phishtank is an anti-phishing site. It contains 11064 records which is in structured data. Our main objective is to notice whether the url is phishing or valid based on the features mined.

In Preprocessing we have done feature mining where The URLs are conveyed to the feature extractor, which extracts feature values through the predefined URL-based features. The features have allotted binary values 0 and. 1 which specifies that feature is present or not and as shown in table below. The extracted feature values are kept as input and passed to the classifiers.

Table 1: Unstructured data

Phish-id	url
4912175	https://srnbc-card.com.60osx47.cn
4912845	https://sbcglobalnet967.weebly.com/
4912842	https://srnbc-card.com.62bgmf0.cn/
4912460	https://srnbc-card.com.70qckck.cn/
4912136	https://srnbc-card.com.68c4103.cn/

A structured dataset is given to the classifiers. For determining whether a URL is phishing or valid, we employ two classification methods: SVM and Nave Bayes. The classifier will now determine whether or not the requested website is a phishing site. When a page request is made, the feature extractor receives the URL of the requested site. It uses predefined URL-based features to cite the feature values. The classifier uses these feature values as an input. After that, we'll be able to tell whether the site is phishing or not.

Table 2: Structured data

Phish-id	IPAddress	URL Length	Shortining_ \Service	havingAtSymbol	double slash
4912175	1	1	1	1	1
4912845	1	1	1	1	1
4912842	1	0	1	1	1
4912460	1	0	1	1	1
4912136	1	0	1	1	1

- **URL Features:** According to Table 3, attributes 1 to 4 are linked to suspicious URL outlines and characters. Characters like _@' and _//' are uncommon in URLs. Feature 5 is used to identify freshly created phishing sites using the expected methodology. The subdomains of these phishing sites are unusually long.

Table 3: URL Features

Sr. No	**Feature name**	**Description**
1	IP address	Whether domain is in the form of IP address
2	URL Length	Length of URL
3	Suspicious character	Whether url has _@`,_//.
4	Prefix & suffix	Whether URL has _-`
5	Subdomain length	Length of subdomain
6	Number of '/'	Number of '/' in URL
7	HTTP tokens	Whether URL use https.
8	Abnormal words in URL	Whether url as phishing terms.

The feature 8 is another new feature that reflects current phishing trends. This feature includes seven phishing terms are secure, websrc, ebaysapi, signin, banking, confirm, login. We employ them in our phishing detection technique. We have already debated the dissimilar classifiers in the above sections.

V. IMPLEMENTATION

This section gives information about the execution situation and throws light on the actual steps for the execution of dataset to get good accuracy to expect phishing by using various classifiers combination.

A. *Hardware requirements*

The following hardware is used for the execution of the system:

- 8 GB RAM
- 256 GB HDD (Minimum)
- Intel core i5 8th Gen

B. *Software requirements*

The following software is used for the execution of the system:

- Windows 7 and above versions
- Python 3.6.0
- Visual Studio Code

C. *Implementation steps*

In this section, we will going discuss about the actual steps which were executed while doing the experiment. We shall discuss the stepwise procedure used to analyse the data and to predict the phishing. We have used unstructured data which consists only url. There are 11064 urls obtained from the internet. Which consists of both phishing and genuine url where most of urls obtained are phishing.

1. First, we have unstructured data of urls from Phishtank website.
2. In Preprocessing, feature generation is done where eight features are generated from unstructured data. These features are length of url, http tokens, suspicious character, prefix/suffix, number of slashes, phishing term, length of subdomain, url IP address.
3. Next, a structured dataset is constructed, with binary values (0,1) for each feature, which is then sent to the various classifiers.
4. Next we train the two different classifiers and compare their performance on the basis of accuracy two classifiers namely SVM, Naïve Bayes.
5. Then classifier detects the given url based on the training data that is if the site is phishing it shows error and if legitimate it opens that page in browser.

6. We compare the accuracy of different classifiers and found Naïve Bayes is the best classifier which gives the maximum accuracy.
7. Below are the screen shots for the execution process.

VI. OBSERVATIONS AND RESULT

A. Observation

As we discussed earlier sections, we have used two different classifiers to predict and identify if the website is Phishing or genuine. Comparisons of these classifiers have been displayed below in the accuracy table.

Table 4: Observation Table

Classifier	Test	Train	Accuracy
SVM	2211	8844	92.5
Naïve Bayes	2211	8844	89.3
RNN	2211	8844	88.6
NN	2200	8855	87.2
DNN	2200	8855	86.9

B. Result

We have got the expected results of testing the site is phishing or not by using two different classifiers.

VII. CONCLUSION AND FUTURE SCOPE

A. Conclusion

It has been discovered that phishing attempts are quite important, and we must develop a technique to identify them. Because phishing websites can expose sensitive and personal information about users, it's even more crucial to address this problem. This problem can be easily solved by using of any Machine learning algorithm with the classifier. We already have the algorithm with classifier which gives good prediction rate of the phishing besides, but after our survey that it will be better to use this approach for the prediction and further improve the accuracy prediction rate of those websites. we have found that our system provides us with 100% of accuracy for SVM and 98% of accuracy for Naïve Bayes classifier. Hence, we found that the best among all the classifiers is SVM Classifier which shows maximum accuracy. The proposed technique is much more secured as it detects new and previous phishing sites.

B. Future Prospects

If we have a structured dataset of phishing in the future, we will be able to detect phishing far faster than any other technique. In the future, we will be able to combine any two or more classifiers to get maximum accuracy. We also intend to investigate other phishing strategies that make use of lexical, network, and content-based aspects, as well as HTML and JavaScript features of webpages, to increase system efficiency.

VIII. REFERENCES

[1] Ms. Sophiya Shikalgar, Mrs. Swati Narwane (2019), Detecting of URL based Phishing Attack using Machine Learning. (vol. 8 Issue 11, November – 2019)
[2] Rashmi Karnik, Dr. Gayathri M Bhandari, Support Vector Machine Based Malware and Phishing Website Detection.
[3] Arun Kulkarni, Leonard L. Brown, III[2], Phishing Websites Detection using Machine Learning (vol. 10, No. 7,2019)
[4] R. Kiruthiga, D. Akila, Phishing Websites Detection using Machine Learning.
[5] Ademola Philip Abidoye, Boniface Kabaso, Hybrid Machine Learning: A Tool to detect Phishing Attacks in Communication Networks. (vol. 11 No. 6,2020)
[6] Andrei Butnaru, Alexios Mylonas and Nikolaos Pitropakis, Article Towards Lightweight URL-Based Phishing Detection.13 June 2021
[7] Ashit Kumar Dutta (2021), Detecting phishing websites using machine learning technique. Oct 11 2021
[8] Nguyet Quang Do, Ali Selamat, Ondrej Krejcar, Takeru Yokoi and Hamido Fujita (2021) Phishing Webpage Classification via Deep Learning-Based Algorithms: An Empirical study.
[9] Ammara Zamir, Hikmat Ullah Khan and Tassawar Iqbal, Phishing website detection using diverse machine learning algorithms.
[10] Vahid Shahrivari, Mohammad Mahdi Darabi and Mohammad Izadi (2020), Phishing Detection Using Machine Learning Techniques.
[11] A. A. Orunsolu, A. S. Sodiya and A.T. Akinwale (2019), A predictive model for phishing detection.
[12] Wong, R. K. K. (2019). An Empirical Study on Performance Server Analysis and URL Phishing Prevention to Improve System Management Through Machine Learning. In Economics of Grids, Clouds, Systems, and Services: 15th International Conference, GECON 2018, Pisa, Italy, September 18-20, 2018, Proceedings (Vol. 11113, p. 199). Springer.
[13] Desai, A., Jatakia, J., Naik, R., & Raul, N. (2017, May). Malicious web content detection using machine leaning. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1432-1436). IEEE.
[14] H. N. Security. (2019). Phishing attacks at highest level in three years. Available: https://www.helpnetsecurity.com/2019/11/07/phishingattacks-levels-rise/
[15] L. James, Phishing exposed. Canada.: Syngress, 2005.
[16] Nagaraj, K., Bhattacharjee, B., Sridhar, A. and Sharvani, G. (2018), "Detection of phishing websites using a novel twofold ensemble model", Journal of Systems and Information Technology, Vol. 20 No. 3, pp. 321-357.
[17] Jain, A.K. and Gupta, B. (2019), "A machine learning based approach for phishing detection using hyperlinks information", Journal of Ambient Intelligence and Humanized Computing, Vol. 10 No. 5, pp. 2015-2028.
[18] Fette, I., Sadeh, N. and Tomasic, A. (2007), "Learning to detect phishing emails", Proceedings of the 16th International Conference on World Wide Web, ACM, pp. 649-656