Phishing Website Detection System Using Machine Learning

N. Mohan Rao[1] ,Namala Lakshmi Sireesha[2]
Assistant Professor[1] and M.Tech Student Department of Computer Science
V.S.M College of Engineering,Ramachandrapuram,East Godavari,Andhra Pradesh

---***---

Abstract - *Phishing is frequently a routine attack on people wherein fake websites are used to trick individuals into divulging all of their personal information. Phishing record keeping process tool URLs are used to steal personal data, including user names, passwords, and online banking activity. Attackers who utilise phishing techniques use websites with rectangular diplomas as a visual and semantic spoof of the real websites. Phishing strategies have advanced swiftly as the age has progressed, however this may be avoided by using anti-phishing tools to spot phishing. A potent tool frequently utilised in the context of phishing attacks is the machine planning to apprehend. This study examines the machine learning skills employed in detecting and detection methods.*

Key Words: *Phishing, Phishing Websites, Detection, Machine Learning, Information.*

1. INTRODUCTION

Phishing mimics the features and options of emails as well as makes them seem identical to the real thing. It resembles the real supply very closely. The consumer believes that this email is from a legitimate employer or business. This forces the user to click on the links provided in the phishing email and visit the phishing website. These phishing websites were developed to imitate the design of a clever website. The phishers coerce people into listing up their private information by sending false messages, asking them to confirm their account, and other means so that they can list up the information they want to use against them. They come up with strategies that prevent users from constantly having a choice but to visit their fake website. The most dangerous criminal activity in the online world is phishing. Since the majority of users log on to access the services offered by governmental and financial institutions, phishing attacks have significantly increased over the past several years. Phishers started using this as a lucrative business to make money.

The reason phishers commit this crime is because it is incredibly trustworthy to do so, it doesn't cost anything, and it works. Phishing may be illegal. The scam will really discover the email address of the person it's genuinely trying to find.

It is currently free to send emails to everyone around the world and mail identification is available every day. These attackers have a terrible lack of resources, making it difficult for them to quickly and effectively advance valuable knowledge. The phishing scams have an impact on fraud, statistics loss, malware infections, etc. The crucial information of a user, such as their password, OTP, credit/debit card numbers, CVV, sensitive knowledge related to business, medical knowing, confidential information, etc., is information that those cybercriminals are interested in at some point. Frequently, these criminals also acquire data that could give them direct access to their social media accounts and emails. [4]

There are numerous programmes, methods, and algorithms that are used to identify phishing. These are employed at the academic and professional levels. Take for instance that to cover the original domain selection the scam assaulter will perceive terribly long and complicated name of said domain. A phishing address in addition to the parallel internet page have numerous characteristics that may be unique from the address. This is frequently horribly obvious.

Usually, they utilise the domain call victimisation website's data science address. On the other hand, if you want to not be relevant to the unique legitimate website, they will also utilise a shorter domain option. There are other unique functions, such as domain-based options, page-based features, and content-based features, which can be used to detect phishing websites in addition to the address-based function of phishing detection. Phishers use a variety of methods, including digital communication, VOIP, faked links, and fake websites, to target unprotected users. Making fake websites that, in terms of design and content, behave like authentic internet pages is incredibly honest. Even their genuine websites' images may be reflected in the substance of those websites. These websites were created with the intention of obtaining sensitive information from users, such as account numbers, login credentials, MasterCard and debit card password, etc.

The total number of phish discovered in the second quarter of 2020 dropped from 263,538 in the first quarter to 233,040, per the APWG 2Q report. These totals are higher than the 190,942 seen 3Q 2021 and the 180,577 found in the 4Q 2021 of 2021. The SAAS/webmail-focused sector saw growth with 21% of typical phishing assaults. The total number of phishing attempts discovered in the first quarter of 2018 was 263, up 46 percent from the 180,577 attempts discovered in the fourth quarter of 2021. Additionally, it was significantly higher than the 190,942 recorded in the third quarter of 2017. During the first quarter of 2018, 262,704 unique phishing reports were filed to APWG, compared to 233,613 in the fourth quarter of 2021 and 296,208 in the third quarter of 2021.

2. RELATED WORK

Numerous studies have been done on security [10–13], including secure routines, intrusion detection, threat detection, and the security of smart grids. Web pharming is the practise of pretending to be a reliable website on the internet in order to acquire sensitive data, like usernames, passcode, and credit card numbers, frequently for illegal purposes.

Researchers offer the following solutions to recognise web phishing attacks:

The most straightforward way to determine whether a specific website is an internet phishing site is to utilise a whitelisting a blacklist. Additionally, we could look up the Urls in a handful of databases before making a choice. By using the blacklist, Pawan Prakash et al. provided two methods to identify phishing websites.

Examining the characteristics of the URL is another method for detecting phishing. For instance, a URL may occasionally resemble a well-known web page URL or include unusual letters. One idea of intra-URL relatedness was employed by Samuel Marchal et al., who computed it by taking advantage of the capabilities found in the words that make up a URL and basing their calculations on search query data from Google and Yahoo. The phishing URLs are subsequently found using these functions in a device mastering-based kind from a genuine statistic set. This method is environmentally friendly and cost-effective since it makes advantage of the URL's built-in knowledge, which has a quick detection rate and a cheaper price. Due to the fact that the core of the method is to commit fraud using online content, we are unable to completely translate the characteristics of phishing into the language of a URL. If only the records of both the URL are checked, this strategy will produce a lower detection rate because phishing attackers are likely comfortable with URLs and simply modify them to avoid detection.

3. LITERATURE SURVEY

a special category approach that makes use of the function extraction method based on heuristics. In this, the extracted capabilities are divided into three groups: third-party-based features, hyperlink-based features, and URL obfuscation features. Moreover, this model is the only

Betting on the quality and quantity of the education set and extracting the Broken links feature has the drawback of longer execution times for sites with more links. A method that prioritises unique frequency capabilities was proposed by Chunlin et al. In order to reveal the result that is more accurate for the type of dangerous URLs, they combined statistical appraisal of URL with a scientifically learning approach.

This study [4] describes a method for identifying phishing emails by analysing linguistic communication and applying machine learning. To find malicious intent, it is customary to search the syntax of the text. Each sentence is decoded using a natural language understanding (NLP) approach, which also detects the semantic roles of the individual words in the sentence. For the purpose of creating the blacklist of malicious pairs, computer supervised learning is used.

Additionally, we compared the accuracy of five machine learning algorithms: Generalized Linear Model (GLM), Generalized Additive Model (GAM), Gradient Boosting (GBM), and Random Forest (RF) [5]. (GAM).

Each algorithm's accuracy, precision, and recall were calculated and compared. Python is used to help with the selection of website properties, while R, an open-source programming language, is used for performance evaluation. Performance of the top 3 algorithms, in particular SVM, Variational Forest, and Naive Bayes, is compared.

They have thought of using a rule-based approach to identify phishing websites. Due to this sincere rule transformation, they argue that relationship classification algorithms are superior to other algorithms everywhere. Although they extracted 16 alternatives and achieved 92.67% accuracy, it is incorrect to say that the purposeful algorithmic rule could be multiplied for an affordable detection rate. Authors [6] developed a version of a method for identifying phishing websites using the universal help locator identification method while abusing the Variational Forest algorithmic software. The three stages of the presentation are performance analysis, heuristic classification of the data, and parsing. The capacity set is examined using parsing. This research [7] plans a framework to recover functions that are adaptable and straightforward using novel techniques. Data is gathered from Google's reliable URLs and PhishTank's authentic URLs.

4. METHODOLOGY

In this section, we'll look at the different classifiers used in system learning to identify phishing. We will also back up our proposed methodology for identifying phishing websites and assaults with evidence.

We will demonstrate numerous classifiers and approaches that can be used to distinguish between legitimate and phishing websites in section 4.1. Our proposed system can be justified in section 4.2.

We will go into more detail regarding the image retrieval of the URL in section 4.3. The data sets will be trained on and tested using the extracted features.

Machine learning classifiers and methods to detect the phishing website

Simply said, spotting and classifying phishing websites is a challenging and constantly changing endeavour. In several fields, machine learning has been heavily used to generate solutions. Hoax attacks can be conducted via a variety of channels, such as email, websites, malware, SMS, etc. Using the Hybrid Algorithm Approach, we focus on identifying website phishing (URL) in this paper. A hybrid algorithmic approach combines different classifier algorithms that operate together to produce an incredible prediction rate and increase the system's accuracy.

Any grouping calculations mentioned will be used, depending on the application and also the principle of the dataset used. We are unable to distinguish whether the algorithms are better or not because they are used in such a variety of applications. Each classifier operates and categorises data in a unique way.

Let us discuss each of them in details. [8]

- **Naive Bayes Classifier:** The term "Generative Learning Model" can also be used to describe this classifier. The categorization in this case relies on the Bayes Theorem and anticipates independent indicators. Simply said, this classifier will anticipate that the existence of explicit highlighted in a class does not imply the existence of further components. If there is any dependence between the strengths of different traits or on their proximity, this will be taken account as a soul commit to the certainty of the yield. Large datasets can benefit greatly from these arrangement calculations, which are also quite simple to use.

- **Random Forest**: This set of classification rules is similar to compiling a certain type of learning strategy. Working with a group of decision bushes formed at the level of training records and for the length of the output of such class, that may be the pattern of class or forecast regression for the regression and many other tasks character wood This decision tree classifier accuracy practise overfits the training data set.

- **Support vector machine (SVM):** This is another algorithm for classification that is straightforward to use and is directed. Each informational point in this calculation is displayed in a region of space, often known as an n-dimensional surface, where the number "n" refers to the number of informational highlights.

After the model has been trained, it is crucial to evaluate and confirm the performance of the classifier that will be used. We have now listed all of the benefits and downsides of each of their classifiers in the section above. Therefore, in order to improve the accuracy of both prediction and classification, we prefer to recommend using a few categories that are capable of being combined. By using the combination explicitly described in this part, we have a propensity to improve the precision and construct it higher. We value each of the models and utilise Naive Bayes and Stochastic Forest. After applying the classification, results are produced, and the URLs are divided into phishing and legitimate URLs. The legitimate URLs may be on a white list in the database, while the phishing URLs are banned in the information.

Proposed System

Pre-processing is done within the application so that the data is in a working format for study. There are roughly 30 characteristics of fake websites that are utilised to separate them from genuine ones. The characteristics and values of each category of phishing are clearly stated. For each URL, the desired features are retrieved, and valid stages of inputs are located. Each risk associated with a phishing website is then given one of these values. The double no 0 and 1 that appears the attribute is present or not is used to address the phishing properties.

The categorised data, which contain samples from both legitimate and phishing regions, must be used in the education phase. In the case that we try this, typing won't ever again be a challenge for identifying the phishing space. We should only employ samples whose recommendations are familiar to us, therefore samples that we identify as phishing should only be identified as phishing. Similar to this, authentic samples are recognised as having a valid URL. These features must actually be present in the dataset being used for machine learning. There are so many different machine learning algorithms, and each set of guidelines has its unique method for operation. We have already seen this in the prior chapter. The winning system forecasts the accuracy of the phishing URL detection and uses all of the allowed system learning strategies. [9]

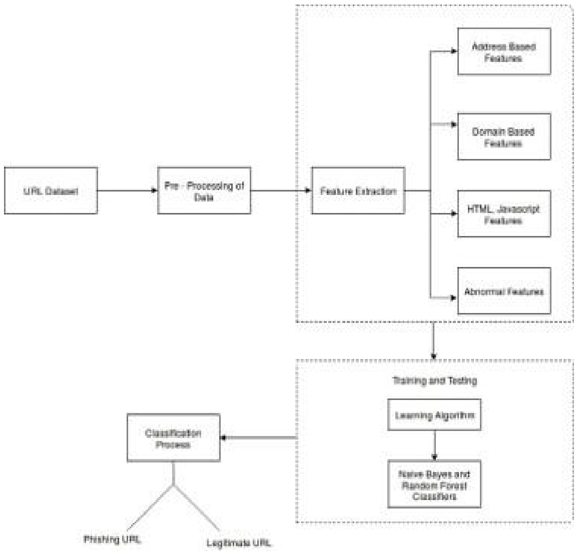


Fig -1: Proposed System block diagram

Lexical Feature Analysis

Lexical functions, as opposed to the topic of the website that they influence, are the textual characteristics of the URL itself. URLs are textual content structures that can be interpreted by user programmes in a common manner at some point. Browsers interpret each URL into command that let them find the server hosting the district and determine where the location or resource is located on that server through a multi-step procedure. URLs contain the following common syntax to help with this AI process.

<protocol>://<hostname><path>

An example of URL resolution is shown below:

Protocol Top Level domain

Host name

https://accounts.google.com/ServiceLogin?service=mail&passive=true&rm=false&continue=https://mail.google.com/mail/&ss=1&scc=1&ltmpl=default&ltmplcache=2

Path

The URL's protocol section specifies the network protocol that should be used to get the requested resource. The most widely used protocols are File Transfer Protocol, HTTP with Secure Channel, and Hypertext Transfer Protocol (HTTP) (ftp). The hostname serves as the web server's unique identifier. It occasionally has a Transmission Control protocol / internet (IP) address that is understandable by computers, but more frequently, and more importantly from the standpoint of the user, it has a name. A URL's "path" is comparable to a file's "trail name" on a public computer. The following steps are utilised in our study to isolate the lexical features from the URL list: The scratchpad is filled with the URLs of legitimate websites that have been collected from alexa.com plus dmoz.org, and as a result, the record is saved inside the computer.

FEATURE EXTRACTION

LONG URL:

To conceal the Suspicious Part, a lengthy URL is employed. The URL is considered to be phished if it has more than or on par with 54 characters.

URL's having "@" Symbol:

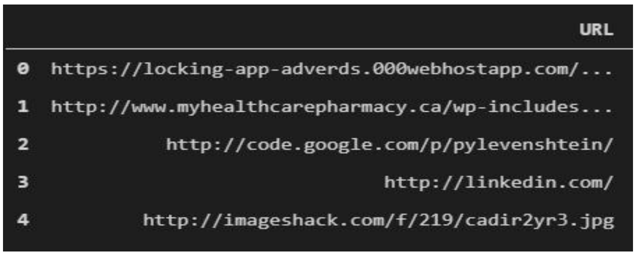
The "@" sign causes the reader to ignore everything before the "@," making the URL appear to be phished. The actual address frequently comes after the "@" sign.

IF {URL Having @ Symbol→ Phishing
URLOtherwise→ Legitimate}.

Sub-Domain and Multi Sub-Domains

The legitimate URL link has two dots within the URL since we will ignore typing "www.". If the number of dots is comparable to three then the website is evaluated as "Suspicious".
However, if the dots are larger than three, then it will be categorized as "Phishy".

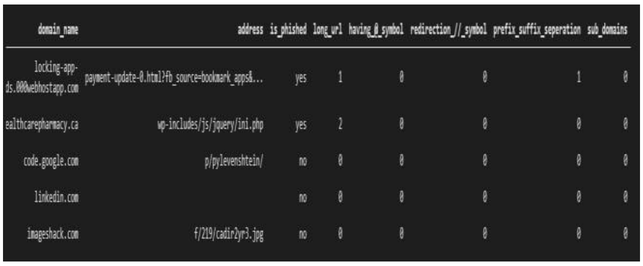
Data set: The information of URLs is gotten from the Phishtank site, where Phishtank is an enemy of the phishing site. It contains 2905 URLs which is in an unstructured structure. Our primary target is to identify whether the URL is phishing or authentic dependent on the highlights removed.



```
                                               URL
0  https://locking-app-adverds.000webhostapp.com/...
1  http://www.myhealthcarepharmacy.ca/wp-includes...
2                http://code.google.com/p/pylevenshtein/
3                                   http://linkedin.com/
4           http://imageshack.com/f/219/cadir2yr3.jpg
```

Fig -2: Unstructured Data

In Preprocessing, we have performed the component extraction where The URLs are transmitted to the element extractor, which concentrates values through the predefined URL-based highlights. The highlights have allocated twofold qualities 0 and 1 which demonstrates that component is available or not as appeared in the figure beneath. A structured dataset is given to the classifiers.



```
         domain_name                                address is_phished long_url having_@_symbol redirection_//_symbol prefix_suffix_seperation sub_domains
        locking-app-
ds.000webhostapp.com payment-update-0.html?fb_source=bookmark_apps&...        yes        1               0                     0                        1           0
ealthcarepharmacy.ca                       wp-includes/js/jquery/ini.php        yes        2               0                     0                        0           0
     code.google.com                                    p/pylevenshtein/         no        0               0                     0                        0           0
        linkedin.com                                                             no        0               0                     0                        0           0
       imageshack.com                                   f/219/cadir2yr3.jpg         no        0               0                     0                        0           0
```

Fig -3: Loading the data in our program

Table -1: URL Features

Sr. No	Feature name	Description
1	IP address	Whether Domain is in the form of an IP address
2	Length of URL	Length of URL
3	Suspicious character	Whether URL has ‗@ ', ‗//'

URL Features: Referring to Table 1., features from 1 to 4 are associated with suspicious Characters such as ‗@ 'and ‗// 'rarely appear in a URL. At present, to keep a client from distinguishing that a site isn't authentic, phishing destinations ordinarily conceal the essential area; the URLs of these phishing locales have curiously long subdomains.

5. IMPLEMENTATION AND TESTING

This segment gives data about the execution condition and illuminates the real strides for the usage of the dataset to show signs of improvement exactness to anticipate phishing by utilizing various classifiers mixes.

Hardware requirements

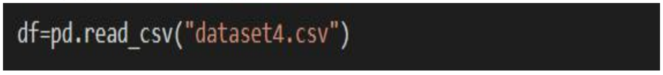
The following hardware was used for the implementation of the system:

- 4 GB RAM
- 10GB HDD
- Intel 1.66 GHz Processor Pentium 4

Software requirements

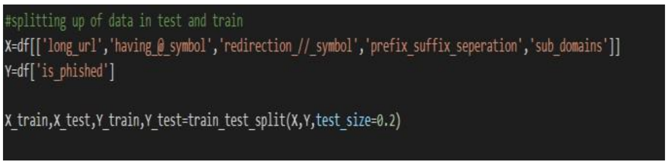
The following software was used for the implementation of the system:

- Windows 7
- Python 3.6.0
- Visual Studio Code



```
df=pd.read_csv("dataset4.csv")
```

In this section, we will talk about the means which were actualized while doing the examination. We will provide a piece of evidence for the stepwise method accustomed to split the knowledge and to foresee the phishing. We have utilized unstructured information that comprises just URLs. There are 2905 URLs gotten from the Phishtank site which comprises of both phishing and genuine URL where the majority of the URLs got are phishing.



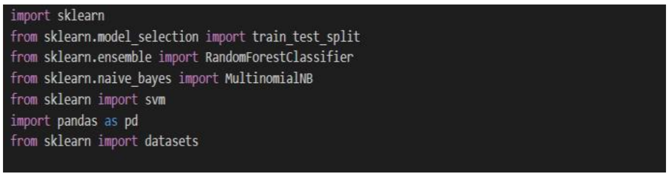
```
#splitting up of data in test and train
X=df[['long_url','having_@_symbol','redirection_//_symbol','prefix_suffix_seperation','sub_domains']]
Y=df['is_phished']

X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.2)
```

1. We have collected unstructured data of URLs from Phishtank website.
2. In pre-processing, feature generation is done where nine features are generated from unstructured data. These features are length of an URL, URL has HTTP, URL has suspicious character, prefix/suffix, number of dots, number of slashes, URL has phishing term, length of subdomain, URL contains IP address.
3. After this, an organized dataset is made in which each detail incorporates the paired (0,1) which is then passed to the various classifiers.
4. Next, we train the three unique classifiers and analyse their presentation based on exactness three classifiers utilized are SVM, Naive Bayes and Random Forest.
5. At that point, the classifier identifies the given URL dependent on the preparation information that is if the site is phishing it prompts the user that the website is phished and if genuine, it prompts the user that the website is legitimate.
6. We look at the exactness of various classifiers and discovered Random Forest as the best classifiers which gives the most extreme precision.

6. RESULTS

We have efficiently calculated the consequences of numerous classifiers which might be SVM, Naïve Bayes, Random Forest.

On comparison of resultant values, we chose to put into effect the Random Forest classifier in our datasets. Steps to obtain the accuracy of various classifiers:



```
import sklearn
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn import svm
import pandas as pd
from sklearn import datasets
```

- Initially, we import all the packages which can be implemented in our project.

Fig -4: Importing the required packages

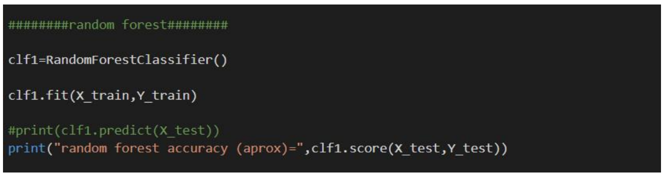
- We will load the data sets for testing and training.

Fig -5: Loading the data set

- Now, we will do splitting up of data for training and testing. We will use 20% of data set for testing.

Fig -6: Splitting up of Data Set for testing and training

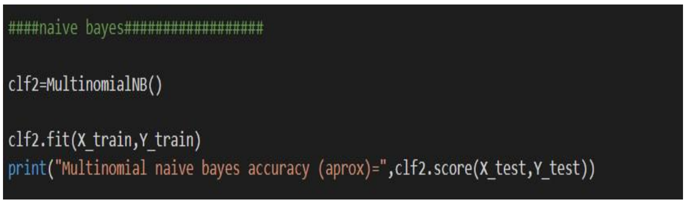
- We will calculate the accuracy of Random Forest classifier.



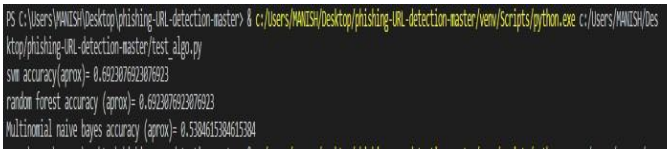
```
########random forest########

clf1=RandomForestClassifier()

clf1.fit(X_train,Y_train)

#print(clf1.predict(X_test))
print("random forest accuracy (aprox)=",clf1.score(X_test,Y_test))
```

Fig -7: Calculation of the accuracy of Random Forest classifier

- We will calculate the accuracy of Naïve Bayes classifier.



```
####naive bayes#################

clf2=MultinomialNB()

clf2.fit(X_train,Y_train)
print("Multinomial naive bayes accuracy (aprox)=",clf2.score(X_test,Y_test))
```



```
PS C:\Users\MANISH\Desktop\phishing-URL-detection-master> & c:/Users/MANISH/Desktop/phishing-URL-detection-master/venv/Scripts/python.exe c:/Users/MANISH/Des
ktop/phishing-URL-detection-master/test_algo.py
svm accuracy(aprox)= 0.6923076923076923
random forest accuracy (aprox)= 0.6923076923076923
Multinomial naive bayes accuracy (aprox)= 0.5384615384615384
```

Fig -8: Calculation of the accuracy of Naïve Baye's classifier

- Now, we will compare the results obtained after calculating the accuracy of various classifiers.

Fig -9: Comparison of accuracy of various classifiers

- Upon comparison, we found that accuracy of Random Forest Algorithm is highest and is considered best for our data set.

7. CONCLUSIONS

It is discovered that phishing assaults are unbelievably essential and it's significant for us to invite an instrument to distinguish it. As fundamental and private data of the client is spilled through phishing sites, it turns out to be progressively basic to require care of this issue. This issue is handily understood by utilizing any of the AI calculations with the classifier. We have just got classifiers that give a decent expectation pace of phishing additionally, yet after our overview that it'll be smarter to utilize a half breed approach for the forecast and further improvement of the exactness expectation pace of phishing sites. We've seen that the current framework gives less precision so we proposed a fresh out of the box new phishing strategy that utilizes URL based highlights and furthermore, we created classifiers through a few AI calculations.

The main findings of our preliminary work include:

- Phishing URLs and domains show some characteristics that are different from other URLs and domains.
- Phishing URLs and domain names have altogether different lengths contrasted with different URLs and domain names inside the Internet.
- A large number of the phishing URLs contained the name of the brand they focused on.

REFERENCES

[1] Web Phishing Detection Using a Deep Learning Framework. Hindawi Wireless Communications and Mobile Computing Volume 2018, Article ID 4678746, 9 pages.

[2] Phishing Websites Detection Using Machine Learning R. Kiruthiga, D. Akila. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019

[3] Detection of URL based Phishing Attacks using Machine Learning. Published by: International Journal of Engineering Research & Technology (IJERT) http://www.ijert.orgISSN:2278-0181 Vol. 8 Issue 11, November-2019R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[4] T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.

[5] Shad and S. Sharma, "A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology," pp. 425–430, 2018.

[6] S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. Icicct, pp. 949–952.

[7] M. Aydin and N. Baykal, "Feature extraction and classification phishing websites based on URL," 2015 IEEE Conf. Commun. NetworkSecurity, CNS 2015, pp. 769–770, 2015.

[8] 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.

[9] Lakshmi, V. S., & Vijaya, M. S. (2012). Efficient prediction of phishing websites using supervised learning algorithms. Procedia Engineering, 30, 798-805.

[10] P. Yi, T. Zhu, Q. Zhang, Y. Wu, and L. Pan, "Puppet attack: A denial of service attack in advanced metering infrastructure network," Journal of Network and Computer Applications, vol. 59, no. 1, pp. 325–332, 2016.

[11] P. Yi, T. Zhu, Q. Zhang, Y. Wu, and J. Li, "A denial of service attack in advanced metering infrastructure network," in Proceedings of the 2014 IEEE International Conference on Communications (IEEE ICC 2014), pp. 1029–1034, IEEE, Sydney, Australia, June 2014.

[12] S. Xiao,W. Gong,D. Towsley,Q. Zhang, andT. Zhu, "Reliability analysis for cryptographic key management," in Proceedings of the IEEE International Conference on Communications (IEEE ICC 2014), Sydney, Austrailia, June 2014.