# An Optimal Technique For Predictive Phishing Detection

**D.N. Goswami , Manali Shukla, Prof. Anshu Chaturvedi**

**Abstract:** Web Security is a major issue and evolving as a regressive challenge in the field of Information Technology. Nowadays besides the existence of several security mechanisms, web threats are taking shape of web attacks which are causing potential damage to web users. Phishing is one of the attack which has been frequently practiced by the attackers over the past decades in order to fulfill their malicious goals. Phishing aims to ascertain confidential information of web users by redirecting them to web replica of legitimate web created by attackers to spoof web users. Identification of Phishing is most significant because this may further lead to evolve into more vulnerable attacks such as Ransomware, botnets etc. This paper proposes an optimal technique for phishing detection that identifies the presence of malicious websites by assessing the characteristics of only those features associated with uniform resource locators. In this paper, we proposes a binary classifier algorithm which predicts the presence of malicious websites to classify Uniform Resource Locator into two categories phishy and nonphishy which is determined by identifying the presence of features responsible for phishing attack. Python IDE is used for assessing accuracy of algorithm over different size of datasets.

**Index Terms:** Botnets ,cybercrime, Internet , Phishers, Ransomware, Subdomain , URLs.

## 1 INTRODUCTION

Internet users are increasing very rapidly and sharing their sensitive data on various digital platforms without knowing the affect of malicious attacks. Several awareness programs are being conducted online and offline but still statistical reports show that the users are being spoofed very frequently over the Internet. Phishing is one of the vulnerable attack which is being used as a weapon by the cyber criminals to perform cybercrimes. Phishing can be defined as a malicious activity which is carrying out to steal sensitive and confidential information of internet users in a very short time. Phishers accomplish their malicious goals by redirecting users to a fake website which looks like the authentic one to capture the user's data without giving any clue to users. Cyber criminals uses this confidential data to do financial frauds and also further take it to the next level of Ransomware attack by encrypting user's files and keeps them on ransom. Therefore there is a huge scope of research in this field to detect and report Phishing as earliest in order to avoid such damages. Phishing can be identified by using various features which are broadly classified into uniform resource locators based features and content based features. In [1] author have determined thirty features which are used by the attackers to perform Phishing. In this paper we have designed an algorithms which employs a subset of features as suggested in [1]. It is being observed several times that the computational complexity of an algorithm increases as number of dimensions (features) increases and as a result of which reduction in quality of results appears. Therefore dimensionality reduction is employed in order to select subset of features without sacrificing the performance of underlying algorithm. The authors in [2] suggested about various techniques of dimensionality reduction available in the present scenario.


- *D.N. Goswami S.O.S.in Computer Science And Applications, Jiwaji University Gwalior, India*
- *Manali Shukla , S.O.S.in Computer Science And Applications, Jiwaji University Gwalior, India*
- *Prof. Anshu Chaturvedi Department Of CSE & IT, Programme M.C.A Madhav Institute Of Technology &Science (M.I.T.S.) Gwalior, India*


The proposed research work has also adapted an optimal approach of analyzing only the affect of significant features of uniform resource locators without sacrificing the performance measures such as accuracy and efficiency of classifier. Attackers are frequently modifying the standard features of Uniform resource locators so that they can mislead web users to their malicious websites. The typical structure of uniform resource locators includes Protocol, hostname and path. Attackers cleverly modifies these parts of uniform resource locators either by increasing or decreasing length of uniform resource locators or by adding special characters in between the host name etc. This research work is performed using uniform resource locators features which includes number of dots, presence of hyphen, length of uniform resource locators, presence of @, presence of double slash, number of subdirectories, number of sub domain, length of domain and presence of internet address in the domain part. Generally attackers increases number of dots in the uniform resource locators in order to modify number of sub domains in the uniform resource locators to navigate towards malicious links. The hyphen is generally not appears in the legitimate uniform resource locators. Intruders have a tendency to add prefixes or suffixes separated by hyphen to the domain name as a result of which users feel that they are using a legitimate website. Attackers uses long uniform resource locators to conceal the suspicious part in the address bar. Usage of special characters such as '@ in the uniform resource locators also helps Phishers because it instruct the browser to overlook all the things former to the '@' symbol on the other hand the legitimate addresses frequently follows this symbol. Presence of double slash inside the uniform resource locator path is responsible for redirection towards the malicious website. An increase in number of subdirectories is also another important feature of phishing websites. Phishers add multiple sub domains in their malicious uniform resource locators These types of uniform resource locators can not be considered as legitimate and may lead us towards the Phishing attack. There is great difference between the length of malicious domains and legitimate domains. Phishers modifies this length feature to fulfil their aim of phishing. Several researchers have given serious attention on the presence of internet address in domain section of the uniform resource locators.

## 2 RELETED WORK

Many researchers have made their contribution in this field and numerous approaches are being implemented to find out Phishing attack . The performance of phishing detection algorithm entirely depends on the feature selection. The use of lexical and host-based features of the associated Uniform resource locators has been proposed by the author in their research work in [3] but it is particularly designed for online algorithms to identify suspicious URLs. The author in [4] proposed an approach which blocks the internet address of the intruders through source and content filters. This approach also educates, train users and report to service providers. Their approach could handle phishing attack in their network approximately 95%, but nowadays intruders are keep on changing different internet address in order to spoof users, In such cases this approach can be proved bit complex and more time spanning. A combinational approach is proposed by the author in [5] which classify the normal and malicious activities using k- means and ID3 algorithm . This approach uses ID3 algorithm due to which sometimes problem of over fitting arises. The author in [6] has compared phishing websites with legitimate websites and observed lack of security in comparison of authentic websites.. A lexical features based model has been developed by the author in [7] that analyzed three steps for phishing identification . Their approach is completely relies on lexical features for phishing identification The research work proposed by the author in [8] for identification of Chinese phishing e- business websites which uses domain features. In another study of research proposed in [9] the use of page signature which is generated using term frequency and provided to search engine to determine the difference between real page and resultant page through tag comparison and cosine similarity and it also uses the Google page ranking information. The use of term frequency-based signatures are easier to compute but are strongly relies on search engine 's ranking system. The use of Lexical based, keyword based, search engine based, and reputation based features has been proposed by the author in [10]. They have employed 138 features in order to determine phishing but in our opinion, instead of having increased number of dimensions or features reduction of dimensions or features can give better results. An another heuristic based technique has been proposed by the author in [11] and they determined that the Random forest as the best classifier in their research work. Their technique is useful but they used number of features. The authors in [12] proposed use of auto update whitelist of legitimate websites and they uses two component in their research first domain and IP address matching and secondly observe the features of the hyperlinks from source code. Their approach of analyzing the features of the hyperlinks from the source code is bit complex process. The author in [13] has proposed neural network based approach for classification which is self structuring and uses 17 features for phishing detection .This approach is suitable only for server side and do not give solutions for client side.

## 3 PROPOSED WORK

This paper proposes an algorithm which determines the presence of malicious uniform resource locators in order to classify it into two categories phishy and nonphishy.

This Binary Classifier takes extracted features of uniform resource locators as input and gives output in the form of result which classifies uniform resource locators into two types phishy and non- phishy which is shown below in the figure 1. This research work is divided into following phases.

i. Extraction of Features

ii. Assignment Of weight

iii. Classification

iv. Prediction

Extraction of features plays an important role in this work because it serves as input to the proposed classifier. According to our approach , Each and every feature which is being used by the attackers for spoofing users has its own significance in the process of phishing identification. Several uniform resource locators are analyzed to determine which features are frequently used in the uniform resource locators for transforming them into illegitimate web addresses. This research work identifies that Nine prime features which are frequently used by the Phishers in the uniform resource locators. Phishtank[14] which is an online repository of phishing uniform resource locators contains numerous such malicious uniform resource locators that signifies the presence of these prime nine features which are extracted from the uniform resource locators in order to perform this classification

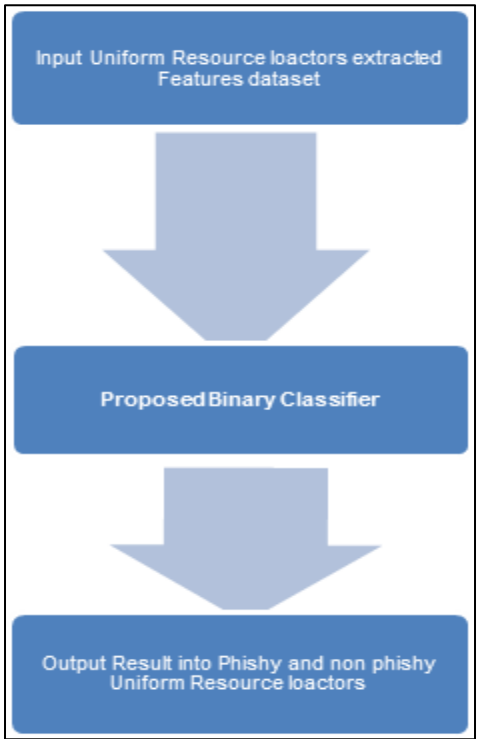


***Figure 1:*** *Proposed Binary Classifier*

This research approach identifies the occurrence of each feature as a baseline for the phishing detection and assigns equal weight for its each occurrence in the uniform resource locators. The proposed binary classifier classifies the uniform resource locators into two types phishy and nonphishy after assessing the weight of each uniform resource locators as

shown in the pseudo code as given below.

```
Binary_Classifier(TWO CLASS SOLUTION)

Input:
      featureSet   // Input the dataset of the
                      Extracted features
Output:
OfeatureSet      // Dataset consisting of output values of predicted label from this algorithm
// Proposed algorithm

Initialize threshold_Count to zero
for each row of the dataframe  check the value of
the  threshold_count

            Read the dataset

           .If threshold_count is positive then


                  output " Phishy "

           elseif threshold_count is negative

                 output "nonphishy"
      else

                  print “error message"
```

The input to this binary classifier is a dataset which contains values of all the nine features which we have extracted from uniform resource locators data set using Python IDE. Initially the threshold_count is set to zero which is used to count the weight of each phishing feature if it appears in the Uniform Resource locators .This threshold count will get incremented with each occurrence of features which results into malicious uniform resource locators and the proposed algorithm evaluates the value of this threshold_count to determine whether it is positive or negative. If the value of threshold_count results into positive, then the binary classifier will classifies the corresponding uniform resource locators as phishy and if it’s value turn out negative then the classifier classifies the corresponding uniform resource locators as nonphishy. This algorithm fits very well during training phase as well as it gives satisfactory prediction accuracy during testing it over different sizes of datasets of malicious uniform resource locatorsIn the next section, we have analyzed the results collected during the prediction phase which is implemented to analyze its prediction accuracy over un-labeled datasets of malicious uniform resource locators.

## 4 RESULTS

In this section, the proposed binary classifier is evaluated to determine its prediction accuracy over varying size of datasets containing malicious Uniform Resource Locators. These results are evaluated to study and analyze whether this binary classifier correctly classifies the malicious uniform resource locators into phishy and non- phishy types. The implementation of this classifier is performed over different sizes of data sets to observe variance in prediction accuracy. The prediction accuracy turn out efficient during testing with varying size of datasets of malicious uniform resource locators and do not decreases its performance with an increase in the size of datasets as given in table- 1 and fig -2 .Number of true positives and false positives is also calculated over these datasets to verify these results which are depicted in fig -3 and table-2

***TABLE 1***
*SUMMARY OF RESULTS*

| *size* | *Prediction_Accuracy* |
|---|---|
| *500* | *87.2* |
| *1000* | *88.6* |
| *1500* | *89.8* |
| *2000* | *90.5* |

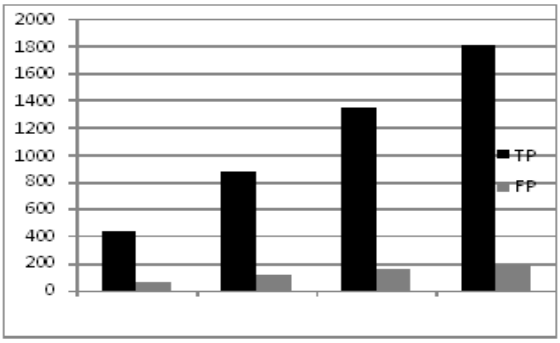


***Fig. 2****: Chart showing the prediction accuracy over different size of Test dataset*

***TABLE 2:*** *CALCULATED VALUES OF TRUE POSITIVE & FALSE*

| Data Set Size | 500 | 1000 | 1500 | 2000 |
|---|---|---|---|---|
| True Positive | 436 | 886 | 1348 | 1810 |
| False Positive | 64 | 114 | 152 | 190 |
| False Positive in Percentage | 12.8% | 11.4% | 10.1% | 9.5% |

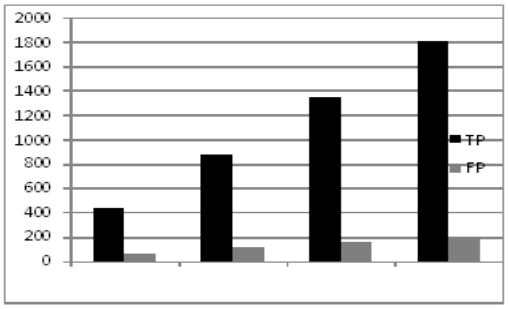


***Fig. 3:*** *Chart showing the number of True positive, False positives over different size of Test datasets.*

These results clearly shows that this binary classifier do not underperform on large size datasets and manages to provide satisfactory prediction accuracy.

## 5. CONCLUSION

The proposed binary classifier is tested over varying sizes of datasets in order to analyze its predictability in the form of accuracy for classifying the uniform resource locators correctly and to provide minimum false positives. In this exploratory process the proposed algorithm has shown satisfactory results over the datasets of malicious uniform resource locators by using only significant attributes.

## 6 REFERENCES


[1] Abdelhamid N and Ayesh A: Phishing detection based associative classification data mining Expert Systems with Applications 41(13) pages 5948- 5959, Oct 2014

[2] Govinda.K, Kevin Thomas (): Survey on Feature Selection and Dimensionality Reduction Techniques International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 03 Issue: 07 | July-2016

[3] Justin Ma Lawrence K. Saul Stefan Savage Geoffrey M.Voelker: identifying Suspicious URLs An Application of Large-Scale Online Learning. Proceeding of 26th International Conference on Machine learning Montreal Canada 2009 .

[4] Dhinakaran, C., Nagamalai, D., Lee, J.-K. (2010). Multilayer approachto defend phishing attacks. ŁaaŁ, 11, 417–425.

[5] K. Hanumantha Rao, : Implementation of Anomaly Detection Technique Using Machine Learning Algorithms. International Journal of Computer Science and Telecommunications [Volume 2, Issue 3, June 2011].

[6] Alkhozae, M.G., Batarfi, O.A., 2011. Phishing websites detection based on phishing characteristics in the webpage source code Int. J. Inf. Commun. Technol.Res. 1.

[7] Le, A., Markopoulou, A., Faloutsos, M., 2011. Phishdef: Url names say it all. In: INFOCOM, 2011 Proceedings IEEE (pp. 191–195). IEEE.

[8] Zhang D, Yan Z, Jiang H, Kim T (2014) :A domain-feature enhanced classification model for the detection of Chinese phishing e- business websites. InfManag 51:845–853

[9] Roopak, S., Thomas, T., 2014. A novel phishing page detection mechanism using html source code comparison and cosine similarity. In: Advances in Computing and Communications (ICACC), 2014 Fourth International Conference on (pp. 167–170). IEEE.

[10] Basnet, R.B., Doleck, T., 2015. Towards developing a tool to detect phishing urls:A machine learning approach. In: Computational Intelligence & Communication Technology (CICT), 2015 IEEE International Conference on (pp. 220–223).IEEE.

[11] Lee, J.-L., Kim, D.-H., Chang-Hoon, Lee, 2015. Heuristic-based approach for phishing site detection using url features. in: Third International Conference On Advances in Computing, Electronics and Electrical Technology - CEET.

[12] Jain, A.K., Gupta, B., 2016. A novel approach to protect against phishing attacks at client side using auto-updated white-list. EURASIP J. Inf. Sec. 2016, 1–11

[13] Mohammad RM, Thabtah F, McCluskey L. Predicting phishing websites based on self-structuring neural network. Neural Computing and Applications 2014; 25(2):443–458.

[14] HTTPS://WWW.PHISHTANK.COM