RESEARCH ARTICLE OPEN ACCESS

Automated Phishing Website Detection Using URL Features and Machine Learning Technique

V.Preethi[1], G.Velmayil[2]
M.Phil Scholar, PG and Research Department of Computer Science,
Quaid-E-Millath Govt. College for Women, Chennai-02[1]
Assistant Professor, PG and Research Department of Computer Science
Quaid-E-Millath Govt. College for Women, Chennai-02[2]

Abstract:

In spite of the development of aversion strategies, phishing remains an essential risk even after the primary countermeasures and in view of receptive URL blacklisting. This strategy is insufficient because of the short lifetime of phishing websites. In order to overcome this problem, developing a real-time phishing website detection method is an effective solution. This research introduces the PrePhish algorithm which is an automated machine learning approach to analyze phishing and non-phishing URL to produce reliable result. It represents that phishing URLs typically have couple of connections between the part of the registered domain level and the path or query level URL. Using these connections URL is characterized by inter-relatedness and it estimates using features mined from attributes. These features are then used in machine learning technique to detect phishing URLs from a real dataset. The classification of phishing and non-phishing website has been implemented by finding the range value and threshold value for each attribute using decision making classification. This method is also evaluated in Matlab using three major classifiers SVM, Random Forest and Naive Bayes to find how it works on the dataset assessed.

Keywords— **Phishing detection, machine learning, URL features, classification algorithm**

I. INTRODUCTION

In this digital day and electronic world, Internet plays a vital role in day-to-day activities like communication, business, transactions, personal needs, marketing, e-commerce etc. Internet is a multifaceted facility which help in completing many tasks readily and conveniently within few seconds. Almost everything is presently accessible over web in this period of progression of advances. Thus increasing usage of internet leads to cybercrime and other malware activities. The information divulged in online leaves digital imprint and if it happens to drop into the wrong hands, it will result in data theft, identity theft and monetary loss. Cybercrime includes many kinds of security issues over the internet and one of the most threatening problems is Phishing. Phishing is a fraudulent technique achieved by phishing web page. Phishing uses e-mails and websites, which are intended to look like from trusted organization, to hoodwink clients into unveiling their own or money related data. The threatening party then use these data for criminal purposes, such as, identity or data theft and extortion. Clients are deceived into revealing their data either by giving touchy data through a web shape or downloading and introducing unfriendly codes, which seek clients' PCs or checking clients' online actions to get data. Luring Internet users by making them click on rogue links that seem trustworthy is an easy task because of widespread credulity and unawareness.

It is important to prevent user's confidential data from unauthorized access. The procedure for the most part includes sending messages that then cause the beneficiary to either visit a deceitful site and enter their data or to visit an authentic site through a phishing intermediary attack or using spoofed website, which then

gathers the details of user leads to several loss. The Phishing problem needs to be mitigated by anti-Phishing approaches. This research provides a solution that helps in detecting and preventing Phishing attacks using the features of phishing URLs and an automated real-time detection of phishing websites by machine learning approach.

II. RELATED WORK

In Phishing E-mail Detection Based on Structural Properties[1], the proposed approach explains to find phishing through appropriate identification and usage of structural properties of email. The experiment is done by SVM and classification technique to classify phishing e-mails. The technique used in this classification method is not large enough and it uses only one approach to identify phishing e-mails, which is low in efficiency and scalability. This is purely based on structural properties of e-mail and it has to extend more structural or content properties to reduce error results.

Discovering Phishing Target Based on Semantic Link Network[2], the paper proposes a novel approach to discover phishing website by calculating association relation among webpages that include malicious webpages and its associated webpages to measure the combination of link relation, search relation, and text relation. The semantic link network proposes a strategy based on four convergent situations to identify the suspicious webpage as phishing. The demerits in this approach are more kind of association has to be done, similarities between visual, layout and domain has to be related. This method is considered as a time consuming approach and also various sub-relations in the combined association relations be studied.

Evolving Fuzzy Neural Network for Phishing Emails Detection[3], deals with zero-day phishing email. It differentiates phishing email and ham email in online mode. It is adopted on feature fetching, rank fetching and grouping similar features of email. The technique is based on binary value 0 or 1 to produce the result for all features used in this method, where 1 denotes a phishing feature and 0 for non-phishing. This technique does not have more dynamic system so it is less in performance to produce accurate results.

Intelligent Phishing Website Detection and Prevention System by Using Link Guard Algorithm[4], proposed a system using link guard algorithm which works for hyperlinks. The algorithm performs certain tests like comparison of the DNS of actual and visual links, checks dotted decimal of IP address, checks encoded links and pattern matching. The drawbacks of this system are, it produce the false positive results if any genuine site has IP address instead of domain name, and it considers some phishing site as normal one if the user does not visit the original site. This results in false negative conclusions.

In Said Afroz, Rachel Greenstadt - Phishzoo Approach[5], the algorithm detects current phishing sites by matching their content with genuine site. This will match images, contents and the structure of website with trusted one in order to avoid phishing. Drawbacks of this algorithm is, it requires matching image site and it is less robust for detecting phishing attacks.

III. THE PROPOSED METHOD

The Proposed PrePhish algorithm is based on an automated real-time phishing detection and a machine learning process. The phishing URLs mostly have couple of connections between the part of the URL which means an inter-relatedness and by using it the features of phishing URLs are extracted. Then the extracted features are used for a machine-learning classification to detect phishing websites on real time.

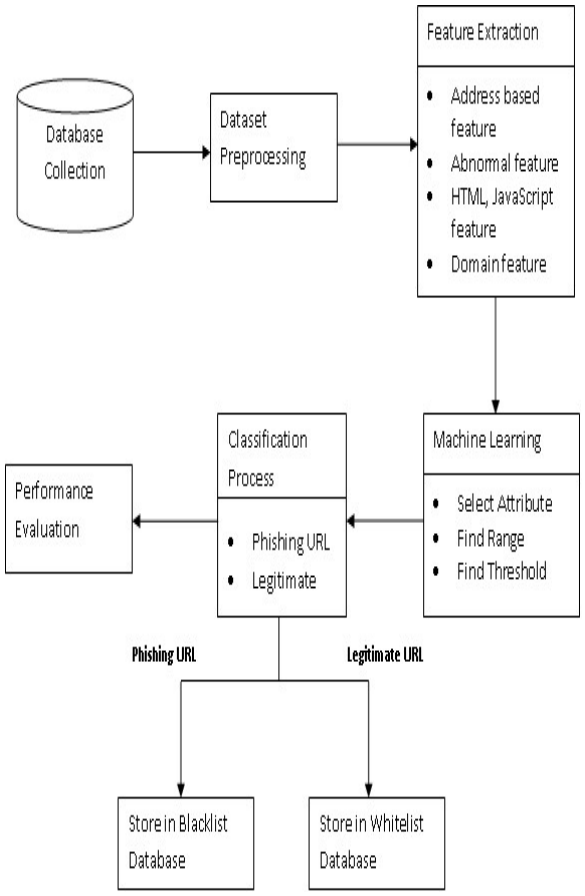

Fig.3.1: Block diagram of PrePhish algorithm

The overview of PrePhish algorithm is shown in Fig.3.1 the dataset of phishing and legitimate URLs are preprocessed for feature extraction method. The preprocessed dataset is used to extract the phishing features for each URL under four categories: Addressed based feature, Abnormal feature, HTML, JavaScript feature and Domain feature. This basic features have number of 30 characteristics of phishing websites which used to differentiate from legitimate website. Each category has its own characteristics of phishing attributes and values are defined. The specified characteristics are extracted for each URL and the valid ranges of inputs are identified. The values are then assigned to each phishing indicator with the range defined for phishing website risk. For each input the values range from 0 to 10 while, for output they range from 0 to 100. The phishing attribute values are represented with binary number 0 and 1 that indicates the attribute is present or not.

TABLE 3.1 ATTRIBUTES AND VALUES FOR PHISHING FEATURE

Feature category	Attributes	Values
Address based Features	having_IP_Address	{ 1,0 }
	URL_Length	{ 1,0,-1 }
	Shortining_Service	{ 0,1 }
	having_At_Symbol	{ 0,1 }
	double_slash_redirecting	{ 1,0 }
	Prefix_Suffix	{ -1,0,1 }
	having_Sub_Domain	{ -1,0,1}
	SSLfinal_State	{ -1,1,0 }
	Domain_registeration_length	{ 0,1,-1}
	Favicon	{ 0,1 }
	Non-standard port	{ 0,1 }
	HTTPS_token	{ 1,0 }
Abnormal Features	Request_URL	{ 1,-1 }
	URL_of_Anchor	{ -1,0,1 }
	Links_in_tags	{ 1,-10 }
	SFH	{ -1,1 }
	Submitting_to_email	{ 1,0 }
	Abnormal_URL	{ 1,0 }
HTML, JavaScript Features	Redirect	{ 0,1 }
	on_mouseover	{ 0,1 }
	RightClick	{ 0,1 }
	popUpWidnow	{ 0,1 }
	Iframe	{ 0,1 }
Domain Features	age_of_domain	{ -1,0,1 }
	DNSRecord	{ 1,0 }
	web_traffic	{ -1,0,1 }
	Page_Rank	{ -1,0,1 }
	Google_Index	{ 0,1 }
	Links_pointing_to_page	{ 1,0,-1 }
	Statistical_report	{ 1,0 }

Table 3.1 represents the feature category, its attribute and values. Some attributes have 3 values which represent its strength ranging from low, medium and high.

IV.IMPLEMENTATION OF PREPHISH METHODOLOGY

The PrePhish methodology which imports dataset of phishing and legitimate URLs from the database and the imported data is preprocessed. Detecting phishing website is performed based on four category of URL features: domain based, address based, abnormal based and HTML, JavaScript

features. These URL features are extracted with processed data and values for each URL attribute are generated. The analysis of URL is performed by machine learning technique which compute range value and the threshold value for URL attributes. Then it is classified into phishing and legitimate URL.

PREPHISH ALGORITHM

1. Import and Preprocess Dataset.
2. Extract the features of URL
3. Compute attribute values, if
   - Attribute present value = 1
   - Attribute absent value = -1
   - Attribute not considered = 0
   - 3.1 Select attribute X and Y
   - 3.2 Compute equation for X and Y
4. Compute threshold value for attribute X and Y
5. Find Range value.
6. Select Attribute to get threshold value.
7. Classify phishing and legitimate site using attribute value.
8. Compute Sensitivity and Specificity.

The attribute values are computed using feature extraction of phishing websites and it is used to identify the range value and threshold value. The values for each phishing attribute is ranging from {-1,0,1} these values are defined as low, medium and high according to phishing website feature. The classification of phishing and legitimate website is based on the values of attributes extracted using four types of phishing categories and a machine learning approach.

A. URL Feature Analysis

The phishing attribute features are extracted for each URL to find whether the website is phishing or legitimate. The URL_of_Anchor tag attribute is selected to find the overlap values which is shown in Fig 4.1. The overlap value is the sum of selected attribute value which is combined with other attributes.

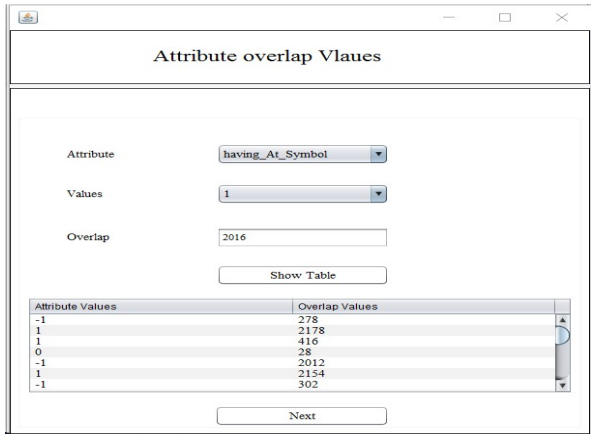

Fig 4.1: URL of Anchor tag

Table 4.1 gives the example for phishing URL based on the phishing feature attributes.

TABLE 4.1 EXAMPLE FOR PHISHING URL FEATURES

URL Features	Example
Having '@' or '//' Symbol	http://harasz.art.pl/images/l/rb=digi/@!_.php http://formulastartup.it//yahoo/index.html
Having long URL	http://nco1925.com/jmhfgh453242sds/amazzon-daazn-amzon-uk-sing-32sdsd-ss12391-htths12-openid-4251-identifier_select=http=fa4udacz-23212ct=checkid-set=aj328aaaa/8ec7ee82ba3e0e2c522f4a3f5ec172c6/
Having Prefix or Suffix	http://bankofamerica-boa.com/
Having IP address	http://59.151.102.220/www.my.commbank.com.au/netbank/
Shortening URL	https://goo.gl/HQx5g

B. Finding Attribute Values

The attribute value for each URL is computed using corresponding set of attribute values {-1,0,1}. Fig 4.2 represents attribute X that URL_of_Anchor tag value and attribute Y that is Prefix_Suffix value. Both the attributes URL_of_Anchor tag and Prefix_Suffix also have inter linked value and that has to be computed for finding range and threshold value.

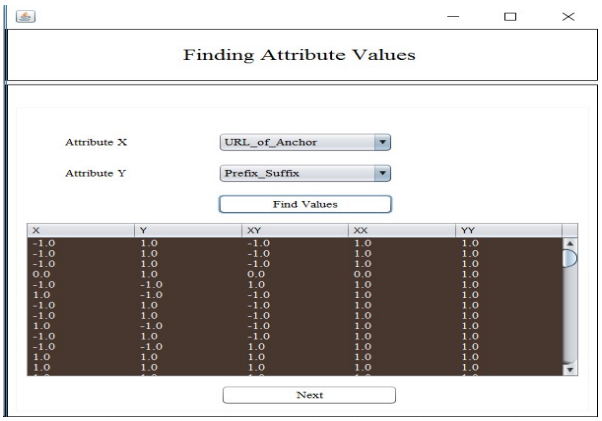

Fig 4.2: Attribute Values for URL of Anchor and Prefix_Suffix

The values of attribute having _At_Symbol and Request_URL is used to find the threshold value computed commonly for both the attribute to find the rate of phishing URLs which are having the selected phishing features. The range and threshold value for attributes URL_of_Anchor and Prefix_Suffix are obtained which is shown in Fig 4.3. The values for each attribute differ from others and thus it has to be computed every time.

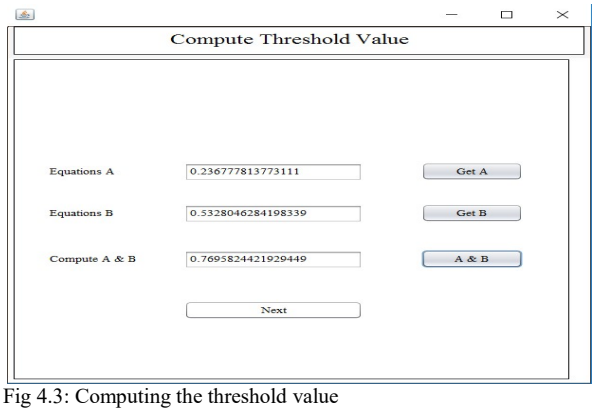

Fig 4.3: Computing the threshold value

The equation to compute A and B value is:

$$A = \frac{(\sum y)(\sum x^2)-(\sum x)(\sum xy)}{n(\sum x^2)-(\sum x)^2} \quad (1)$$

$$B = \frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^2)-(\sum x)^2} \quad (2)$$

$$A + B = \text{Range value} \quad (3)$$

C. URL Classification

The computed threshold value is used to classify the phishing and legitimate ULRs. The positive value 1 for the attribute Prefix_Suffix represent as phishing and the negative value -1 as legitimate. Fig 4.4 classifies the URL based on 1 and -1 values of Prefix_Suffix attribute.

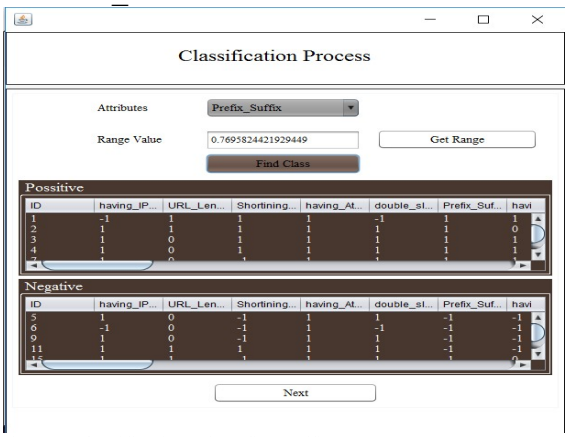

Fig 4.4: Classification of Prefix_Suffix

The true positive/negative and false positive/negative value shown in Fig 4.5 which computed using k-fold cross validation technique by splitting the data into two sets as known and unknown. The known dataset is a training dataset and the unknown dataset is a testing dataset. A general rule to assess the minimum size for a training set is to dimension it six times the number of used feature.

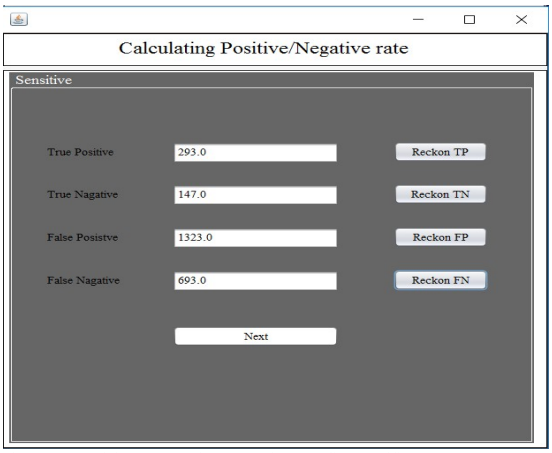

Fig 4.5: Positive-Negative rate for Prefix_Suffix

Formula to find TP/TN and FP/FN :

Phishing classified as phishing: true positives (TP) and
TPrate = TP/TP+FN *(4)*

Legitimate classified as phishing: false positives (FP) and
FPrate = FP/TN+FP *(5)*

Legitimate classified as legitimate: true negatives (TN) and
TNrate = TN/TN+FP *(6)*
Phishing classified as legitimate: false negatives (FN) and
FNrate = FN/TP+FN *(7)*

Sensitivity is also called the true positive rate, which measures the proportion of positives that are correctly identified. Sensitivity is calculated to find the number of phishing websites which are classified correctly as phishing.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (8)$$

Specificity is also called the true negative rate, measures the proportion of negatives that are correctly identified. Specificity is calculated to find the number of legitimate website which are classified correctly as legitimate.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (9)$$

The positive and negative predictive values (PPV and NPV respectively) are the proportions of positive and negative results in statistics tests that are true positive and true negative results, respectively.

Positive predictive value is the probability of phishing website that has classified using phishing properties.

$$\text{Positive Predictive} = \frac{\text{TP}}{\text{TP+FP}} \quad (10)$$

False predictive value is the probability of legitimate website which does not have phishing properties.

$$\text{False Predictive} = \frac{\text{TN}}{\text{TN+FN}} \quad (11)$$

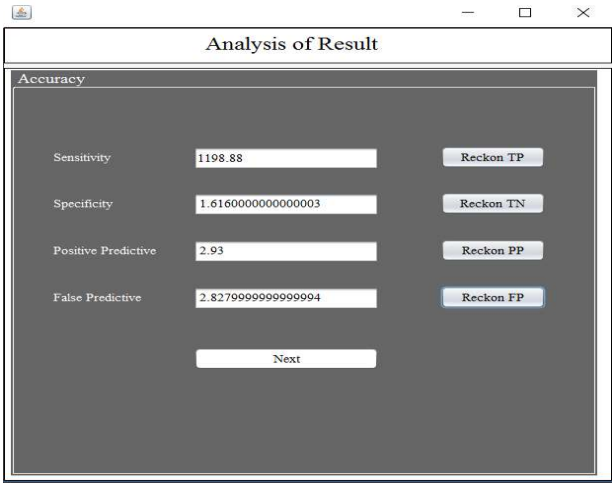

Fig 4.6: Analysis of Prefix_Suffix

The sensitivity, specificity and positive predictive, false predictive values are computed using the following formula and the Fig 4.6 represents computed value for given dataset that is used in this process.

V.RESULT AND DISCUSSION

The classification of phishing and legitimate URL using the categories domain based, address based, abnormal based and HTML, JavaScript features are finally obtained. Table 5.1 shows the rate of positive and negative values for classified URLs. The rate of phishing classified correctly is 97.83% and the rate of phishing incorrectly classified as legitimate is 2.17%. The legitimate URL correctly classified is 98.18% and legitimate incorrectly classified as phishing is 1.82% .

TABLE 5.1: CLASSIFICATION OF URL

Class	Class. Phishing	Class. Legitimate
Phishing	97.83 %	2.17 %
Legitimate	1.82 %	98.18 %

The group level analysis of phishing and legitimate URL is shown in Fig 5.1 and this elucidate the total number of phishing and legitimate websites using the combination of all attributes values. The 30 characteristics of phishing URLs are extracted and analyzed by the classification process.

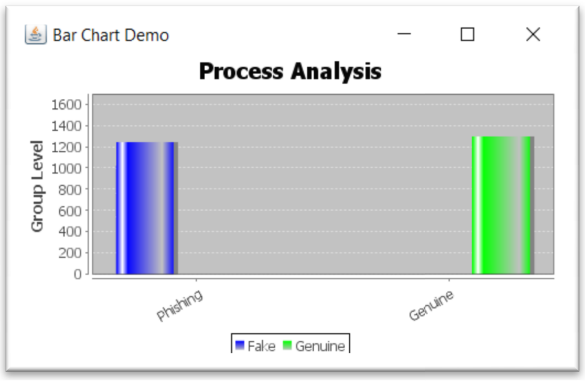

Fig 5.1: Group level analysis

The comparison between proposed PrePhish methodology and existing PhishStorm methodology is shown in Fig 5.2 The existing PhishStorm analyzes URLs using five types of classification that are URL obfuscation with other domain, URL obfuscation with keywords, Typo squatting domains or long domains, URL obfuscation with IP address and Obfuscation with URL shortened which extracts 12 features for each URL. In the proposed PrePhish methodology it uses 30 features based on four categories and that are extracted for each URL to examine.

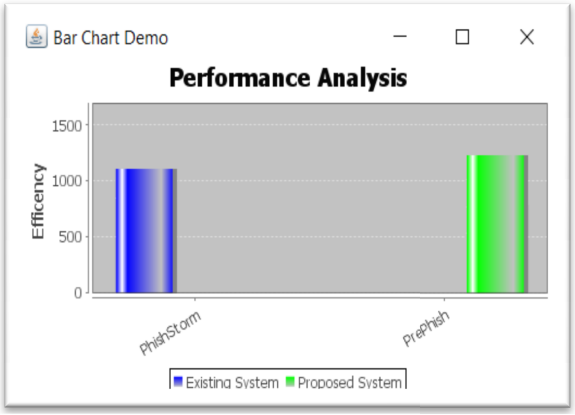

Fig 5.11: Comparison with existing system

TABLE 5.2 CLASSIFICATION RESULT FOR DATASET

Methodology	Class. Phishing	Class. Legitimate
PrePhish	97.83 %	98.18 %
PhishStorm	83.97%	99.22%

The PrePhish methodology classifies 97.83% of phishing URL and 98.18% of legitimate URL when compared to existing PhishStorm method that identified 99.22% legitimate URL and 83.97% of phishing URL that shown in Table 5.2

A. Analysis of Proposed Method in Matlab Using Classifiers

The proposed method is implemented in Matlab and the classification process is done by using three major classifiers that are SVM, Random Forest and Naive Bayes. Each URL is assigned by its corresponding feature attribute and it is used for classification process which are the input to machine learning technique to identify phishing and legitimate URLs. In classification process, a classifier tries to learn several feature variables as inputs to predict an output. In the case of phishing website classification, a classifier rule tries to classify a website as phishing or legitimate by learning defined characteristics, features and patterns in the website. The classification is performed by using three classifiers that are SVM, Random Forest and Naive Bayes. The performance evaluation is shown in the Fig 5.3

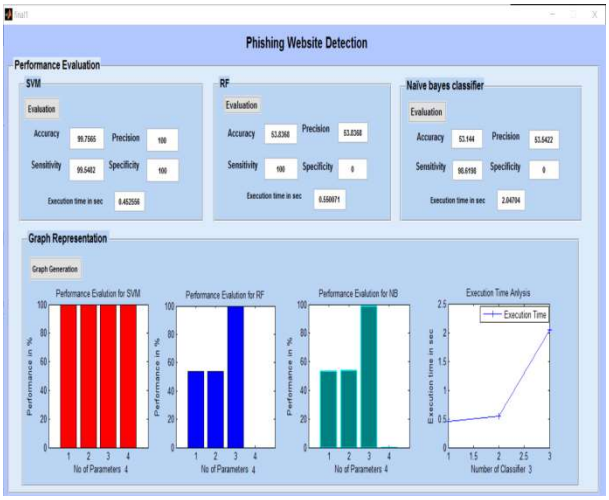

Fig 5.3 Performance Evaluation

Table 5.3 shows the results of classifiers used for the classification process in Matlab. From the table it is shows that the classifier SVM produce maximum result at the minimum rate of time when compared to Random Forest and Naive Bayes.

TABLE 5.3 CLASSIFICATION RESULT FOR DATASET

Classifiers	Accuracy	Sensitivity	Precision	Time Sec.
SVM	99.7565	99.5482	100	0.452556
Random Forest	53.8368	100	53.8368	0.550071
Naive Bayes	53.144	53.5422	98.6198	2.04704

VI. CONCLUSION AND FUTURE WORK

This research proposes the PrePhish algorithm to acquire an efficient phishing URL detection system relying on URL lexical analysis. The PrePhish methodology is an empirical phishing experimental case study that has been implemented to gather and analyze range of different phishing website features and patterns, with all its relations. The proposed method has been implemented on dataset of 2456 phishing and legitimate URLs. The set URLs are analyzed using inter related features and the experiment furnish a classification of phishing and legitimate URL with 97.83% of accuracy and 1.82% of false predictive rate. This is an automated machine learning approach that rely on characteristics of phishing URL properties to detect and prevent phishing websites and to ensure high level security. The classification is done in Matlab using SVM, Random Forest and Naive Bayes classifiers. As a future work the same technique is used to develop a tool, based on a web browser add-on component which can detect and prevent phishing websites on real time in addition to, implementing data mining techniques to discover new patterns of phishing URL.

REFERENCES

1. Chandrasekaran, Madhusudhanan, Krishnan Narayanan, and Shambhu Upadhyaya. "Phishing email detection based on structural properties." NYS Cyber Security Conference. 2006.

2. Wenyin, Liu, et al. "Discovering phishing target based on semantic link network." Future Generation Computer Systems 26.3 (2010): 381-388.

3. Almomani, Ammar, et al. "Evolving fuzzy neural network for phishing emails detection." Journal of Computer Science 8.7 (2012): 1099.

4. Madhuri, M., K. Yeseswini, and U. Vidya Sagar. "Intelligent phishing website detection and prevention system by using link guard algorithm." Int. J. Commun. Netw. Secur 2 (2013): 9-15.

5. Afroz, Sadia, and Rachel Greenstadt. "Phishzoo: Detecting phishing websites by looking at them." Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on. IEEE, 2011.

6. Wenyin, Liu, et al. "Discovering phishing target based on semantic link network." Future Generation Computer Systems 26.3 (2010): 381-388.

7. Bergholz, A., Chang, J. H., Paass, G., Reichartz, F., & Strobel, S. "Improved Phishing Detection using Model-Based Features." CEAS. 2008.

8. Sananse, Bhagyashree E., and Tanuja K. Sarode. "Phishing URL Detection: A Machine Learning and Web

Mining-based Approach." International Journal of Computer Applications 123.13 (2015).

9. Mishra, Madhuresh, Anurag Jain Gaurav, and A. Jain. "A Preventive Anti-Phishing Technique using Code word." International Journal of Computer Science and Information Technologies 3.3 (2012): 4248-4250.

10. Shreeram, V., et al. "Anti-phishing detection of phishing attacks using genetic algorithm." Communication Control and Computing Technologies (ICCCCT), 2010 IEEE International Conference on. IEEE, 2010.

11. Parrish Jr, James L., Janet L. Bailey, and James F. Courtney. "A Personality Based Model for Determining Susceptibility to Phishing Attacks." Little Rock: University of Arkansas (2009).

12. Gupta, Rajendra, and Piyush Kumar Shukla. "Performance Analysis of Anti-Phishing Tools and Study of Classification Data Mining Algorithms for a Novel Anti-Phishing System." International Journal of Computer Network and Information Security (IJCNIS) 7.12 (2015): 70.

13. V. M. Vasava and Rupali A. Mangrule. " Detection and Prevention of Javascript Vulnerability in Social Media." International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE) 2015: 708-713.

14. James, Divya, and Mintu Philip. "A novel anti phishing framework based on visual cryptography." Power, Signals, Controls and Computation (EPSCICON), 2012 International Conference on. IEEE, 2012.