Implementation of Phishing Detection using SVM

Dr. A. Christy Jeba Malar[1], Dr.R.Kanmani[2], Vijayavarman R[3],PraveenKumar R[4], Poorna Bharathi G[5]

[1]currently working as Associate Professor in Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. *Email: a.christyjebamalar@skct.edu.in*

[2]currently working as Associate Professor in Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India. *Email: r.kanmani@skct.edu.in*

[3,4,5]Students, Department of Information Technology, Sri Krishna College of Technology, Coimbatore, India.

Article Info
Volume 83
Page Number: 3287 - 3295
Publication Issue:
March - April 2020

Article History
***Article Received**: 24 July 2019*
***Revised**: 12 September 2019*
***Accepted**: 15 February 2020*
***Publication**: 22 March 2020*

Abstract:
Phishing is a criminal deception of phishers intended to steal sensitive information such as username and password of financial account details in a pool of internet users and proliferate it through e-mails and social media. In this paper, we have proposed a machine learning based anti-phishing system based on lexical features, host properties with some evaluation techniques. The proposed system is trained using more than 10,000 phishing and licit URLs. For accurate and reliable evaluation of our proposed system, we have used different types of SVM Kernel. Experimental results of our proposed system show more than 90% accuracy in detecting phishing site using SVM Kernel.

Keywords :*cyber security, Phishing Detection, Support Vector Machine.*

I INTRODUCTION

The recent years of study on cyber security reveals that there is an escalation of phishing sites. Phishing is a social engineering technique attempted by phishers to steal personal and sensitive information such as credit card pins, user names, passwords of an individuals or group of users. There were many approaches have been implemented like signature, novel, content, anomaly and Ensemble based approaches to restrict the loss of sensitive information of users. Nowadays lifespan of a phishing site is scaled down to be less than 24hrs. As the technology develops So there is an advancement in phishing techniques. Typical a diverse way of attempting phishing attack such as address bar spoofing, cross site scripting, HTML frame injection, key loggers, shorten URLs for redirection to malicious site have been grown. We have taken some of this criterion as an input to train our proposed system.

According to Verizon Data Breach Investigation Report (DBIR) 74% of cyber-espionage actions within the public sector involved phishing. The stealing of personal information of users also occurs through installation of malicious virus into their systems and it launches keylogger that captures everything you type within the login interface. Machine Learning provide accurate way of analysing vast volumes of data. In Machine Learning SVM (support vector machine) is a supervised machine learning algorithm which can be used for both classification and regression challenges. And SVM use a set of mathematical function that are defined as the kernel, it returns the inner product between two different types of classified data.

In SVM there were many kernels used depending on set of input data given to a machine. Every kernel having own functions such as Gaussian, Gaussian Radial Basis Function (RBF), Laplace RBF Kernel are general-purpose kernel and RBF is used when there is no prior knowledge about the data.

1.1 Methods of Phishing

Different number of phishing techniques have been performed to lure the users for obtaining the information from them. Basic techniques are:

In the phishing site, the contents are made with the same phrasing, typefaces, logos and signatures to appear like similar contents in the legitimate site.

Our proposed system generates less false positive or less false negative classifications.

1.2 Statistics of phishing

Phishing is one of the rapidly growing types of identifying a fraudulent or a deceptive act which causes more damage in both short and long terms. In the early stages, there have been 33,000 phishing attacks are recorded globally each month which causes more losses in the year 2012. Financial services are the targeted industry sectors for the phishers. Verizon's data breach investigation report provides approximate percentage of data breaches for every year.

1.3 Neural Networks

Neural network could be a branch of machine learning technique that is employed to acknowledge the identical patterns between legitimate and phishing sites. Having the flexibility to implicitly observe complicated non-linear relationships between dependent and freelance variables and conjointly to observe all doable interactions between predictor variables. The first and spoofing sites are loaded into the neural networks supported the feature extractions like address examination, Host Based mostly Analysis, secure certification checking, website structure checking and page ranking.

II RELATED WORKS

Many researchers have examined the statistics of illegal URL's in different ways. In this we have borrowed more important ideas from previous studies.

Ma et al(2017). performed several machine learning algorithms to classify the phishing URL's and stated that the lexical features and the host-based approaches results in classification with high accuracy. In the study carried out by He et al., the provided a system based on page content, search engine results and HTTP transactions, which could examine spoofing pages with an accuracy of 97% (He, et al., 2011). SVM algorithm also used to classified the datasets for the valid output.

In (Arade,Bhaskar,& Kamet 2011), the author mainly used a different algorithm for approximate string matching in order to similar the web page address and addresses in the database. If the similarity is more than 60%, then the webpage will consider as original or else, the web page content will be verified and the previous algorithm will run for all web page links.

Ramesh et al., proposed a possibilities to phishing web pages by taking web page under indicate and inspect all the direct and indirect links closed with the page(Ramesh, Krishnamoorthy and Kumar, 2014).

III DESIGN FLOW

The work consists of Host, Lexical and Page based features which are extracted from the collected legitimate and phishing site database providers. This database gained knowledge using different machine learning techniques. The Design Flow is shown in Figure 1.

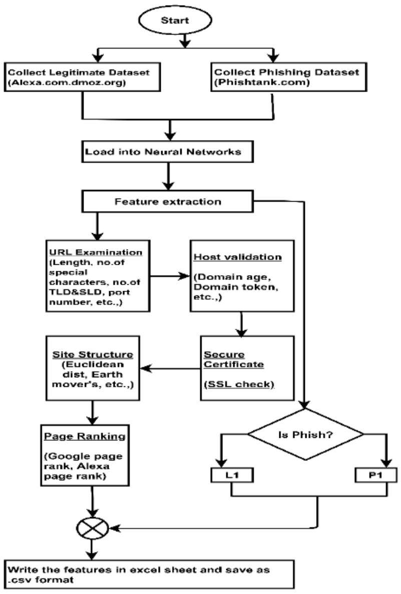

Start
Collect Legitimate Dataset (Alexa.com.dmoz.org)
Collect Phishing Dataset (Phishtank.com)
Load into Neural Networks
Feature extraction
URL Examination (Length, no.of special characters, no.of TLD&SLD, port number, etc.,)
Host validation (Domain age, Domain token, etc.,)
Site Structure (Euclidean dist, Earth mover's, etc.,)
Secure Certificate (SSL check)
Page Ranking (Google page rank, Alexa page rank)
Is Phish?
L1
P1
Write the features in excel sheet and save as .csv format

Figure 1. Design Flow

2.1 Cluster Of URLs

We have collected legitimate URLs from different sites like www.alexa.com and www.dmoz.com. The cluster of phishing sites are taken from www.phishtank.com etc.The dataset consists of 17,000 phishing URLs and 20,000 benign sites. And we have obtained PageRank of 200 legitimate and 200 phishing sites through PR checker. And we have collected WHOIS information of 200legitimate and 200 phishing sites. Examination of a URL,

Syntax Of URL:<protocol>://<hostname><path>

Examination Of URL:

https://www.blog.hotspot.com/marketing/parts

- https – Protocol
- Hostname – www.hotspot.com
- Subdomain - blog
- Secondary Level Domain (SLD) – hotspot
- Top Level Domain (TLD) – com
- Subdirectory - /marketing

As of now research says that a site URL which contains more than two subdomains having a greater chance of being a phishing site. Each domain is separated by dots(.). Some of the criterion which led some sites into phishing sites and also helpful in identifying the phishing sites they are, numbers of @ symbols, number of Dash(-), position of top level domain, HTTPS- (S word for secure check), length of the URLs DNS Lookup is used to retrieve all DNS records of a domain name.

Number of attacks on URLs some of them are, Spamming, Phishing, Malware, DOS attack, SQL injection, Fast flux, URL Manipulation, Semantic URL, Tampering attack, Cross side scripting.

Classification of URLs through types of URLS (original or phishing), Features (lexical, webpage content, link popularity, Host based), Classification Models (SVM, logistic Regression, Naïve Bayes), Datasets (dmoz phish tank, DNS-BH).

The URL properties includes Domain token count, Average domain token length, Average path token length, Longest domain token length, Longest path token length, Brand name presence, Length of hostname, length of entire URL, IP address presence of binary, Security sensitive word presence binary and tokens in the path URL delimited.The design flow shows that the list of URL features are extracted from the legitimate and phishing site which is loaded in the neural networks.

1.4 Host based analysis

1. *Geographic properties:*

The geographical properties give us the country/continent/city for the respective IP address.

2. *WHOIS properties:*

It gives the details about the date of registration, updating and expiry and also ability to know the registrar and the registrant. Several phishing sites contains IP address in their Host name. So getting such a details of a site will led easier path whether the site is legitimate or phishing.

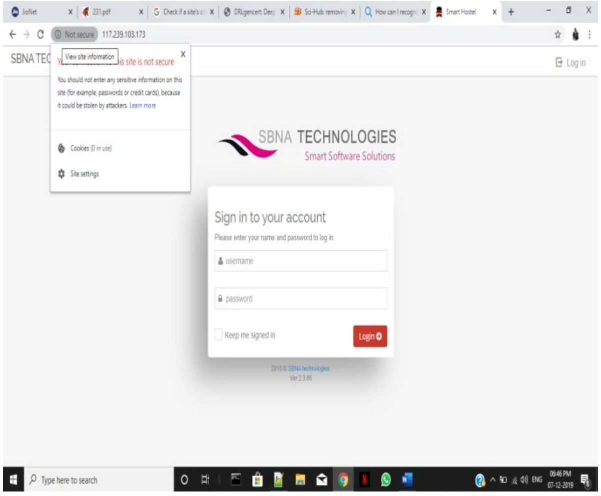
3. *Domain Validation*

The phishing sites have very short life span so it adds more difficulties to detect with accuracy. Domain age willdefine the life span of a site.

4. *Validation of SSL*

The SSL certificate is able to establish encrypted HTTPS connections. A invalid certificate will not receive Ranking Factor. If a site might look as untrustworthy because of self-signed certificate, no known root certificate, expired certificate, accessed domain does not match the valid range of the domain registered in the certificate, a site may consists of unencrypted resources, SNI(server name indication) support is missing in the certificate, the usage of old protocol version due to use of an old version of open SSL, the evolution of SSL are, SSL v2 is unsafe and should not be used, SSL v3 and TLSv1.0 are wide-spread, although their security is being challenged, TLS v1.1 and v1.2 are the newest, safest standards and the key length is always less than 2048 bit. And we have examined almost all the SSL certificates such as Extended Validation Certificates(EV SSL), Organization Validated Certificates(OV SSL), Domain Validated Certificate(DV SSL), Multi-Domain SSL Certificate(MDC), Unified Communications Certificate(UCC) through existing Deep Learning-Based Automated Testing of Certificate verification in SSL/TLS.

It uses Deep Reinforcement Learning network, which enables them to learn independently and automated. And it's a framework consists of three components certificate set, the deep reinforcement learning network, and the differentialtesting module. A non-secured site as shown below,



You should not enter any sensitive information on this site (for example, passwords or credit cards), because it could be stolen by attackers.

Sign in to your account

Please enter your name and password to log in.

☐ Keep me signed in

2018 © SBNA technologies
Ver 2.3.05

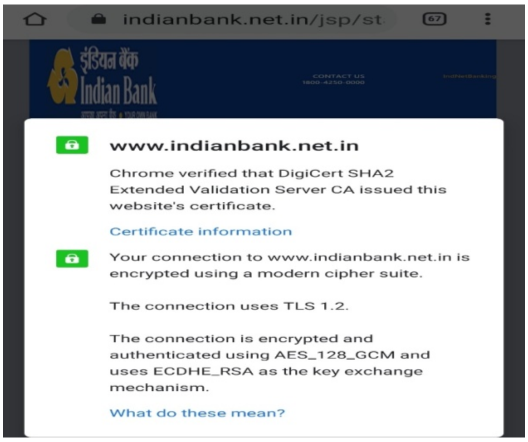
Figure 2. Non-secure site

The Figure2 which represents hostname as URL for a site and it doesn't have any SSL/TLS certificate within it, so it is non secure site to enter any sensitive or personal information of the targeted users because the personal information could be stolen by phishers.

The secure site having the users information as a encrypted form and also hold the certificate authority issued valid SSL certificates. Secured site as shown given below,

5. *Page Ranking*

Page rank is a system developed by google for ranking webpages. Page rank provides absolute importance and authority by checking their quality and quantity of its links. Each page has a link connect to another casts, and weight depends on weight of the pages that link to it. Page rank calculation is made without any prior knowledge about the value of other pages that link to it. We will get a closer estimation of final value when we run the calculation. We have to note each calculated value and keep on repeating the calculations more number of times till the numbers stop changing much.



www.indianbank.net.in

Chrome verified that DigiCert SHA2 Extended Validation Server CA issued this website's certificate.

Your connection to www.indianbank.net.in is encrypted using a modern cipher suite.

The connection uses TLS 1.2.

The connection is encrypted and authenticated using AES_128_GCM and uses ECDHE_RSA as the key exchange mechanism.

Figure 3. Secure Site Certification

In order to discard some pages from having more influence, it uses a term called dampening factor. According to this, dampening factor is defined as the probability that person will continue clicking at any particular step. The total value of pages are sparged down by multiplying with generally assumed value(0.85).

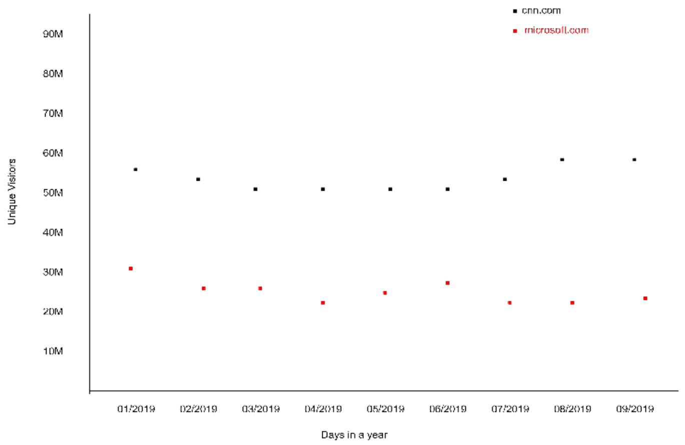
It is noted that average sum of all pages is equal to one. Even if a page doesn't have any votes, it still provided with average score of 0.15.

Google will recalculate score of page rank after each and every trail of web. In this, page rank always provides importance to older pages as new pages cannot have better quality of backlinks than older pages. So ,new pages receive lower values.

$Wi = (1-d) + d \sum li, j(w/n)$

The page rank of a particular page equals to 1 by subtracting and adding dampening factor(d) for every links, add page rank of every pages which is divided by number of links on page and decreased by dampening factor.

The estimation and ranking of Alexa's are based on the people's actions in our total data panel which acts as sample of all internet users. The ranking of site is based on combination of page visitors and page views. The site with best combination of unique visitors and pageviews is ranked as 1.



Alexa Traffic Rank for 2019

6. *Site structure checking*

The organization of website's content is referred as a site structure. It deals with how this content is grouped, presented and linked to the user. The site structure can be defined by some categories like CCH (Contrast Context Histogram), Euclidean distance and the EMD (Earth Mover's Distance). These categories are effectively used to defined how the site structured can be checked properly and also detect the scams.

CCH (Contrast Context Histogram) uses the properties of contrast of a local region, but not store the weighted edge orientation histograms of salient edges. This is the process of gaining the similarities from analog to image, which can be performed by different aspects. The image is running count of the number of pixels found at each intensity value is kept and scanned in a single pass. This is used to construct a suitable histogram.

Euclidean distance is the square root of the sum of squared differences between corresponding elements of the two vectors like x and y. The distance between vectors x and y is defined as follows:

$D (x, y) = \sqrt{} (x - a)^2 + (y - b)^2$

It is a live of similarity ought to show a discrepancy variable underneath the admittible information transformations. Activity of the designed for interval information, like the acquainted consumer coefficient of correlation, mechanically disregards variations in variables that may be attributed to variations in scale. If the consumer recall, all valid interval similar objects, will reworked into the one another by a linear transformation. This is often similar two interval variables square measure standardizing the information or to be making an attempt constants that reworked variable like Mx + b is as similar as doable. This is often what the r-square live of regression does). Similarly, a activity of designable for normal information ought to respond solely to variations within the rank ordering and to not absolute the size of scores.

EMD (Early Mover's Distance) is a method to evaluate the dissimilarities between two multi-dimensional distributions in some feature space where a distance measure between single features, which the base distance is given. It has some bounded variable between the centers of mass of the two signatures when the ground distance is induced by a norm. By using the lower bound in retrieval systems could be reduced the number of EMD computations. The EMD has a special category to define the site structured checking.

It allows the partial matches in every normal way. This is more efficient to image retrieval and in order to deal with occasions and clutter. If it is a true metric and also the ground distance is metric and if the total weights of the two signatures are equal.

This would allows the endowing image space with a metric structure, these are the categories to define the site structured checking procedures.

All the Features are extracted from the cluster of legitimate sites from Alexa.com and demoz.com in addition to phishing Datasets are from Phishtank.com and openphish.com by neural networks. The features are extracted in a sequence manner. Then the neural network contained python program is executed. The program processes the legitimate list and the feature list is obtained and it's denoted as variable L1. List is saved in excel and csv format at a location in a computer as specified in the program. The process is repeated for the phishing Datasets. And it's denoted as P1. The program flow is given below,

Figure 4-Represents the performance analysis phase by analysing the given input URLs with respect to the feature extracted by neural networks in .csv file. The csv file having individual columns and rows to store every feature which were extracted by neural networks.

IV MACHINE LEARNING ALGORITHM

SVM is a class of algorithm which collaborate the principal of statistical theory with kernel mapping and optimisation techniques. SVM kernel is the important factor for kernel-based learning and it is an effective of extracting explicit features. The building blocks of kernel-based learning method is known as kernel function. It is very difficult to fit all the kernels into the SVM and also difficulties in fitting the appropriate values of some kernels. So, we have used some of the required kernels which are effective for our anti-phishing model. It is an effective way to analyse the non-linear data.

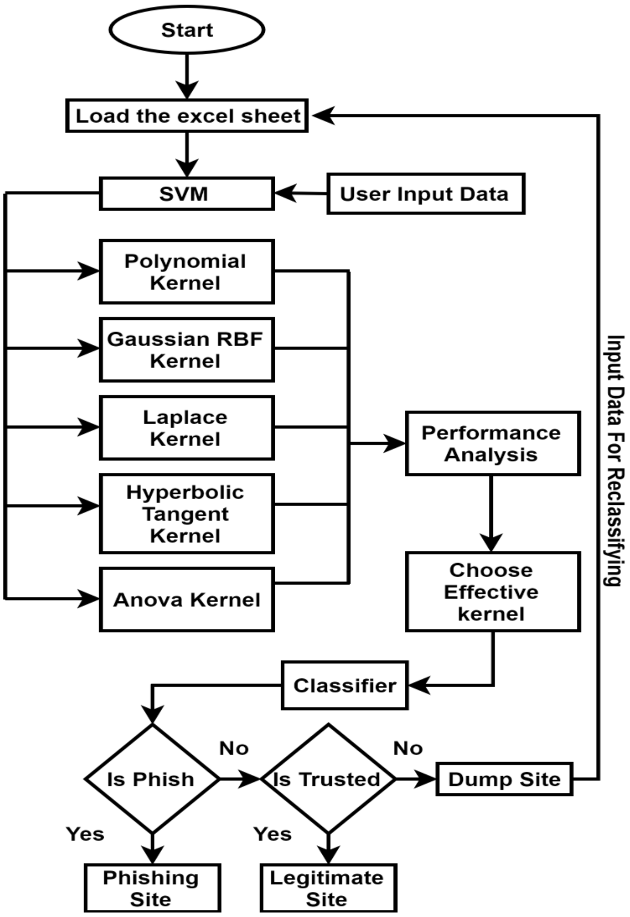


Figure 4. Program Flow

a. Polynomial Kernel

In machine learning, a polynomial kernel is widely used for many other functions but the main function is image processing. In our model we have used polynomial kernel for examination of Site Structure which is defined by CCH (Contrast Context Histogram) which includes the process of identifying and analysing the similarities between the pixels of the picture.

$K\ (x_i,\ y_i) = (x_i.\ y_i{+}r)^{\ d}$

In this equation, x represents the loaded features of legitimate and phishing site in the .csv file format on the excel sheet and y represents input URL by the user for examining phishing or legitimate site and d represents the degree within the analysing factor of x and y. r represents the constant function. The iterations(i&j) are continuous until the features are ended.

b. Gaussian Radial Basis Function(GRBF)

The main function of GRBF is to examine the feature without any prior knowledge about the data. In our model. This is the effective kernel function than any other if any input arrives with more complexity in determining features.

$$K\ (a_i,\ a_j) = \exp\ (-\gamma[a_i\text{-}a_j]^2)$$

Here, 'a_i' represents the continuous iterations on the loaded features of legitimate and phishing sites in the .csv file format on the excel sheet, 'a_j' shows the input URL given by the user. 'γ' denotes optimization parameter for the kernel function. The function of k provides exponentiation of square of a_i &a_j associates with optimization parameter. GRBF is a real valued function and the value depends only on distance between the given input and some framed input.

The Laplace RBF (Radial Basis Function) which also works without any knowledge about the data, the equation is,

$$(x_n,y_i) = \exp(-(x_n\text{-}y_i) / \sigma\ )$$

x_n portray the stored excel sheet input and y_i is the input given by the user, the function k defines exponentiation of the ratio of function parameter to free parameter(σ).

c. Hyperbolic tangent kernel

Hyperbolic Tangent Kernel is also been named as sigmoid kernel or tanh kernel and it is mainly used for neural networks which convert an input vector into some output vector and it is widely used for pattern recognition. In our model we had used it for pattern analysis in SSL(Secure Socket Layer).

$$K(a_i\ ,\ a_j\ ) = \tanh(k\ a_i\ -\ a_j\ +c)$$

a_i denotes the design flow output and a_j is given user input URL. tanh represents the hyperbolic tangent function, k denotes gram matrix of 'a' and c represents the optimization parameter.

d. Anova radial basis kernel

Anova radial basis kernel is used for regression problem to decrease the approximation error while calculating the highest percentage of features extracted from the given input with the help of dataset in the excel sheet.

$$K(x,y) = \sum \exp(-\sigma(x^k\ -\ y^k\ )^2\ )^{\ d}$$

In this equation, 'd' denote degree, x^k and y^k symbolize stored the dataset in .CSV file format on the excel sheet and given input URL by the user for examination. Sigma of exponentiation of free parameter with square of inputs.

e. WORK FLOW

After the formation of .csv file ,excel sheet is inclusion with the given input URL is loaded into the SVM Kernels for analysing the input URL. By broke down the site features and made comparison with .csv file for feature percentage calculation of legitimate and phishing separately named as L1 percentage and P1 percentage. The performance of the kernels is analysed with respect to the percentage. According to the percentage level of features in every kernel, the respective kernel feature has been taken to predict the site is phishing or legitimate by the classifier. If a site URL is said to 100% legitimate then it is intimate or popup the users that the site whom visiting is safe to give the personal credentials within the site and vice verse if it's phishing site it popup it's not a safe to give their personal details. If a site is 50% legitimate and 50% phishing in that case the site is stored in the Dumpsite List where the unpredictable sites are stored, it's function is to reload the respective dump site into the SVM Kernel and process it into the kernel again and again to get the exact result either the site is legitimate or phishing. The storage area which stores the unpredictable sites are called Dump sites. After obtained the exact result of dumpsites the features are extracted it is stored and update the legitimate or phishing site list with respect to the site is original or fake site. And the feature is reused as a reference to solve a new unpredictable site which is similar to stored feature sets. The level of secure site group is shown below,

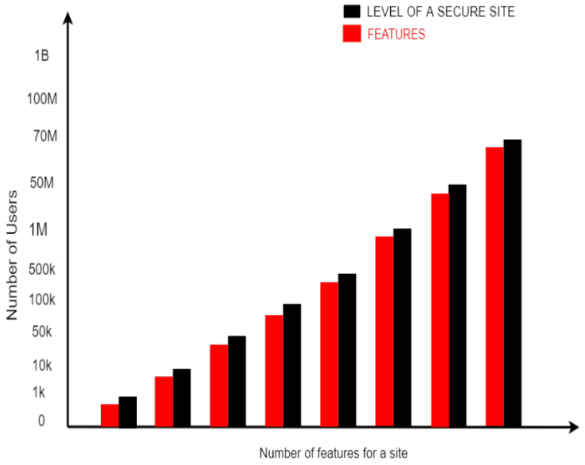


Figure 5. Level Of Secure Site

Level of secure site is mainly comprises of number of features valid within a site. If a site having huge number of views then it must have secure level of features. Security level for a site increases with increases in number of features and number of users.

Secure Site α $\sum$(number of users + number of features)

V CONCLUSION

The main contribution of this work is to provide some features to reduce the possibilities of phishing attack. The approach used in proposed system is to identify the relation between the page content and page address, which based on the aforementioned feature sets. Our experiment indicates that the presented model by using the proposed feature sets along with some related features can detect phishing sites in the internet banking with the accuracy of 95.4% for positive and the 0.84% for negative.

To classify the webpages by using our feature, we use the SVM(Support Vector Machine) algorithm to classify the webpages and also to classify the legitimate site and the spoof site. The SVM kernels has the ability to classify the two vectors and also can detect the spoof variables.We will apply algorithms to cut back the number of options and thereby improve performance. In addition. We will examine a replacement phishing detection technique that uses not solely URL based options but also HTML and JavaScript options of website to enhance performance.

VI REFERENCES

[1] Gastellier-Prevost, S., Granadillo, G. G., and Laurent, M. “Decisive heuristics to differentiate legitimate from phishing sites”, In 2011 Conference on Network and data Systems Security ,pp. 1-9. IEEE., 2011.

[2] Altaher A. “Phishing websites classification using hybrid SVM and KNN approach”, International Journal of Advanced Computer Science and Applications, vol. 8(6), pp. 90-95.(2017).

[3] James J., Sandhya L., Thomas C. “Detection of phishing URLs using machine learning techniques”, In 2013 International Conference on Control Communication and Computing (ICCC). pp. 304-309, IEEE, 2013.

[4] Lee J. L., Kim D. H., Chang-Hoon, L. “Heuristic-based approach for phishing site detection using url features”, In Proc. of the Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology-CEET, pp. 131-135, 2015.

[5] Jain A. K. and Gupta B. B. “Comparative analysis of features based machine learning approaches for phishing detection”, In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 2125-2130, IEEE, 2016.

[6] Afroz S., Greenstadt R. and Phishzoo: “Detecting phishing websites by looking at them”, In 2011 IEEE Fifth International Conference on Semantic Computing , pp. 368-375, IEEE, 2011.

[7] Sharifi M., Siadati S. H, “A phishing sites blacklist generator”, In 2008 IEEE/ACS International Conference on Computer Systems and Applications, pp. 840-843. IEEE, 2008.

[8] Wenyin L., Huang, G., Xiaoyue, L., Deng, X., and Min, Z. “Phishing Web page detection”, In Eighth International Conference onDocument Analysis and Recognition (ICDAR'05), pp. 560-564, IEEE, 2005.

[9] Chu W., Zhu B. B Xue F., Guan X., and Cai Z. “Protect sensitive sites from phishing attacks using features extractable from inaccessible phishing URLs”, In 2013 IEEE International Conference on Communications (ICC), pp. 1990-1994, IEEE. 2013.

[10] Weider D. Y., Nargundkar S., and Tiruthani, N. “A phishing vulnerability analysis of web based systems”, In 2008 IEEE Symposium on Computers and Communications, pp. 326-331, IEEE, 2008.

[11] Khonji M., Jones A., and Iraqi, Y. “A novel Phishing classification based on URL features”, In 2011 IEEE GCC conference and exhibition (GCC) (pp. 221-224). IEEE. 2011.

[12] http://www.antiphishing.org/

[13] http://www.phishtank.com

[14] Moghim, M., and Varjani, A. Y. “New rule-based phishing detection method”, Expert systems with applications, vol. 53, pp. 231-242, 2016.

[15] Fu A. Y. “Web identity security: advanced phishing attacks and counter measures. unpublished doctoral dissertation, City University of Hong Kong, Hong Kong, 2006.

[16] Bhuvaneswari, K., Rauf, H. A. “Edgelet based human detection and tracking by combined segmentation and soft decision” In 2009 International Conference on Control, Automation, Communication and Energy Conservation. pp. 1-6, IEEE. 2009.

[17] Verification, U. F. F. (2011). Efficient multimodal biometric authentication using fast fingerprint verification and enhanced iris features. Journal of Computer Science, 7(5), 698-706.

[18] Punithavathani, D. S., & Sankaranarayanan, K. (2009). IPv4/IPv6 transition mechanisms. *European Journal of Scientific Research*, *34*(1), 110-124.

[19] Melnichuk, M., 2018. Psychosocial Adaptation of International Students: Advanced Screening. International Journal of Psychosocial Rehabilitation. Vol 22 (1) 101, 113.

[20] Daly, A., Arnavut, F., Bohorun, D., Daly, A., Arnavut, F. and Bohorun, D., The Step-Down Challenge. International Journal of Psychosocial Rehabilitation, Vol 22(1) 76, 83.

[21] Knapen, J., Myszta, A. and Moriën, Y., 2018. Augmented individual placement and support for people with serious mental illness: the results of a pilot study in Belgium. International Journal of Psychosocial Rehabilitation, Vol 22(2), pp.11-21.