www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 5 Issue 09 September 2016 Page No.18197-18204

Phishing Detection in Websites Using Neural Networks and Firefly

Swetha Babu K.P[1], Dr.Radha Damodaram[2]

[1]Cms College Of Science And Commerce ,Bharathiyar University, Mphil Scholar, Coimbatore,Tamilnadu,India
swethababukp@gmail.com
[2]Cms College Of Science And Commerce ,Bharathiyar Universit, Associate Professor, Coimbatore,Tamilnadu,India
Radhabalaji10@gmail.com

Abstract: Nowadays phishing become popular in the internet. Phishing is a website forgery where the attackers steal sensitive information of users like username, password, bank details and security details without the knowledge of users. Phishers are the one to create website same as the trusted website with the same content and designs of the trusted website. Phishing can be done through email, websites and malicious software to get intellectual information, business secrets or military information etc. In order to prevent user from phishing websites PhishShield application is used. It detects phishing website with replacing content by images based on heuristic solutions. In this application an URL is given as input and it gives the status of URL whether it is legitimate or unknown or phishing websites. In this few features are used to detect phishing websites but in the proposed system we considered more features including Google PageRank, Google Position, Alexa rank and other URL based features and its accuracy and performance can be improved by using neural networks where optimum weight is calculated based on firefly algorithm. The experimental results are conducted to prove that the proposed technique works more effectively than the existing technique in terms of accuracy, true positive rate, true negative rate, false positive rate and false negative rate.

Keywords: Phishing, neural networks, firefly, PhishShield, firefly.

1. INTRODUCTION

Internet has become an essential thing in our day today life where the users can exchange their more sensitive information like username, password, credit card pin number, bank account details etc. The attackers steal the user sensitive information for their personal use or some organizational use. Phishing is one type of attack happen in the internet. In the phishing websites attackers creating more appropriate website as legitimate website by replacing some content of website or image in the legitimate website. The user must be aware of phishing websites to protect their sensitive information from the phishers. But it is more complex task to predict and detect the phishing websites in the internet. Moreover phishing may result financial loss for an organization. Hence it is more important to detect and prevent the user from the phishing websites. There are various types of phishing techniques such as social engineering and technical subterfuge [1].

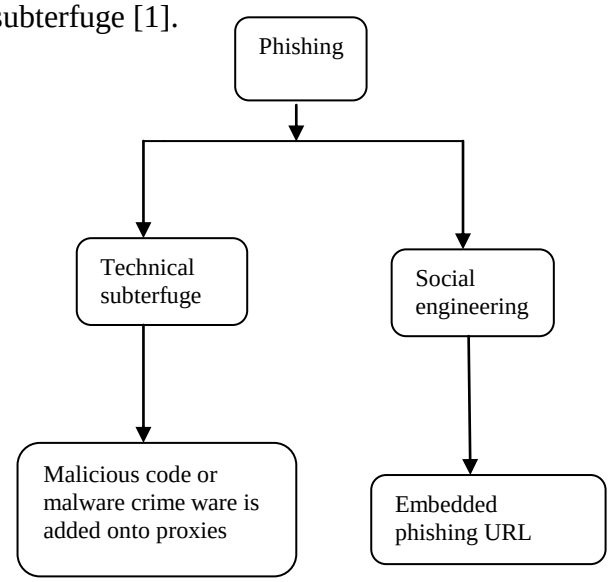

Figure1. Phishing Types

An existing application heuristic approach [2] is used where extracts the features of phishing websites to detect phishing websites. It detects the normal phishing website along with that replacing content with images in a website called image phishing site can also detected by this application. The website URL is given as input to the application and it found the website is legitimate website or not. In this legitimate website and phishing website can be found. It can be enhanced by including more number of features such as Google PageRank, Google Position, Alexa rank, Similarity of primary domain and Google Suggestion, Similarity of sub domain and Google Suggestion, Safety of Google Suggestion , Length of URL, Suspicious character, Prefix and suffix, Number of sub domain and an efficient classifier called neural network is used to classify the legitimate website from the phishing websites. The optimum weight is calculated based on the firefly algorithm it enhances the performance in detecting the phishing websites.

2 LITERATURE SURVEY

Rami M. Mohammad et al [3] proposed new model for the prediction of phishing websites based on artificial neural network model. It identifies phishing attacks in websites through the neural structure of network. The neural network is better than the other classifiers to classify the websites due to several reasons such as identical designing steps, Generalization, Nonlinearity, Fault tolerance and adaptive are used to automating the process of structuring a neural network scheme with the help of some user defined parameters. Thus the proposed model provides adaptive model for learning rate and network structure.

Abdelhamid Neda et al [4] developed an associative classification method called Multi-label Classifier based Associative Classification (MCAC) through describing the problem involved in phishing websites. It also determines the features that able to differentiate the phishing website from the legitimate website. The association classification used to found the correlation coefficient among the discovered features and created an effective rule based on the correlation coefficient. The websites were classified based on the rules created by neural networks. These rules improved the accuracy of classification.

He Mingxing et al [5] proposed a new method to detect the phishing websites based on some heuristic method. It makes use of websites parameters such as search engine results, website content and HTTP transaction. This method was comprised with blacklist-based method that determines certain identity of websites and behavior of website will depends upon the certain identity. The phishing website has fake identity and its behavior depends on their fake identity. From the different identities phishing websites can be easily identified without the prior knowledge of users.

Prakash Pawan et al [6] proposed a new system called PhishNet to observe the attackers. It consists of two components are URL prediction components and approximate URL matching component. In URL prediction components used various heuristics to create new URL based on observing current blacklists then another component found the semantic and syntactic variation of new URL and existing blacklist with the help of hash maps and regular expression to protect against phishing websites. It differentiates the phishing website from the legitimate website with the help of two components.

Mao Jian et al [7] proposed a new solution called BaitAlarm detects the phishing websites based on the features. In this solution, it observes visual appearance features of web pages such page layout and content and detect the similarities in Cascading Style Sheets. It compares the similarity measure between the phishing website and legitimate website from the similarity measure it can detect the phishing websites.

Ramesh Gowtham [8] proposed an anti phishing technique that take web pages for analysis in order

to determine the phishing websites. Form domain group sets from the collected web pages that determine the direct and indirect link involved in the collected web pages. Then determine the target domain set from the domain group sets and it will feed as input to Target Identification algorithm to found the phishing target domain. This approach doesn't require training data to identify the phishing web pages.

Barraclough P.A. [9] developed an online toolbar to address the phishing attack in online transaction. This online toolbar processed concurrently as a background of Internet Explorer web browser that it analysis the websites requested by servers with a set of data in real time. It used six set of inputs to found legitimate websites are phishTank, legitimate site rules, pop-up windows, user behavior profile, user credential profile and pop-up windows are extracted from 300 features by using an extractor algorithm. Then it compares the requested websites with the extracted features. Finally it returns linguistic value which intimates the phishing websites.

Chang Ee Hung et al [10] proposed a new anti phishing method based on the content based image retrieval mechanism to detect the phishing websites and alert the users. This method will provide the screenshot of the web pages. Every web page contains certain logo the screenshot of web pages are segmented region of interest. Then determine the website identity by analyzing Google image database it utilized the content based image retrieval mechanism in the database that retrieves the correct web site identity with the website logo. This content based mechanism effectively differentiates the phishing website from the legitimate website.

3. IMPROVING DETECTION OF PHISHING WEBSITES THOUGH NEURAL NETWORK AND FIREFLY

Phishing is a curious one in websites and it is tedious process to detect. In this proposed work first include the more number of features about the website it increases the accuracy rate and reduce the error rate of prediction.

Feature selection process:

The features play an important role during classification of websites. Based upon the features in the websites identify the phishing website in the internet. Whereas in the internet contains huge volume of websites and it increases dramatically from one second to next second. So for the detection phishing websites more number of features is needed. In this the features Google PageRank, Google Position, Alexa rank, Similarity of primary domain and Google Suggestion, Similarity of sub domain and Google Suggestion, Safety of Google Suggestion, Length of URL, Suspicious character, Prefix and suffix and Number of sub domain are included.

A) Google PageRank:

Page rank is a feature is used to calculate web pages importance based on the vote casted for the corresponding web page. It is a numerical value ranges from 0 to 10 that define importance of web page involved on the web. The votes for each page are increased when another page linked to that page. More number of votes for a particular web page concluded that the page has more importance. PageRank is calculated by using following formula:

$$PageRank\,(W) = (1-f) + f\left(\frac{PageRank\,(T_1)}{Outbound\,(T_1)} + \frac{PageRank\,(T_2)}{Outbound\,(T_2)} + \cdots + \frac{PageRank\,(T_n)}{Outbound\,(T_n)}\right)$$

Eqation (1)

With the help of equation 1, page rank of each web page is calculated. where $PageRank\,(W)$ is the Page rank of page w, $PageRank\,(T_i)$ is the page rank of pages $T_i$which link to page A, f is the damming factor between 0 to 1 and $Outbound(T_i)$ number of outbound links on page T_i.

B) Google Position:

The Google position is the position of the web page in Google search engine. It is defined based on the Google PageRank with numerical value ranges from 0 to 300.

C) Alexa Rank:

Alexa Rank is very much similar to PageRank feature but it fully based on number of users view the page, page views and historical data of about three months. This feature is used to tracking traffic to web sites. It is a numerical value ranges from 0 to millions. 0 is the lowest possible score of a website and million is the highest score of website.

D) Similarity of primary domain and Google Suggestion & Similarity of sub domain and Google Suggestion:

It is an URL based features it is calculated by using Levenshtein distance which is string metric which measures the string difference between the primary domain and sub domain. It is saved in a particular variable used to detect the phishing websites. It can be calculated by using following formula:

Let x and y be two strings then Levenshtein distance is given by

$$l_{x,y}(c1,c2) = \begin{cases} maximum(c1,c2) \ \ if\ minimum(c1,c2) is\ 0 \\ minimum \begin{cases} l_{x,y}(c1-1,c2)+1 \\ l_{x,y}(c1,c2-1)+1 \\ l_{x,y}(c1-1,c2-1)+1 \end{cases} \quad else \end{cases}$$

Equation (2)

where $l_{x,y}(c1,c2)$ is the distance between first c1 characters of x and the first c2 characters of y.

E) Safety of Google suggestion:

In this feature it contains either yes or no value. If the search results of URL is present in the whitelist it hold yes value otherwise it contain no value. It can be for both primary domain and sub domain of a web page.

F) Length of URL, Suspicious character, Prefix and suffix, Number of sub domain:

In the length of the input URL feature defines the number of characters involved in the website URL. It may be up to 2083 characters in an URL. Then in the Suspicious character feature it count the number of suspicious characters in an URL such as @, //, ;. The prefix and suffix feature holds the value of number of prefix and suffix values in an URL. In the number of sub domain feature contains the number of sub domain for a web site. Thus the more number of efficient features are added to define the phishing and legitimate website in internet.

Neural networks and firefly:

Artificial neural network is also called as neural network. The features are given as input to neural network and it classified the websites based on the features. The ANN is computational model that changes its structure based on both external and internal information that flows in a network. The weighting factors in the neural network are adjusted by firefly algorithm it enhances the performance of neural network classification.

G) Neural network:

Artificial Neural Networks (ANN) consists of systems that are purposefully built to make use of some group of values similar to those of the human brain. They signify the promising new creation of information processing systems. Neural Networks are good at tasks such as data clustering, pattern matching and classification, and optimization. They have a large number of extremely interconnected processing elements called neurons, which generally function in parallel and are structured in regular architectures. In this neural network is used for classification process. In this each and every unit of neural network performs two computations based on two functions are input function and activation function. Initially the input function $linear_i$utilized to calculate weighted sum of the unit's input values. Then in another function called activation function converts the output of input function into final values that provides the unit's activation value v_i. The sum of the input activations times their respective weights gives total weighted input.

$$linear_j = \sum_x w_{x,j}\ v_x = w_j\, v_x \qquad (3)$$

Initially the neural network is initialized with random weights from the range of -0.5 to 0.5. Then the weights of each network are updated in this

way to decrease the value of observed values and predicted values. In this weight of each network is updated with the help of an efficient algorithm called firefly it creates an optimum weight for classification.

H) Firefly:

The firefly algorithm chooses the optimum weight for neural network based on the concept of firefly characteristics by using objective function. A firefly is attracted to another firefly despite the consequences of its sex, attractiveness is proportional to their brightness and brightness of each firefly is decided by the landscape of the objective function. Initially the population of fireflies is assigned. Then considered two significant points are formulation of the attractiveness and variation in light intensity. The attractiveness is determined by the brightness of the fireflies in which the objective function is associated. Also, the objective function is used to determine the brightness (I) of the firefly in a specific position. The objective function is to decrease the false error rate (FER) to determine the phishing websites. It can be explained by algorithm 1.

Algorithm 1:

Input: weight values

Output: optimal weight values

1. Begin
2. Create the initial solution at random
3. Formulate light intensity L so that it is associated with $I(y)\ i.e.\ L = I(y)$
4. While(Stopping criterion satisfied)
5. For m=1 to n do
6. For n=1 to n do
7. Find $L(x)$ using equation

   L (x) =minimum (FER)
8. $If(L_m > L_n)$
9. Calculate attractive fireflies

   $\beta(r) = \beta_0 . ew^{-\gamma . r^2}$
10. Compute the distance between the fireflies

$$r_{mn} = ||Y_m - Y_n|| = \sqrt{\sum_{i=1}^{d} (Y_{m,i} - Y_{n,i})^2}$$

11. Move all firefly to the best solution

$$y_m = y_m + \beta_0 * \exp(-\gamma r_{mn}^2) * (y_n - y_m) + \alpha * (rand - \frac{1}{2})$$

12. End if
13. End for
14. End for
15. Move best solution at random
16. Find the best solution from the new population
17. End while

In algorithm 1, the initial population for optimal weight calculation is initialised. Then calculate the intensity and attractiveness function. The attractive value of the firefly at r=0 is represented by β_0 and γ defines the media light absorption coefficient. The distance between two fireflies m, n at the position Y_m and Y_n is calculated and represented by r_{mn}. Move all fireflies to the best solution based on the movement of fireflies $Y_{m,i}$ denotes the i^{th}element of the spatial coordinates Y_m in the m^{th} firefly and d refers the number of dimensions. Thus the optimal weight is calculated it is given to neural network it classify the websites based upon weight.

The classification error is determined by using following equation

Error= P-C (4)

Where P is the predicted output and C is correct output. If the output of this equation is positive then the value of C has to be decreased and vice versa. Each input unit donates $w_x v_j$ to the total input, hence if v_j is positive an increase in w_x will tend to increase of C and if v_j is negative a decrease in w_x will tend to decrease of C. It can be achieved by using following rule:

DOI: 10.18535/ijecs/v5i9.68

$$w_x \leftarrow w_x + v \times v_x \times Error \qquad (5)$$

Thus the classification accuracy of websites is determined by the error equation. In equation (5) v is the learning rate. Using the features in the website rules is generated to determine the phishing website with the help of neural networks and firefly. Once the neural network is created it trained with the existing data and then initially random weights are initialized then adjusted the weights based on the firefly algorithm consequently so that whenever the testing data sees something looking like the existing data it outputs the same result as that data.

4) RESULTS AND DISCUSSION

PhishTank is an anti phishing website where we collect 1800 phishing websites for our experimental purpose. In the PhishTank anyone can submit, verify or tracking phishing data. To estimate the performances of our proposed technique collect 1800 phishing invalid, online, valid and offline phishing sites URL. Moreover we collect 300 legitimate websites out of which 214 are taken from PhishTank and the remaining is taken randomly.

A) True positive

This is a measure used to find the rate of phishing websites correctly classified as phishing websites. It can be calculated by

$$\text{True positive rate} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

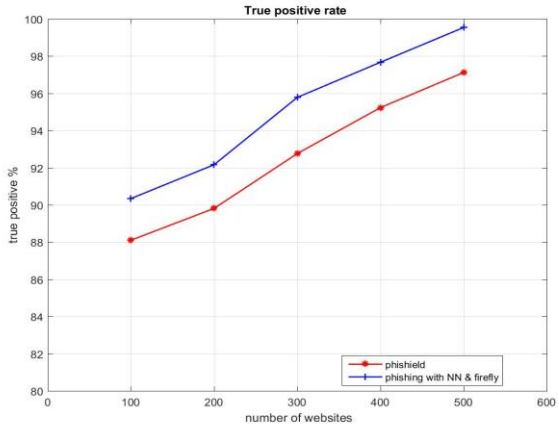

Figure 2: True Positive Comparison Between Phishield And Phishing With NN and Firefly.

In figure 2, x axis represents the number of websites and y axis represents the true positive rate. It is proved that the proposed technique shows 2.42 higher true positive rate at 600 websites than the existing method.

B) True negative

This is a measure used to find the rate of legitimate websites is correctly classified as legitimate websites. It can be calculated by

$$\text{True negative rate} = \frac{\text{True negative}}{\text{True negative} + \text{False positive}}$$

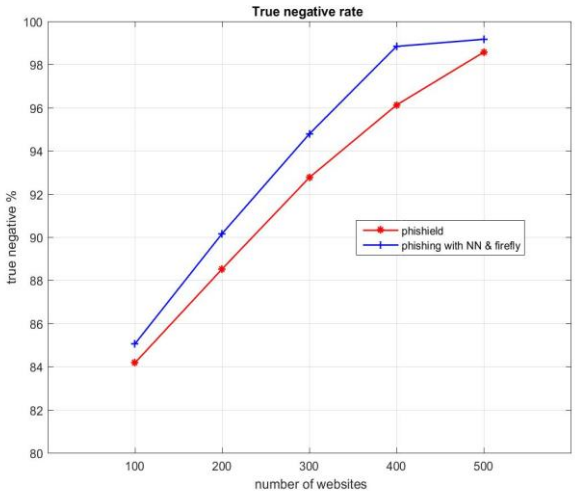

Figure 3: True negative comparison between phishield and phishing with NN and firefly

In figure 3, x axis represents the number of websites and y axis represents the true negative rate. It is proved that the proposed technique shows 0.6 % higher true negative rate at 600 websites than the existing method.

C) False positive

This is a measure used to find the rate of legitimate websites is wrongly classified as phishing websites. It can be calculated by

$$\text{False positive rate} = \frac{\text{False positive}}{\text{False positive} + \text{True negative}}$$

DOI: 10.18535/ijecs/v5i9.68

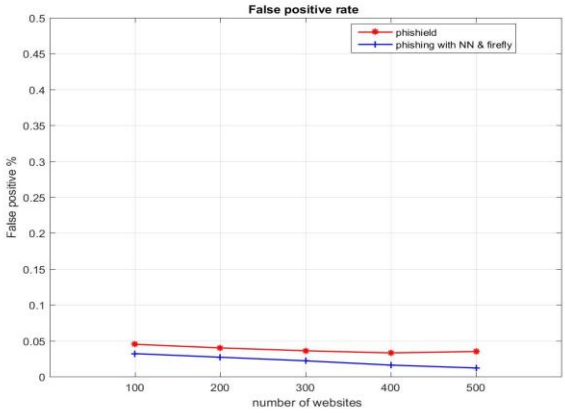

Figure 4: False Positive Comparison Between Phishield And Phishing With NN And Firefly.

In figure 4, x axis represents the number of websites and y axis represents the true negative rate. It is proved that the proposed technique shows 0.023 % lesser true negative rate at 600 websites than the existing method.

D) False negative

This is a measure used to find the rate of phishing websites is wrongly classified as legitimate websites. It can be calculated by

$$\text{False negative rate} = \frac{\text{False negative}}{\text{False negative} + \text{True positive}}$$

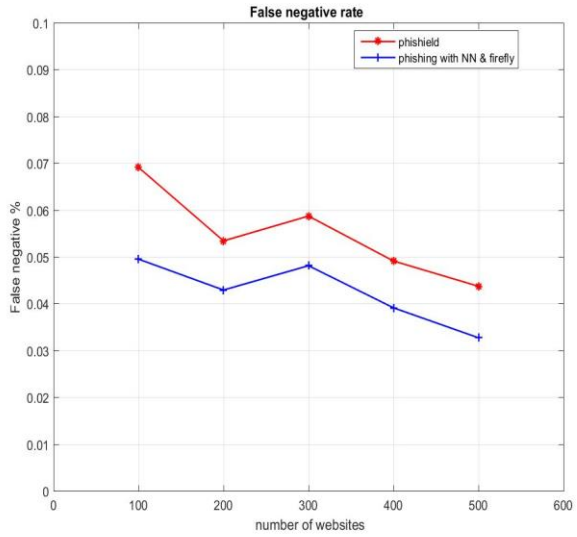

Figure 5 : False Negative Comparison Between Phishield And Phishing With NN and firefly

In figure 5, x axis represents the number of websites and y axis represents the false negative rate. It is proved that the proposed technique shows 0.011 % lesser false negative rate at 600 websites than the existing method

E)Accuracy

Accuracy is the measure of correctly detected phishing websites and legitimate websites in all instances. It can be calculated by

$$Accuracy = \frac{(True\ positive + True\ negative)}{(True\ positive + True\ negative + False\ positive + False\ negative)}$$

In figure 6, x axis represents number of websites and y axis represents accuracy. it is proved that the proposed technique shows 99.52 accuracy and the existing technique shows 96.57 accuracy which is 2.95% higher than the existing technique.

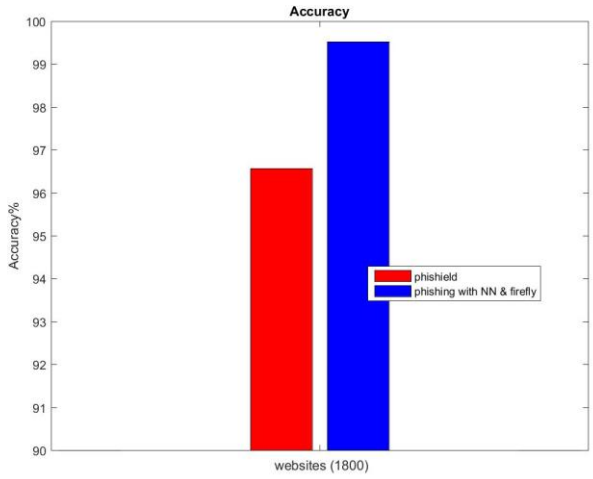

Figure 6: Accuracy Comparison Between Phishield And Phishing with NN And Firefly

5) CONCLUSION

This paper presents an efficient technique to find the phishing website from the huge volume of websites in the internet. This process carried over in three step are feature selection, neural network and firefly. In thefeature selection more number of features related to websites is added to identify the phishing website. Then the features are given as input to neural network to classify the websites. It can be enhanced by using firefly algorithm to find

the optimal weight for classification. Thus the phishing websites are easily detected by this proposed technique. The experimental results show that the proposed technique works more effectively than the existing technique.

6) REFERENCES

[1] Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2013). A survey of phishing email filtering techniques. *IEEE communications surveys & tutorials*, *15*(4), 2070-2090.

[2] Rao, R. S., & Ali, S. T. (2015). PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach. *Procedia Computer Science*, *54*, 147-156.

[3] Mohammad, R. M., Thabtah, F., & McCluskey, L. (2014). Predicting phishing websites based on self-structuring neural network. *Neural Computing and Applications*, *25*(2), 443-458.

[4] Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). Phishing detection based associative classification data mining. *Expert Systems with Applications*,*41*(13), 5948-5959.

[5] He, M., Horng, S. J., Fan, P., Khan, M. K., Run, R. S., Lai, J. L., ... & Sutanto, A. (2011). An efficient phishing webpage detector. *Expert Systems with Applications*, *38*(10), 12018-12027.

[6] Prakash, P., Kumar, M., Kompella, R. R., & Gupta, M. (2010, March). Phishnet: predictive blacklisting to detect phishing attacks. In *INFOCOM, 2010 Proceedings IEEE* (pp. 1-5). IEEE.

[7] Mao, J., Li, P., Li, K., Wei, T., & Liang, Z. (2013, September). BaitAlarm: Detecting phishing sites using similarity in fundamental visual features. In*Intelligent Networking and Collaborative Systems (INCoS), 2013 5th International Conference On* (pp. 790-795). IEEE.

[8] Ramesh, G., Krishnamurthi, I., & Kumar, K. S. S. (2014). An efficacious method for detecting phishing webpages through target domain identification.*Decision Support Systems*, *61*, 12-22.

[9] Barraclough, P. A., Sexton, G., & Aslam, N. (2015, July). Online phishing detection toolbar for transactions. In *Science and Information Conference (SAI), 2015* (pp. 1321-1328). IEEE.

[10] Chang, E. H., Chiew, K. L., & Tiong, W. K. (2013, December). Phishing detection via identification of website identity. In *IT Convergence and Security (ICITCS), 2013 International Conference on* (pp. 1-4). IEEE.