IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 7, Issue 02, September 2018)

Phishing URL Detection using Cuckoo Search Optimized Neural Network Classifier

Varsha Patel
M. Tech. Scholar
Department of Computer Science
Vindhya Institute of Technology & Science, Indore
(M. P.), India
varshapatal@gmail.com

Ashish Tiwari
Asst. Prof. & HOD
Department of CSE & IT
Vindhya Institute of Technology & Science, Indore
(M. P.), India
ashishtiwari205@gmail.com

Abstract –**This paper aims to collect, map and model elements that will lead to the finding of phishing URL automatically, for this purpose data mining is used as basic tools, in this sense, it is considered that the existing patterns in a URL make it possible to distinguish the legitimate link for pages, the identification of these patterns will serve to model a successful classification method, for this purpose, the attributes found in the database "phishing web" that correspond to patterns of phishing pages will be validated, at the same time will be evaluated algorithms extracted from the literature that allow a better classification of records, finally, a model with the highest precision results is delivered which consists of cuckoo search optimized neural network classifier.**

Keywords – **Cuckoo Search Optimization, Neural Network, Phishing, URL.**

I. INTRODUCTION

Phishing fraud - that is, the theft of banking or personal information by phishing techniques and their conversion into money or goods and services - has been steadily increasing for several years and the phenomenon does not seem to have occurred. On the contrary, it has become a widespread practice among web crooks due to the increased use of social networks, e-commerce, mobile devices [1] [2] and cloud solutions to store and manage sensitive data. To convince oneself of this, simply type the terms "bank", "fraud", "Scam" and "phishing" in Google or Google Scholar. We get almost a million results in Google and 10,800 in Google Scholar. This is how important the topic is on the Internet and arouses the interest of researchers and organizations that fight against phishing. Among these organizations, there is the Anti-Phishing Working Group (APWG) which published in its 2017 report [3] that more than 91% of all phishing attacks in 2016 targeted five types of industries in particular, financial institutions, cloud-based data hosts, web hosts, online payment services and e-commerce services. This figure of 91% represents an average increase of 33% per type of industry compared to 2015. An increase which is, however, abnormally high for Canadian companies that have experienced, among the developed countries, the strongest phishing growth in 2016, nearly 237% according to the Phishlabs 2017 report [3], mainly in the financial institutions sector, where the 444% ceiling was reached [4]. Trademarks targeted by phishing campaigns reached an average 2016 record of 380 per month, 13% higher than the previous year.

In addition to targeting businesses and trademarks, fraudsters target consumers who connect to the Internet.

Other statistics from the 2016 and 2017 APWG reports show that the number of detected websites that were dedicated to phishing attacks increased from 393K in 2014 to 1.22M in 2016, an increase of 310%. As for the number of domains where these sites reside, it would be 170K in 2016, which represents an increase of 23% compared to 2015.

Since March 2016, 93% of all phishing emails had a "ransomware" encryption system, according to a report published by Phishme Inc. [5]. Also, there is an increase in the types of attack targets. Attackers increasingly prefer to attack online payment systems like PayPal, Boleto, Bitcoin [6], and businesses that manage personal information.

Another, and not least, indicator of the extent of the phenomenon of phishing is the multiplicity, both in America and in the rest of the world, of national organizations and multinational coalitions of companies fighting this scourge. Their goal is to share information and know-how to reduce or even eliminate identity theft and fraud that result from the

IJDACR
International Journal Of Digital Application & Contemporary Research
International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 7, Issue 02, September 2018)

growing problem of phishing. These organizations include the FBI and NW3C partner Anti-Phishing Working Group, the Internet Crime Complaint Center (IC3), The Coalition on Online Identity Theft, the SCAMwatch website, The Federal Trade Commission of the United States, The 419 Coalition Website [7].

In summary, what we can learn from these numbers is that phishing fraud:

- Makes more and more victims;
- In the short term, at the individual level, losses and undermines confidence in the Internet for online transactions. And, at the corporate level, it would undermine the trust of customers and account holders in them and damage their images [8];
- Adapts more and more to new information technologies (e.g. SMS);
- Target new economic sources such as companies that manage information.

Given these constantly evolving numbers and knowing that more than two-thirds of Canadians (68%) use the Internet to access banking services, we find that it is justified to try to understand why, despite the intensification the efforts of companies that manage information and national organizations and multinational coalitions of companies that fight against this scourge, despite online banking very secure, this phenomenon is not fading and the impacts at victims are still rising [9].

It appears from the literature review that research in this field of activity has evolved a great deal for nearly a decade. It has gone from purely technological research that combines technological and human aspects [9]. It invites researchers to borrow notions from other disciplines such as criminology, economics and the social sciences to design countermeasures that take into account both information technologies and human aspects. The major challenge of research is there and it grows as attacking schemes become more refined and techniques more sophisticated.

The statistics presented above reveal a number of problems. First, whether the number of phishing attacks or the number of victims or the magnitude of the financial impact of these attacks, the figures recorded between 2013 and 2016 are constantly increasing. This raises a number of questions about the effectiveness of the countermeasures implemented to combat this phenomenon. According to Singh et al., No single technical measure will stop phishing altogether, but a combination of good organizational practices, correct use of current technologies, and improved knowledge and challenges. Security can reduce the occurrence of phishing attacks and the resulting losses [10]. However, the literature consulted reveals that there is a gap between this desire to design technological solutions centered on the human and the daily practice. For example, the fact that "anti-phishing" warning messages are not adapted to the dangerousness of the situation or to the characteristics of the user does not facilitate their understanding [11]. And, therefore, the user tends to ignore these messages. The challenge here for research is to work upstream so that the technical solution integrates human concerns. To do this, it is necessary to thoroughly study the strategies and attack techniques used by hackers to trap their victims, identify the predictors of bank fraud using real survey data so as to make current countermeasures more effective.

Secondly, the literature consulted separately studies the phishing attack and the clandestine markets of cybercrime products. It offers phishing taxonomies generally based on an attacker's perspective. For example, Aleroud and Zhou provide an integrated view of phishing that includes four dimensions: communication media, target environments, attack techniques, and countermeasures [12]. This way of characterizing phishing does not establish the necessary bridge between the theft of information and its exploitation for criminal purposes. However, we believe that these are two complementary facets of the same industry, that of phishing fraud because one has, on the one hand, the activities that contribute to the theft of information and, on the other, those who exploit this information for criminal purposes. The necessary bridge is the victim. She is the one who suffers the injury. Whether she receives spam (hooks), or when her information is stolen or when she gets money from her account. In this sense, it would be interesting and innovative to study the information monetization process and to integrate the key risk elements that result from it into a complete taxonomy of phishing, taking into account the perspective of the victim. What we did not find in the literature consulted. Probably for two reasons: because it is very difficult to obtain quantitative data on criminal activity from banks, police departments or in secret forums, but also because we have not found theoretical models analyze clandestine markets to identify key factors that influence the monetization process.

A few thoughts were acquired from Spoofguard and extra checks were added to make sense of the patterns inside the phishing sites. Notwithstanding, regardless of various situations it is hard to give most extreme precision. The center issue is to diminish the identification of false positives and

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 7, Issue 02, September 2018)

increase the true positives along these lines increase the accuracy of the framework.
The real worry of this exploration is to outline a system expected for appraisal of the lexical features to show signs of improvement through comprehensively studying the components of the URLs which promote phishing, by the methods of cuckoo search optimized neural network classifier.

II. PROPOSED METHOD

A. System Model

The classifier takes unclassified URLs as input, and returns a predicted binary class as output (either Phish or Benign). Our aim is to evaluate the effectiveness of URL features as discriminating features.

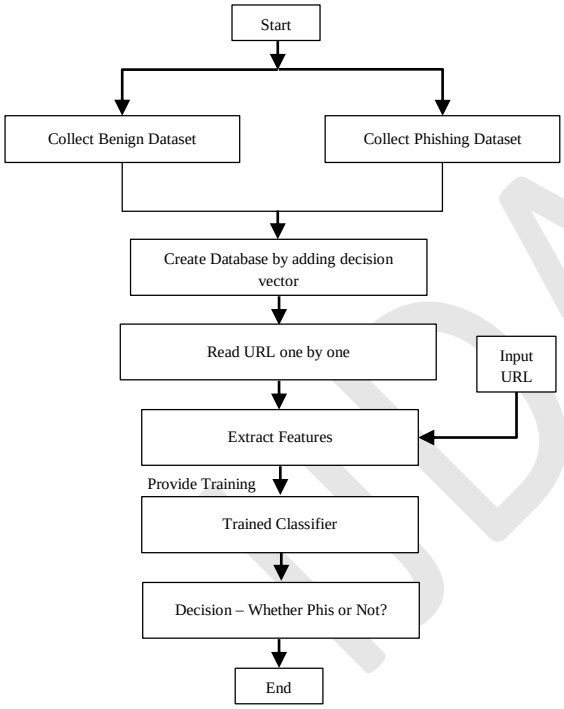

Figure 1: Flow diagram of proposed architecture

We started with collection of URLs and then after loading the URLs we started by reading URLs one by one for feature extraction. To facilitate feature extraction, each URL was split into three sections: protocol, domain, and path. All subsequent feature extraction was performed on these sub-regions. After collecting of URL features, the classifier's life initiates by a supervised learning phase. During this phase, the classifier is fed with pre-classified URL along with their pre-defined class. The classifier is then able to perceive a classification model. Once the learning phase is complete, the classifier is given unclassified URLs as input, and a predicted class is returned as output.
Architectures also hold room for checking a particular URL for Phishing. A random URL is provided to the trained classifier for recognizing the class (Phishing or Benign) of the given URL.

B. Collection of URLs

Here in this research work, we have taken URLs of benign websites from www.alexa.com [13] www.dmoz.org [14] and personal web browser history. The phishing URLs were collected from www.phishtak.com [15].

C. Lexical Feature Extraction

Lexical features are the textual properties of the URL itself, not the substance of the page it indicates. URLs are human-readable text strings that are parsed in a standard manner by customer projects. Through a multistep determination process, programs make an interpretation of each URL into guidelines that find the server facilitating the site and indicate where the site or asset is set on that host.

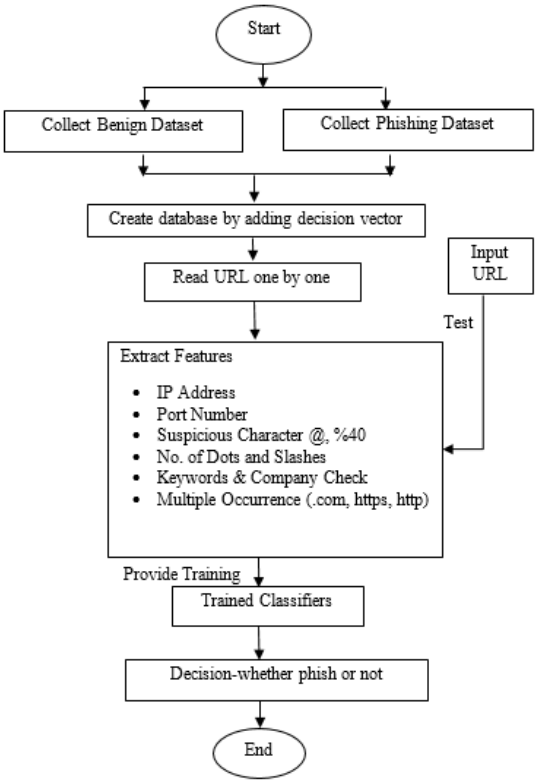

Figure 2: Flow diagram for lexical feature extraction

D. Classification Algorithm

The input to the classifiers in MATLAB is two .txt files; newben.txt and newphis.txt. The classification

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 7, Issue 02, September 2018)

algorithm considered for processing the feature set is describes as follows:

Cuckoo Search Optimized Neural Network

The training problem of an artificial neural network (ANN) is formulated as an optimization problem. Formally, given a function $f(w, X)$ that measures the error of the network when evaluating a set of training patterns X, where $w \in \mathbb{R}^d$ is the vector of weights or parameters of an ANN, the optimization problem is defined as:

$$\hat{w} = \min_{w \in \mathbb{R}^d} f(w, X) \quad (1)$$

$$f(w, X) = \frac{\sum_{i=1}^{|X|} (\hat{y}_l - y_i)^2}{|X|} \quad (2)$$

Where $\hat{y}_l$ and y_i are the expected output and the actual output of the network respectively for the pattern x_i of the set X. The definition of the objective function is also known as the mean square error (MSE).
The ANNs is variations of a distributed parallel processing model that is characterized by a group of aspects of which the most important for our analysis are listed below:

Processing Units (Neurons): Each processing unit performs a relatively simple job: receiving an input from its neighbouring units or external sources and using this input to produce an output signal that is then propagated to other processing units or to the output of the network. This research uses a type of processing units known in the literature as sigma-units whose propagation rule corresponds to Equation (3) [16].

$$S_k(t) = \sum_y w_{jk}(t) \times y_j(t) + \theta_k(t) \quad (3)$$

In Equation (3) w_{jk} is the weight associated with the input y_j and θ_k is the bias corresponding to the neuron k at a time instant t. Subsequently the value S_k is evaluated in an activation function to limit the contribution of the net input in the activation of the neuron. Frequently a non-decreasing function is used as shown in Equation (4).

$$F(S_k) = \frac{1}{1+e^{-s_k}} \quad (4)$$

Pattern of Connectivity between Processing Units: The processing units are connected to each other. The way in which these connections are established determines what the network is capable of representing and learning. Among the most frequent connection architectures are the feed-forward, recurrent and convolutional networks.
The feed-forward networks are the simplest and most used (Figure 3). This architecture is based on a group of layers of units organized in cascade. The units located in the same layer do not have connections between them, receive their input from the output of the units located in the previous layer, and send their outputs to the units in the back layer. For simplicity in the following it will be assumed that an ANN of the feed-forward type is formed by a layer of input neurons, which does not perform processing, a layer of intermediate or hidden neurons and a layer of output neurons. This particular configuration is widely recognized as a network of the multilayer perceptron type.

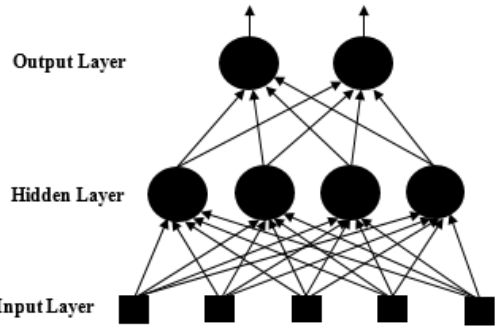

Figure 3: Neural network structure based on back propagation [17]

Learning Rule: For a network to recognize a particular problem, a procedure is necessary to modify the connectivity patterns based on the experience gained from the training patterns. This means training the network or, what is the same, modifying the weights w that weigh the importance of the inputs of each neuron [17].

Traditionally, the stochastic gradient descent (SGD) algorithm is used for training ANNs. This algorithm traces the parameters space of a network in such that the error function (w, X) is reduced, iteratively following the direction of an error gradient calculated at a random initial point of the parameter space. In each step small movements are made in the opposite direction of that gradient until it meets a minimum. Due to this behaviour, the family of algorithms based on a descending gradient is sensitive to converging in local minimums of the parameter space, and therefore is sensitive to the initial values of the network weights [17].

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 7, Issue 02, September 2018)

Neural Network Optimization

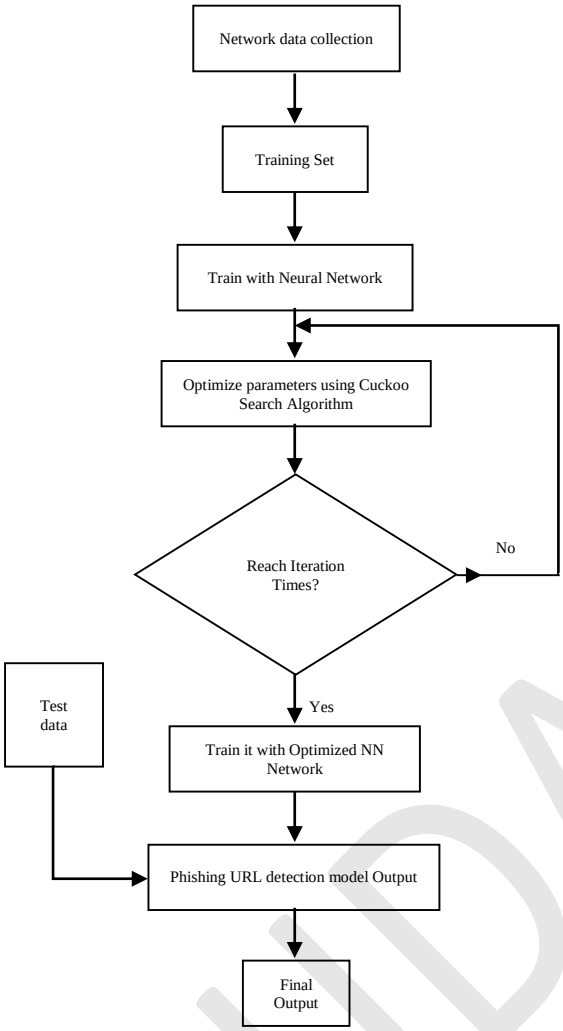

Figure 4: Flow diagram for optimized model based on Neural Network and Cuckoo Search approach

The implementation of the Levy flight in the Cuckoo Search approach is aimed to produce a novel resolution during the exploration process [18].

$$x_i^{t+1} = x_i^t + \alpha \oplus Levy(\lambda) \quad (5)$$

Where, $\alpha(\alpha > 0)$ is the jump size, x_i^{t+1} is the new solution and x_i^t is the current solution. This equation represents a random step called the Markov chain. This means that the next solution depends on the current solution and the probability of transition. $Levy(\lambda)$ follows a Levy distribution with infinite mean and infinite variance $(1 < \lambda \leq 3)$, Equation (6). This allows a part of the generation to move away from the current solution, preventing the algorithm from being trapped in the local minimums [18].

$$Levy(\lambda) \sim u = t^{-\lambda}, \quad (1 < \lambda \leq 3) \quad (6)$$

III. Simulation And Results

The performance of proposed algorithm has been studied by means of MATLAB simulation.

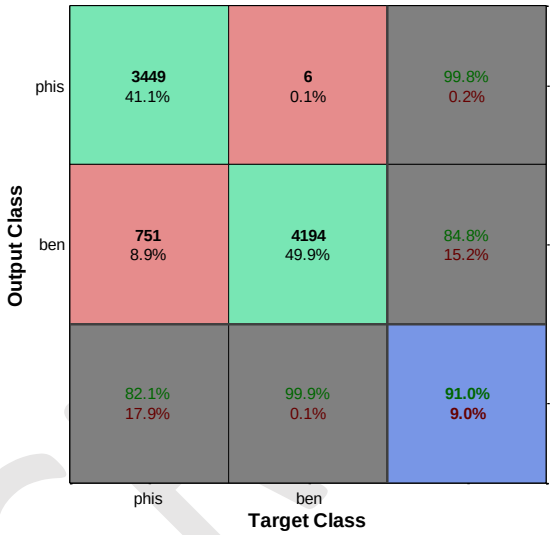

Figure 5: Confusion matrix plot for Neural Network classifier based method

Figure 6: Confusion matrix plot for cuckoo search optimized neural network method

IV. Conclusion

The presented model is an approach to the needs presented by the phishing URL, since it explores the entries and the appropriate algorithms giving a margin of success of approximately 100%, in this sense it is possible to obtain a high level of classification. For this reason, the classification and identification of phishing pages may be approaching a model.

The results presented in the presented model require obtaining results with a higher level of accuracy,

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 7, Issue 02, September 2018)

since the need in terms of safety should be close to 100% with a fault tolerance of 0.001%.

The database "phishing web" offers a number and variety of attributes established by all the literature, however, the tests carried out show that of the 60-40 split case is presented in the simulation, nevertheless, it is proposed a possible consensus on the attributes that can come to clearly define a phishing URL. On the other hand, the amount of consigned attributes turns out to be an inconvenience due to the "curse of the dimension", since understanding and processing all these attributes translates into space, time and costs.

In this work, the gain that occurs when using classification techniques such as cuckoo search optimized neural network classifier is revealed at the theoretical level, even though no technique is superior to the others in a general way, since they have limitations and own advantages that are coupled according to the model we are working with.

REFERENCE

[1] Kritzinger, E., & von Solms, S. H. (2010). Cyber security for home users: A new way of protection through awareness enforcement. *Computers & Security*, *29*(8), 840-847. doi:http://dx.doi.org/10.1016/j.cose.2010.08.001

[2] Hille, P., Walsh, G., & Cleveland, M. (2015). Consumer fear of online identity theft: Scale development and validation. *Journal of Interactive Marketing*, *30*, 1-19.

[3] Phishlabs. (2017). *2017 Phishing Trends and Intelligence Report: Hacking the Human*. Retrieved from https://pages.phishlabs.com/rs/130-BFB-942/images/2017%20PhishLabs%20Phishing%20and%20Threat%20Intelligence%20Report.pdf

[4] Wagner, Thomas D., Esther Palomar, Khaled Mahbub, and Ali E. Abdallah. "Relevance Filtering for Shared Cyber Threat Intelligence (Short Paper)." In *International Conference on Information Security Practice and Experience*, pp. 576-586. Springer, Cham, 2017.

[5] PhishMe. (2016). *PhishMe Q1 2016 Malware Review*. Retrieved from https://phishme.com/project/phishme-q1-2016-malware-review/

[6] Jaeger, J.-M. D. (2016). Des pirates volent 72 millions de dollars à une plateforme de Bitcoin. Retrieved from http://www.lefigaro.fr/secteur/high-tech/2016/08/03/32001- 20160803ARTFIG00143-des-pirates-volent-72-millions-de-dollars-a-une-plateforme-de-bitcoin.php

[7] Phishing-Initiative. (2017). Phishing Initiative. Retrieved from http://www.phishing-initiative.com/

[8] Symantec. (2014). *Internet Security Threat Report 2014, volume 19*. Retrieved from http://www.symantec.com/fr/ca/security_response/publications/threatreport.jsp

[9] Chaudhary, Sunil. "The Use of Usable Security and Security Education to Fight Phishing Attacks." (2016).

[10] Singh, A. C., Somase, K. P., & Tambre, K. G. (2013). Phishing: A Computer Security Threat. *International Journal of Advance Research in Computer Science and Management Studies*, *1*(7)

[11] Mayhorn, C. B., Murphy-Hill, E., Zielinska, O. A., & Welk, A. K. (2015). The social engineering behind phishing. *The next wave*, *21*(2).

[12] Aleroud, A., & Zhou, L. (2017). Phishing environments, techniques, and countermeasures: A survey. *Computers & Security*, *68*, 160-196.

[13] "The Web Information Company," [Online]. Available: www.alexa.com.

[14] "DMOZ Open Directory Project," [Online]. Available: http://www.dmoz.org.

[15] "PhishTank," [Online]. Available: https://www.phishtank.com/.

[16] Lakshmi, V. Santhana, and M. S. Vijaya. "Efficient prediction of phishing websites using supervised learning algorithms." *Procedia Engineering* 30 (2012): 798-805.

[17] Mohammad, Rami, T. L. McCluskey, and Fadi Abdeljaber Thabtah. "Predicting phishing websites using neural network trained with back-propagation." World Congress in Computer Science, Computer Engineering, and Applied Computing, 2013.

[18] Rajabioun, Ramin. "Cuckoo optimization algorithm." *Applied soft computing* 11, no. 8 (2011): 5508-5518.