PERFORMANCE COMPARISON OF PREDICTIVE MODELS BASED ON REDUCED PHISHING FEATURE CORPUS

Abdul A. Orunsolu

MOSHOOD Abiola Polytechnic Abeokuta, Department of Computer Science

Corresponding author: orunsolu.abdul@mapoly.edu.ng

ABSTRACT *– Phishing is currently one of the severest cybersecurity challenges facing the cybercommunity. Either during good times or bad times, phishers exploit the vulnerabilities within the communication chain to lure their victims to fake websites where their sensitive personal credentials are harvested. Although various anti-phishing approaches have been proposed, the problem of phishing continues unabated. Based on the foregoing, the solution to phishing demand constant investigation to win the arms race through predictive models which have shown promising performance results in extant literature. In this paper, the performance comparison of some selected predictive models is presented using a reduced feature set to ensure their deployment on mobile devices. The approach is evaluated using a dataset consisting of 10,000 phishing instances. The approach provides the performance metrics of various state of art machine learning approaches popular with phishing detection.*

KEYWORDS *– Anti-Phishing, Cyber-attacks, Identity theft, Middleware, Threats*

1. INTRODUCTION

Cyberspace remains the single largest space where different services are enjoyed on the click on the mouse by the human race. This is because the introduction of the Internet has brought tremendous changes to humanity. Virtually all aspects of human endeavours are available online. This change has no doubt improve the relationship, enhance communication, promote governance, speed up entrepreneurial skills etc. However, these significant positive changes are been currently challenged by myriads security problems. These security problems have rendered great havoc on the potential benefits on the internet. One of such security challenges is the problem of phishing attacks.

Phishing attacks are criminal attempts that fraudulently deceived unsuspecting online users through fake websites or message into divulging their sensitive personal credentials. These credentials are then used by phishers to commit identity theft on behalf of the victims. These attacks have resulted in significant damages ranging from online brand damages to huge financial losses (Abdelhamid et al., 2014; Qabajeh et al. 2018). A phishing attack involves the phishers setting up a counterfeit website that perfectly mimic the appearance of a known online brand or message icon. The online users are then deceived to access the fake website through a message from unfamiliar information sources. In this process, most online users get their sensitive credentials harvested by cybercriminals. The credentials harvested by the phishers normally include bank account numbers, passwords or PINs, credit card numbers, security questions, security codes etc. With the harvested credentials, the phishers can log in into the genuine websites to steal the victim's money or launch other related attacks. Most online users are vulnerable to phishing threat because they are so comfortable navigating Web pages and the URL that name them. Besides, as the number of mobile-connected devices using social networking sites such as Facebook continues to grow, the motivation for phishers to target the platform also increases as malicious links can be easily embedded into e-chat (Aggarwal et al. 2012; Kumar and Kumar, 2014; Orunsolu et al. 2018).

Due to the numerous threats posed by phishing attacks, the online security community and industry have come with several solutions called anti-phishing systems (Kumar and Kumar 2014). One of the promising anti-phishing techniques is the adoption of machine learning model in mitigating the problem of phishing attacks (Hamid and Abawajy 2014; Tan et al. 2017). Numerous anti-phishing predictive models have been developed to combat phishing attacks. These predictive models have shown significant performance results in terms of high accuracy, low false positive and false negatives and zero-day detection capability (Sonowal et al. 2017; Adebowale et al. 2018; Mao et al. 2019; Orunsolu et al. 2019). However, the performance of these predictive models is heavily dependent on the types of machine learning algorithm adopted and the type/size of heuristics in the feature set corpus (Qabajeh et al. 2018).

In this work, we proposed an approach to examining the different state of art predictive model using reduced phishing feature corpus to resolve the uncertainties that result from performance issues and

other inconsistencies in feature set corpus. The primary element of this approach is the composition of the feature set. It considers various factors that have been examined in the literature for the most representative features set (Varshney et al 2016; Fadheel et al. 2017). Specifically, this approach leverage on the feature frequency analysis technique for resultant feature selection (Orunsolu et al. 2019). Besides, our choice of ML algorithms included in the performance measurement is informed by their existing results in extant literature (Basnet et al. 2007; Fadheel et al. 2017). The contribution of this paper is to improve the deployment of predictive models through slight tuning of feature set with significant performance accuracy. The paper also presents the advantage of improving the discoverability of choice of feature set corpus.

The rest of the paper is organized as follows: Section 2 presents a literature review on classification algorithms in phishing detection. The reduced feature set algorithm is discussed in Section 3. In Section 4, the application and results of the different predictive model on the proposed feature set are presented. Conclusions and future works are presented in Section 5.

2. CLASSIFICATION ALGORITHMS IN PHISHING DETECTION APPROACH

A classification algorithm automatically learns how to make accurate predictions of unknown instances based on the past or trained observations. Given an identity (i.e. malicious or legitimate) and a set of features, the task of determining the genuineness of a transaction is executed by a classification algorithm. The accuracy (i.e. True Positives, False positives, Recall Rate, True Negatives etc.) and the resources-requirements (i.e. training time, response time, memory overhead, etc.) with which the predictions are made determine the efficiency of such classification algorithms. Some classification algorithms have been employed for efficient classification of web documents e.g. email, spam messages, webpage, e-chat etc (Hota et al. 2018; Orunsolu et al. 2019). Such classification algorithms include Decision Tree, K-nearest neighbour, Naïve Bayes, Support Vector Machine, CART, C4.5 etc. These classification algorithms have advantages and disadvantages which necessitates their adoption for a particular problem (Qabajeh et al. 2019; Pham et al. 2014). Some machine learning algorithms used in this approach are described as follows:

i. Naïve Bayes Classifier: This is a simple prediction and classification algorithm which use the joint probabilities of certain features to estimate the conditional independence assumption of other unknown attributes. This classifier is more practical because it does not require a very large training set and can easily handle missing attribute values. It has been researched in many anti-phishing systems with significant performance accuracy. For instance, Han et al. 2012 used NB algorithm on login user interface information of whitelisted websites to achieve an efficient anti-phishing system. Besides, Orunsolu et al. 2019 used NB on certain heuristics from the URL, Webpage properties and webpage behaviour to design an efficient anti-phishing predictive model.

ii. RandomTree: This is another classifier that has been widely used in phishing detection (Mao et al. 2019; Garera et al. 2007). It consists of an ensemble machine learning method used for the purpose of classification, regression and other data mining tasks. The approach operates basically by constructing a multitude of decision trees at the training time and produces the output as a class that is the mode of the classes or mean prediction of the individual's trees.

iii. Support Vector Machine: This is one of the most popular classifiers in designing machine-learning based phishing detection model (Orunsolu et al. 2019; Hota et al 2018). The classifier takes a set of marked training samples, each as belonging to one or the other of two categories and the SVM model builds an algorithm that assign new examples to one or other category. The model is therefore referred to as a non-probabilistic binary classifier. For instance, Zouina and Quttaj (2017) examined a SVM predictive model using URL features with a remarkable performance results.

iv. Neural Network: This is a series of predictive algorithms that attempt to recognize the underlying relationships define in a set of data using a scientific process or model that mimics the operation of human brain. This classification model generates the best possible result without redefining the output criteria. Basically, a neural network consists of layers of interconnected nodes and each node is a perceptron. This perceptron is similar to a multiple linear regression.

v. Decision Tree: This is a classification algorithm whose goal is to create a machine learning model that correctly predicts the value of a target sample based on some input samples. Decision Trees consists of basically two main types namely the classification tree and regression tree. In phishing detection system, the term Classification and Regression Tree (CART) analysis has been used to describe most research in this area. Notable examples of decision tree algorithms include Iterative Dichotomiser 3, C4.5, Conditional Inference Trees, Chi-square automatic interaction detection etc. For instance, Li et al. 2019 investigated an anti-phishing approach where Decision Tree was used on features from URL and HTML. The approach indicated the superior performance of this classifier in phishing detection.

Related Works

Since phishing problem is a typical example of classification puzzle, the machine learning and data mining methods have been employed in a number of research works. Machine learning based anti-phishing scheme are enhanced through classification algorithms to detect or predict phishing activities using certain features usually called *fingerprints.* The class of anti-phishing design remains so popular because of its advantages of minimizing false positives and ability to generalize phishing detection using known instances. This is possible as the ML algorithm can produce powerful predictive model once the initial feature sets have been chosen.

Several works have been published using several classification algorithms to demonstrate the effectiveness of this approach. For example, Han et al. (2012) investigated a whitelist approach in which Naïve Bayes algorithm was employed to capture login information to make prediction of the status of a loading page. The approach produced significant phishing detection model. However, the approach is susceptible to new login problem and pharming attacks. In other related works, Orunsolu et al. (2019) proposed a predictive model for phishing detection using frequency analysis of existing feature corpus to create a more discriminative feature class. The system used an aggregate of 15-dimensional feature set trained using Naïve Bayes and Support Vector Machine. The system achieved a remarkable performance with 99.96% accuracy with low false positive. Jain and Gupta (2016) proposed a list-based anti-phishing scheme using SVM approach. Their approach produced a fast access time method and detection rate of 826.02%. However, the false negative rate of 1.48% limited the application of the approach in a critical financial web transaction. In another application of SVM model, Mao et al. (2019) investigated an anti-phishing system based SVM machine learning approach using the visual analysis method. The approach considered webpage layouts using property vector extraction, property vector generation and comparison vector generation. The method produced a significant accuracy of more than 93.0%. Zouina and Outtaj (2017) studied URL features using SVM model to produce a lightweight phishing detection system. Their method considered six features extracted from the domain address of a querying page. Using evaluation dataset from PhishTank and Alexa, the system produced an accuracy rate of 95.80%.

Using ensemble machine learning approach, Hamid et al. (2011) analyzed various machine learning models like Bayesian Net, AdaBoost, Decision Tree and Random Forest. In their evaluation, phishing dataset consisting of two separate partitions are used for training and testing purposes. The results indicated that Random Forest produced the highest accuracy of 93%. Similarly, Hota et al. (2018) investigated an approach where features are remove and replace from original feature set in a random manner until a certain accuracy threshold is achieved. This method is called Remove-Replace Feature selection technique (RRFST). The approach achieved an accuracy of 99.27% with an ensemble of C4.5 and CART. In an earlier related work, Mohamed et al. 2014 examined the problem of phishing detection using a number of rule induction algorithms. The authors evaluated their approach with dataset tested on C4.5, CBA, RIPPER and PRISM. Similarly, Khadi and Shinde (2014) investigated the problem of email phishing detection system by combining a RIPPER ML algorithm with fuzzy logic on a number of features from fingerprints. The approach produced a prediction rate of 85.4%. Recently, Li et al., 2019 considered a stacking approach with 20 features extracted from the URL and HTML. The extracted features were subjected to training using an ensemble model of Gradient Boosting Decision Tree, XGBoost and LightGBM. The approach which was evaluated using a large dataset achieved a remarkable accuracy of 98.60% accuracy and 1.54% false alarm rate. In a similar vein, Adebowale et al. 2018 investigated an integrated approach consisting of 35-dimensional features set where an Adaptive Neuro-Fuzzy Inference System was employed. The integrated features consists of text, images and frames selected using Chi-Square Statistics and Information Gain technique. The scheme was evaluated with predictive model consisting of SVM, K-NN and ANFIS. This system achieved 98.3% accuracy.

Chin et al. (2018) presented an approach called PhishLimiter where a deep packet inspection (DPI) and software-defined networking method was used to identify phishing activities in email and web-based communication. The approach adopted an Artificial Neural Network model with accuracy of 98. 39%. Similarly, Seymour and Tully (2018) considered a new ML based on NN called Long Short Term Memory Artificial NN to combat the problem of spear phishing on online social networks. The model which presented word vectors after training process consisting of different post messages. The approach provided experimental results that indicated that the proposed system was superior to other manual classification approaches. In one of the earlier approach to NN, Mohammad et al. (2014b) developed a Neural Network-based anti-phishing model that improve the learned predictive model based on system's previous training experiences. The authors posited the use of self-structuring Neural Network classification approach to cope with the changing

nature of phishing fingerprints. The authors considered about thirty features to investigate the accuracy of their model. The evaluation process involved more than 10000 instances with a remarkable accuracy.

3. REDUCED FEATURE GENERATION ALGORITHM

Feature generation algorithms are used to identify or create certain characteristics of a particular dataset on which prediction of non-class member can be based. In phishing detection, the generation of relevant features remains central to the performance of data mining and machine learning algorithms. Although, several features have been proposed in extant literature, the task of generating the most representative feature remains a big task in any anti-phishing studies. While some works (Zouina et al. 2017; Mao et al. 2019; Hota et al. 2018), consider a single class of feature such as URL, Visual Similarity etc. in their studies, others considered integrated features involving two or more categories (Adebowale et al. 2019; Orunsolu et al. 2019; Li et al. 2019). In either cases, efforts are toward obtaining classifier with greater performance accuracy while resources requirements are reasonable. It is therefore imperative to continuing evaluating the performance of different classifiers on a number of features in other to keep anti-phishing model efficient and relevant. For instance, Gupta et al. 2016 and Toolan et al. 2010 provided the ranking categories for different features used in phishing and spam detection. It is based on the premise that we identified the following features in this study of certain classifiers to increase the coverage of anti-phishing studies.

In this study, the phishing dataset includes 13 features extracted from 10,000 instances as captured in a WEKA applications. The feature set consists of 85% URL-based category and 15% non-URL category. This is because of the popularity of URL-based features in most anti-phishing studies (Sahingoz et al. 2019; Qabajeh et al. 2018; Orunsolu et al. 2019; Adebowale et al. 2018). The features are selected due to their frequency of occurrence in most existing works. This popularity is not unconnected with the superior performance of these features with negligible response time (Zouina and Quttaj (2017); Orunsolu et al. 2019). The other non-URL features were randomly chosen without any regard to their underlying contributive significance. The purpose of this is to investigate the effect of these features on the other features which has exhibited a more superior performance i.e. URL-features. Table 1 contains the selected features and their short description.

Table 1. Selected Feature set

S/N	Feature name	Description
1	Number of Dots	This is used to elongate domain name address by adding irrelevant prefix or suffix to genuine URL
2	URL Length	Phishers use long domain name to disguise fake website
3	@ Symbol	This is used by phishers to redirect to phishing domain
4	No HTTPS	Most phishing website are hosted on non-https domain by phishers due to its non-expensive nature
5	Domain in path	Phishers make use of the domain name in the links in order to hide the identity of malicious link in the address
6	Https in Hostname	Fraudsters make use of subdomain to make malicious link look legitimate
7	Path Length	Phishers add the domain mane of genuine site within the path length of a URL to deceive users
8	IP address	This involves the use of IP address in order to obscure a server's identity by phishers
9	Popup Window	Phishers used pop-window to circumvent data validation during authentication process
10	Submitting to Email	This involves phishers using servers that are different from the loading page to obtain users credentials
11	Missing Title	Phishers often host their domain name on compromised domain whose domain keywords do not relate to its brand.
12	IFrame redirection	Phishers use Html tag that display additional pages invisible without a frame border
13	Return URL Length	Phishers use URL that does not return to a particular whois server by obfuscating web address using unrelated information in the URL path

4. APPLICATION AND RESULTS

The evaluation procedure used in this paper is aim to compare different classification models on the proposed reduced phishing features selected from URL and non-URL fingerprints. The dataset used for the evaluation consists of 10000 phishing instances that were imported into a WEKA application. The JSoup HTML parser was used to extract the feature set from the experimental dataset instances. On the other hand, the Weka application provides the environment where the extracted features are trained

and tested with different classification algorithms. A typical Weka preprocess interface for the proposed model is presented in Figure 1. The figure indicated the extracted features, size of evaluation dataset and other defaults settings in Weka application. These features can be reverted in WEKA to show the contribution of each or group of selected features.

The evaluation metrics consists of True Positive (TP) rate and False Positive (FP) rate. The TP is the rate of phishing instances that are correctly predicted as phishing out of the total phishing instances. On the other hand, the FP is the rate of phishing instances that are misclassified as legitimate out of the total phishing instances. The experimental dataset instances were separated into training and testing data using a 10-fold cross validation techniques. Usually, cross validation technique is a predictive model that evaluate the performance of a machine learning model on new instances in relation to certain portion of the dataset. Thus, the 10-fold cross validation randomly split the test dataset into ten equal portions where a single portion is used to validate the training of the other remaining portions. This process is necessary to generalize the performance of the predictive model to independent data corpus while providing error performance verification for the machine learning model (Orunsolu et al. 2019).

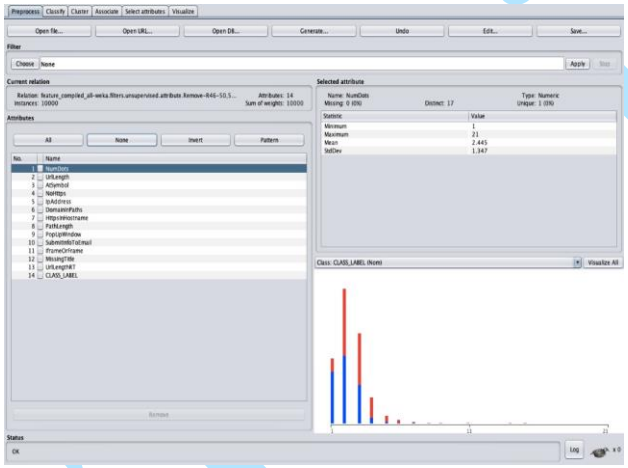

Figure 1. Feature Interface in WEKA

Figure 2 presented the visualization effects (VE) of different features used in the proposed system. The VE clearly shown that the URL features have more discriminative predictive power than the non-URL features. Specifically, the HTTPS in hostname obviously separated the data instances into two points while the other features produced significantly different color patterns of the experimental data instances. This function can be extended to construct confusion matrix and Receivers Operating Curve model of the approach. A confusion matrix is a table that is often used to describe the performance of a classification model while ROC estimate the predictive accuracy of the proposed model.

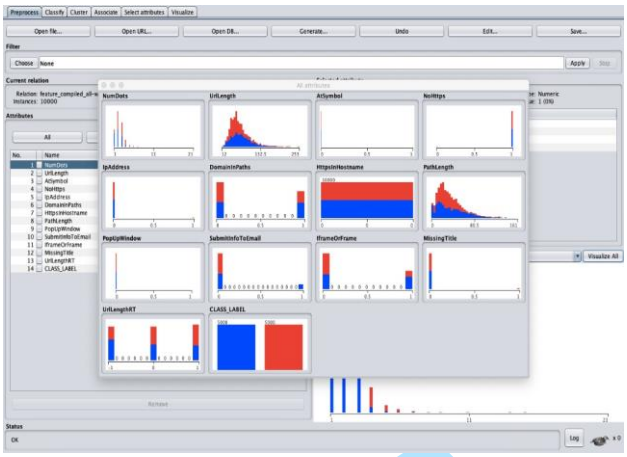

Figure 2. Feature Visualization in Weka

Table 2 presents the experimental results for the different classifiers used in evaluating our phishing *fingerprints*. The classifiers used in the experimental process are Naïve Bayes, SVM, ANN, RF and DT. The results indicated that Random Tree outperform other classifiers with significant accuracy of 96.1% and a ROC value of 98.7%. This is followed by the Decision Tree classifier with accuracy of 78.2% and a ROC of 85.7%. The Multilayer perceptron model (ANN) performed next to DT with 74.6% accuracy and 82.4% ROC value. The SVM classifier produced accuracy of 72.9% and a ROC value of 72.9%. The least performed classifier was NB with 69.9% predictive accuracy and a ROC of 77.9%.

Table 2. Performance statistics of proposed classifiers

Classifier	TP	FP	Precision	Recall	F1-Score	ROC
RT	96.1	0.39	96.1	96.1	96.1	99.7
DT	78.2	2.18	78.6	78.2	78.1	85.7
ANN	74.6	2.50	75.6	74.2	73.9	82.4
SVM	72.9	2.71	74.2	72.9	72.8	72.9
NB	69.9	3.01	70.2	69.9	69.8	77.8

These results indicated that even the least performed classifier hover well-above average in experimental results. In addition, the range of the ROC values from 98-77% indicated a good predictive accuracy of the selected classifiers and features. Similarly, the low FP of RT models is a promising feature that indicates that the model can be integrated into critical web transaction for determining the status of a loading website. Thus, the predictive models based on the reduced phishing feature sets can produced a good generalization model on which a good classifiers can be built. This is because these results indicated that the selected features present a good fingerprints on which anti-phishing model can be investigated. Figure 3 and 4 presented some interesting interfaces from our experimental procedure. The confusion

matrix can be seen from Figure 3 as indicating the distribution of classification data points with regard to predictive accuracy. Figure 4 presented the Multilayer Perceptron of the ANN predictive model with respect to feature input and its binary output value.

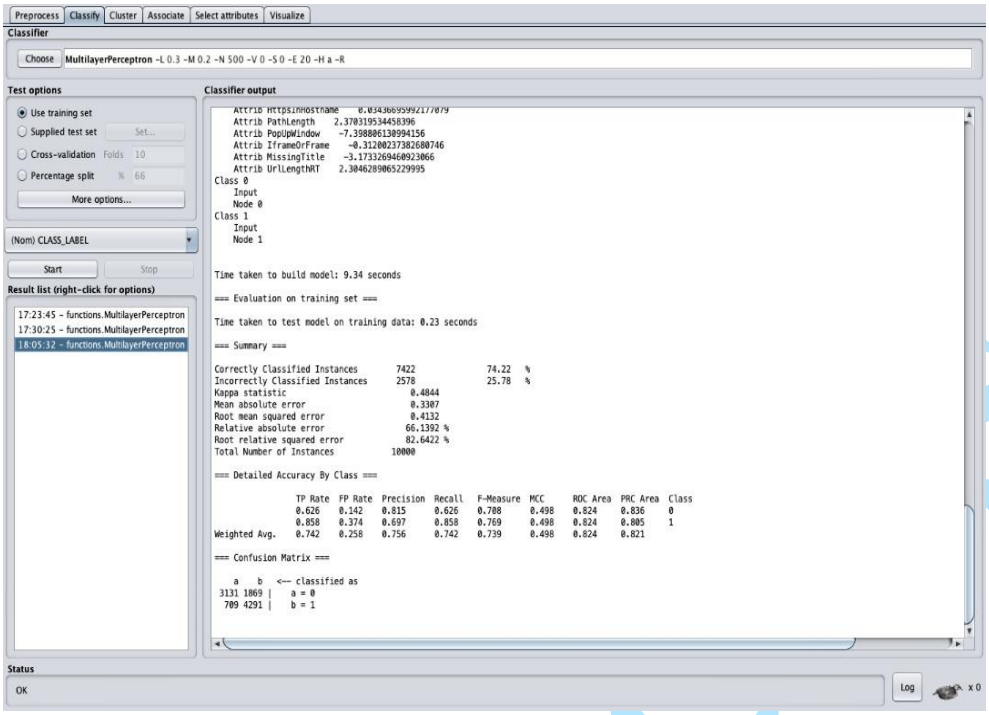

Figure 3. ANN performance statistics and cost matrix

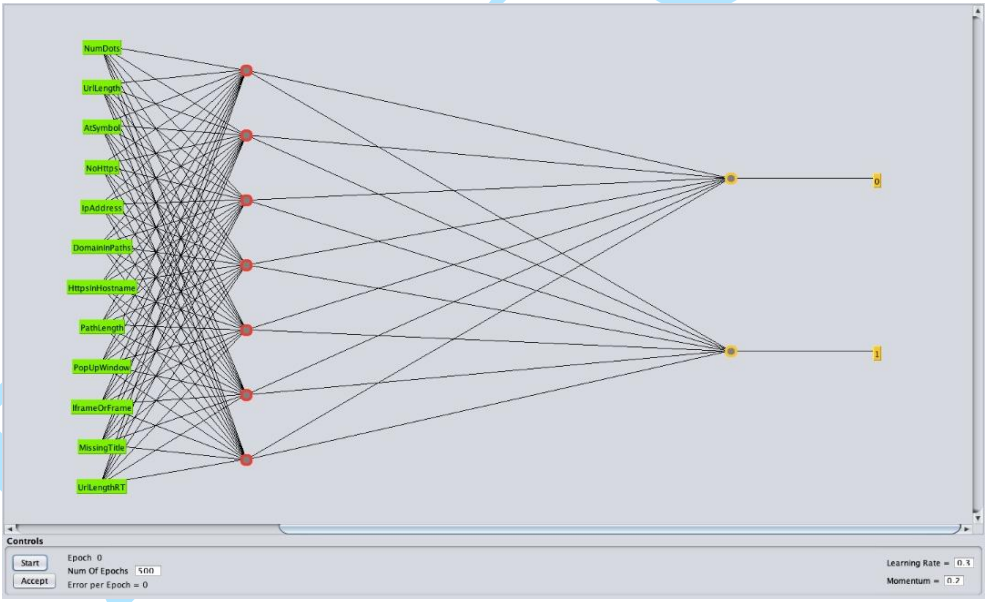

Figure 4. Visualization of ANN predictive model

CONCLUSIONS

In this work, performance evaluation of different classification models are considered with respect to a smaller feature set. The features are selected from extant literature with particular consideration for URL features due to their sterling performance in existing works where they have been applied. These features are then trained and tested using 10000 phishing instances on five different classifiers. The experimental procedure was implemented using JSoup Parser and WEKA application. The JSoup Parser was used to extract the selected features from the loading experimental instances while WEKA provide the running environment for the preprocessing and evaluation of the different classifiers adopted in this work. Cross validation experiment was used to generalize and verify error performance associated with the different classifiers. Specially, a 10-fold cross validation experiment was

chosen. The experimental results indicated that Random Tree outperform other classifiers with a remarkable accuracy and low false positive. These results showed that this approach presents a more accurate predictive models for mitigating phishing attacks.

REFERENCES

[1] Adebowale M., Lwin K., Sanchez E and Hossain M. (2018). Intelligent Web-Phishing Detection and Protection Scheme using integrated Features of Images, Frames and Text. Expert System with Applications.

[2] CSO Online report on phishing activities. Accessed 2016 (www.csoonline.com/articles)

[3] Chiew L., Chang H., Sze N and Tiong K. (2015.) Utilization of website logo for phishing detection. Computer and Security Journal.

[4] Gowtham R and Krishnamurthi I. (2014). PhishTackle-a web services architecture for anti-phishing. Cluster Compt.

[5] Han W, Cao Y, Bertino E and Yong J. (2012).Using automated individual white-list to protect web digital identities. Expert Systems with Applications.

[6] Hamid A and Abawajy, J. 2014. An approach to profiling phishing activities. Journal of computer and security. Elsevier Press

[7] Hota H.S, Shrivas A.K and Hota R. (2018). An Ensemble Model for Detecting Phishing Attack with Proposed Remove-Replace Feature Selection Technique. International Conference on Computational Intelligence and Data Science. Procedia Computer Science. Vo. 123, pp. 900-907

[8] Jain A and Gupta B. (2017). Two-level authentication approach to protect from phishing attacks in real-time. J. Ambient Intell Human Comp. DOI 10.1007/s12652-017-0616-z

[9] Jain AK, Gupta BB (2016) A novel approach to protect against phishing

[10] attacks at client side using auto-updated white-list. EURASIP J Inf Secur 2016:1–11

[11] A. Khadi, S. Shinde, Detection of phishing websites using data mining techniques, Int. J. Eng. Res. Technol. 2 (12) (2014).

[12] Mao J, Bian J., Tian W., Zhu S., Wei T., Li. A. and Liang Z. (2019). Phishing Page detection via classifier from page layout feature. EURASIP Journal of Wireless Communication and Networking. Vol 43,

[13] Mohammad R and Thabtah L and McCluskey. (2014). Tutorial and critical analysis of phishing websites methods. Comp Sci. Rev. J

[14] R. Mohammad, F. Thabtah, L. McCluskey, Predicting phishing websites based on self-structuring neural network, J. Neural Comput. Appl. (ISSN: 0941-0643) 25 (2) (2014) 443–458. Springer.

[15] Orunsolu A. Sodiya S. and Akinwale A. (2018). A Predictive Model for Phishing Detection. Journal of King Saud University-Computer and Information Sciences.

[16] Phishtank dataset (2018). http://www.phishtank.com.

[17] Qabajeh I., Thabtah F. and Chiclana F. (2018). A recent review of conventional vs. automated cybersecurity anti-phishing techniques. Computer Science Review.

[18] Tan C., Chiew L and Sze N. (2017). Phishing Webpage Detection Using Weighted URL Tokens for Identity Keywords Retrieval. Lecture Notes in Electrical Engineering. Vol. 398.

[19] J. Seymour, P. Tully, Generative Models for Spear Phishing Posts on SocialMedia. Technical report, 2018.

[20] Varshney G, Misra M. and Atrey P. (2016). A survey and classification of web phishing detection Schemes. Security Comm. Networks

[21] Varshney G, Misra M., and Atrey K. (2016). A phish detector using lightweight search features. Comput Secur;62:213–28.