Using Attribute-based Feature Selection Approaches and Machine Learning Algorithms for Detecting Fraudulent Website URLs

Mustafa AYDIN
Informatics Institute
Middle East Technical University (METU)
Ankara, Turkey
Banking Regulation and Supervision Agency
Istanbul, Turkey
e-mail: maydin@bddk.org.tr

Ismail BUTUN
Department of Computer Science and Engineering
Chalmers University of Technology
Goteborg, Sweden
e-mail: ismail.butun@chalmers.se
ORCID ID: 0000-0002-1723-5741

Kemal BICAKCI
Computer Engineering Department
TOBB University of Economics and Technology
Ankara, Turkey
e-mail: bicakci@etu.edu.tr

Nazife BAYKAL
Computer Engineering Department
Northern Cyprus Campus of METU
Guzelyurt, Cyprus
e-mail: nzbaykal@metu.edu.tr

***Abstract*—Phishing is a malicious form of online theft and needs to be prevented in order to increase the overall trust of the public on the Internet. In this study, for that purpose, the authors present their findings on the methods of detecting phishing websites. Data mining algorithms along with classifier algorithms are used in order to achieve a satisfactory result. In terms of classifiers, the Naïve Bayes, SMO, and J48 algorithms are used. As for the feature selection algorithm; Gain Ratio Attribute and ReliefF Attribute are selected. The results are provided in a comparative way. Accordingly; SMO and J48 algorithms provided satisfactory results in the detection of phishing websites, however, Naïve Bayes performed poor and is the least recommended method among all.**

Keywords — Attribute-based feature selection, Cyber theft, Data analysis, Fraudulent website detection, Machine learning algorithms

I. INTRODUCTION

Phishing attack is a one type of the cyber theft that aiming hijacking Web user's sensitive information, for instance Personal Identification Information (PII) such as citizen ID number, passport number, etc., or passwords of online banking accounts, credit card information or other financial data [1]. In this type of attack, users are lured that they are using a legitimate Web service and provide their confidential information to fraudulent websites that mimic the original one. In recent years, many internet users lost their money from the phishing attacks, which are growing in numbers every day.

Web applications that use the browser execution model are the most common form of delivering software nowadays. Because of the necessities in many different areas and devices, having a standard application delivery and development method is leading to the appearance of the Web application model based on client-side code execution. On the contrary, several security challenges and client-side code execution presents difficulties that are non-trivial to overcome. Unless security measures are taken on the Web application interfaces, such as the Web application integrity protection method mentioned in [2]; phishing attacks will exist and need to be detected/prevented before they can harm clients.

Therefore, many anti-phishing methods have developed in the recent past. On the contrary, the statistics show that the phishing attacks are still increasing. The purpose of this study is to identify decisive features from the URL structure of phishing web pages. In this study, we analyzed attribute-based feature selection methods and three different machine learning algorithms for phishing website analysis based on URL properties to specify the most effective features.

II. LITERATURE REVIEW

Several researchers have worked on this specific topic of "Phishing Website detection". Here in this section, we introduce most of the significant related work within the literature according to chronological order:

2010: Aburrous *et al.* [3] presented a resilient layered system in detecting phishing counterparts of the on-line banking websites, which is based on fuzzy logic combined with five different Data Mining (DM) algorithms (C4.5, JRip, RIPPER, PART, PRISM, and CBA). The approach is based

978-1-7281-3783-4/20/$31.00 ©2020 IEEEE

on a total of 27 features along with 6 criteria: (a) *URL and Domain Identity* (5 features); (b) *Security and Encryption* (4 features); (c) *Source Code and JavaScript* (5 features); (d) *Page Style and & Contents* (5 features); (e) *Web Address Bar* (5 features); and (f) *Social Human Factor* (3 features). Their assessment is performed in three steps: (1) *Fuzzification* – for each feature a range of values are assigned to descriptors related to linguistics such as low, medium, high, etc., and valid ranges of the inputs divided into fuzzy sets; (2) *Rule generation using classification algorithms* – using DM classification and association rules to automatically obtain significant phishing characteristics in the archived data of phishing websites (a total of 606 phishing websites from APWG archive and PhishTank); and (3) *Defuzzification* - converting a fuzzy output into a scalar value which determines if a website is Very legitimate, Legitimate, Suspicious, Phishy or Very Phishy. They used a test mode which employs 10-fold cross-validation as a testing mode and achieved a detection accuracy of 86.381% with PART.

2011: Xiang *et al.* developed a layered phishing detection system named CANTINA+ which uses machine learning techniques, search engines, HTML Document Object Model (DOM), third party services along with the expressiveness of a rich set of 15 webpage features including eight novel ones proposed with the goal of improving the true positive rate [4]. They also designed two novel filtering algorithms (a near□duplicate phish detector that utilizes a login form filter and hashing) to reduce the false-positive rate and human effort. CANTINA+ comprises three main modules: (1) the cross-comparison similarity of the website to the known phishing attacks via hashing; (2) using heuristics to filter websites with no login forms that request sensitive information before the classification phase; and (3) using machine learning techniques over 15 highly expressive features organized in three categories: URL (6 features), HTML content (4 features) and searching the web for information about that website (6 features), to classify websites. The authors have tested their proposed system over 8,118 phishing and 4,883 legitimate websites.

2011: Alkhozae and Batarfi introduced a phishing detection approach based on verifying and checking the website source code for phishing features out of W3C standards including https, images, suspicious URLs, domain, IFrame, script and popup windows [5]. Their method involved calculating the security percentage of a website based on the final security weight, which is obtained by decrementing the initial secure weight of a website in case a phishing characteristic is encountered within each line of the source code. Their proposed algorithm is quite primitive and includes simple computations. The authors, additionally, did not test their method on real data set; thus, the accuracy and success rate of their algorithm were not calculated and are questionable at the moment.

2011: Martin *et al.* [6] presented a revolutionary e-banking phishing website detection framework that makes use of Neural Network (NN) techniques based on 27 phishing features and indicators that can be grouped under 6 criteria: (i) URL and Domain Identity; (ii) Security Encryption; (iii) Source Code and JavaScript; (iv) Page Style and Contents; (v) Web Address Bar; and (vi) Social Human Factor. Their method involves initializing a NN with arbitrary weights and then training the network with a set of inputs from an archived data in the machine-understandable format. Their contribution is a novel algorithm based on the current-best-hypothesis for updating weights in a multi-layer NN, in which at each stage the output is checked, and weights are adjusted accordingly. They envisioned that this method would reduce the prediction error rate and offer better classification due to the parallel nature of the NN.

2011: Basnet *et al.* introduced a rule-based phishing detection approach inspired by an approach that involves monitoring networks by matching each observed packet against a set of rules [7]. They generated their rule set based on heuristic observations of various techniques and tricks used by phishers using machine learning features, which they applied on temporal data sets by using DT and Logistic Regression (LR) learning algorithms. They grouped their rules into the following categories: (a) *Search Engine-Based Rules*: using their crawling and indexing of webpages to check if the URL is listed or not (top 30 results of Google, Yahoo! and Bing); (b) *Red Flagged Keyword-Based Rule*: checking if any of the words in the URL is on the list of 62 frequently occurring words popular among phishers, which they had generated from their training data set; (c) *Obfuscation-Based Rules*: checking if certain characters such as '-', '_', '@', etc. which are commonly used by phishers to obfuscate URLs are present; (d) *Blacklist-Based Rule*: checking if the webpage is blacklisted by Google Safe Browsing API; (e) *Reputation-Based Rule*: checking if the URL is listed as a top phishing target or if its IP or domain are reported by PhishTank, StopBadware.org, etc.; and (f) *Content-Based Rules*: examining HTML contents of the webpage for password input field without using TLS/SSL or if it has more external than internal links or bad HTML markups, META tag and external domain that is on a blacklist, and IFrame with a URL that is on a blacklist. They pointed out that one of the main benefits of using a rule-based approach is that rules can be easily adapted when necessary to detect constantly changing phishing tactics. They tested their approach on more than 40,000 phishing and legitimate URLs, and achieved accuracy of 95-99%, with a false positive rate of 0.5-1.5%.

2011: Shahriar and Zulkernine analyzed and classified the existing phishing detection procedures based on the 5 most typical data sources: *whitelisted*, *blacklisted*, *hybrid*, *standalone*, and *random* [8]. Their goal was to perform a comprehensive analysis of these methods which would help address the issues such as unsuccessfulness of anti-phishing approaches to detect fresh phishing URLs; i.e. issues that remain unresolved in anti-phishing. Their study showed that whitelisted information-based approaches are beneficial when the data provided by the user is rather modest, but they require storing enormous quantities of information and they are not practical for dynamic webpages. On the other hand, blacklisted information-based approaches require timely fresh

information distributed across all parties for effective detection, which may not be the case in practice. However, hybrid approaches have much better performance and accuracy, since they combine the advantages of both approaches, but they also need to deal with the maintenance of white and blacklisted information. Approaches based on random information are not efficient in cases when phishing websites allow only a limited number of random credential submissions. To conclude, the results showed that combining different information sources offers better protection overall.

2012: Maurer and Höfer developed a method for phishing websites detection based on URL similarity [9]. Their method involves the extraction of four different URL terms and their validation using the search engines' capability of spelling correction and suggests querying of the original website's name in case of a suspicious query. They tested the proposed method on a big data set of 8,370 real phishing URLs. They obtained a result of 54.3% phishing URL detection accuracy; which led them to conclude that this method is not enough as a phishing detection mechanism on its own but should rather be used in combination with other methods. Furthermore, they pointed out that enhanced extraction of correctly spelled domain names at unexpected positions in the URL and a better brand name checker could also improve their method.

2012: Lakshmi and Vijaya developed a phishing website prediction system based on machine learning techniques [10], which employs website-identity/feature extraction from a website's URL and HTML code (a total of 17 features), and leverages third party services for example search engines, 'Whois' Lookup, and 'Blacklist' database of phishing and suspected websites. The learning model is created by training the features of both legitimate and phishing websites using Multi-Layer Perceptron (MLP), DT induction and Naïve Bayes (NB) classification supervised learning algorithms. By analyzing the evaluations of the models by employing 10-fold cross-validation and 2 performance-criteria (ease of learning, predictive accuracy), the DT classifier was revealed to give the best prediction results achieving up to 98.5% of prediction accuracy.

2012: Balamuralikrishna *et al.* developed an anti-phishing method established upon two stages: *URL Domain Identity* and *Image-Based Webpage Matching* [11]. They consider three webpage features: (i) text pieces and their style; (ii) images embedded in the page; (iii) the complete visual appearance of the page. For identifying URL domains and IP addresses they used the divide rule approach and approximate string-matching algorithm. If IP addresses are different then the suspected webpage's snapshot is passed on to the second stage for image matching. In the second stage, the corner detection method is used to calculate snapshot's salient points and Contrast Context Histogram (CCH) descriptor is used to extract their features which are then compared to those same features of the authorized webpage by computing the distance between their vectors. If the chosen threshold level is crossed during this matching, the suspected website is diagnosed as a phishing website. The authors claim that their method performs better than other existing tools but did not provide any testing results as evidence.

2013: Barraclough *et al.* developed a hybrid Neuro-Fuzzy phishing detection and prevention system for web transactions which employs both numeric and linguistic properties by combining a Fuzzy Logic that can deal with high-level reasoning and a Neural Network that handles raw data well [12]. Their main contribution is five inputs that completely represent phishing techniques and methods which allow for highly accurate detection of phishing sites in real-time. Those five inputs include: (i) *Legitimate site rules* (with 66 features); (ii) *User-behavior profile* (with 60 features); (iii) *PhishTank* (with 72 features); (iv) *User-specific sites* (with 48 features); and (v) *Pop-Ups from emails* (with 42 features). A total of 288 extracted features and 5 inputs are employed in the proposed model through 2-fold training and testing of cross-validations. The achieved results showed that the model's accuracy is high (98.5%) allowing for accurate distinction between phishing, suspicious and legitimate websites.

2013: Aburrous and Khelifi developed a phishing e-banking website detection method based on supervised machine learning which uses a fuzzy logic model with basic data mining associative classification algorithms to handle the phishing data features and patterns, for determining classification rules into the data miner [13]. Fuzzy reasoning provides the capability to determine imprecise and dynamic phishing features and to classify the phishing fuzzy rules. Their system extracts 27 phishing website features and patterns based on 6 criteria (URL and Domain Identity, Security and Encryption, Source Code and Java Script, Page Style and Contents, Web Address Bar, Social Human Factor). The proposal approves the phishing vulnerability based on particularized fuzzy data sets by cross-checking each extracted feature with corresponding fuzzy variables (low, moderate and high). They designed their system in three layers and used a removal (pruning) procedure to optimize the processing-time so that if a layer contains one high-value fuzzy input variable, controlling other features on the same layer is disregarded. To evaluate their approach, the authors designed a plug-in toolbar and tested it by using a representative example of 160 various online banking sites. They achieved the detection accuracy of 86% with very low false-positive alarms.

2015: Aydin and Baykal analyzed fraudulent URL`s features, subset-based feature selection techniques and machine learning algorithms for classifying phishing and legitimate websites [14]. As a first step, authors extracted the features about the URL of the pages and composed feature matrix. After creation of the feature matrix, CFS/Consistency subset-based feature selection techniques are employed to detect most prominent features. The number of properties is set to 17 and 25 for the CFS and Consistency subset-based feature selection techniques, respectively. As a next step after the having two matrixes which includes most prominent features, Naïve Bayes and Sequential Minimal Optimization (SMO) machine learning algorithms are used to classify websites. According to results, Naïve Bayes machine learning algorithm achieved the 88.17% accuracy. Besides that, the

SMO algorithm showed the best result with the 95.39% accuracy. The Consistency subset-based technique showed the weakest result in Naïve Bayes with the 83.69% accuracy. On the other hand, this technique showed its best result in SMO. The SMO revealed better results in both two subset-based techniques when compared with the Naïve Bayes method.

III. Research Methodology

This study is done for analyzing the performance of some specific machine learning classification algorithms on a given data set. After setting up the data set, we used data mining feature selection methods. Then arrange data set according to feature selection methods. After making sets we applied classification algorithms for evaluating the performance of each algorithm with used feature selection methodology.

A. Feature Selection Methods

If the data set is appropriate for machine learning, later the assignment of identifying regularities can be done simpler and faster by excluding features of data set which are not-relevant concerning the task to be acquired. This process is called **feature selection**. It often builds a model that generalizes better to unseen points. It can also increase the comprehensibility of resulting classifier models significantly.

Attribute-based feature selection approaches are used in this study. These techniques assess every feature individually and autonomously. In this study 'Gain Ratio' and 'ReliefF' attribute-based approaches are used.

B. Classification Algorithms

In this study, performances of 3 machine learning methods are analyzed as the basis in caparisoning the effects and compatibility of multiple feature selection methods. In the following subsection, an overview of all used classification algorithms is given.

B.1 Naïve Bayes

Naïve Bayes classifier runs on the simple yet relatively instinctive idea. In some instances, it further performs than other complex algorithms. It uses variables of a data sample, by observing them separately and independently.

B.2 J48 (Decision Tree)

The J48 decision tree builds a tree based on training set attributes that discriminate most clearly. This feature can show us more about the data occurrences in order that we can arrange the best of them to have the most eminent information gain.

B.3 SMO

SMO proposed by John C. Platt is an agile responding method for training the SVM, where SVM is a hyperplane that divides a set of positive samples from a set of negative ones with a maximized margin.

It performs great for big problems since it is training with a set volume that is greater and better than chunking.

IV. Implementation

To successfully detect a broad variety of fraudulent websites, we extract and analyze many of the features associated with them. We have composed feature group based on our analysis and several available literatures on detecting of phishing attacks. The purpose behind the using feature-based technique is to make the phishing attack detection technique as simple as possible.

A. Data Collection

As in our previous study [14] data set used in this study focuses on fraudulent URLs that are related to the most targeted websites. We specified the most targeted websites and their fraudulent URLs from the Phish Tank database (www.phishtank.com). Phish Tank website is managed by the OpenDNS platform which is a collaborative clearinghouse for the data on the Internet. Additionally, it provides information about detecting and preventing Phishing websites.

After determined the most targeted websites, we run the Google search engine to get real URLs regarding these websites. As a first step, we typed the company names on Google and found the legitimate link from the results as a website's real home page link. After having home page links of the most targeted websites we began to have more links by using web crawling method. In this study, we analyzed 8,538 URLs including 3,622 legitimate and 4,919 fraudulent ones.

B. Feature Extraction

The workflow of our performance analysis consists of several separate processes. The first process extracts properties related to the URL`s of the websites and creates feature matrix. At this stage, we classified the properties into five different group as shown in Fig. 1. To have textual properties we run software codes with using C# language at Microsoft Visual Studio. On the other hand, we performed some online processing works to get other properties from free web information providers. At this stage we coded R language scripts to have "whois record" and at the end the remaining data is collected by manual work. In consequence, we have 133 separate features related to the website URL`s which in our data set.

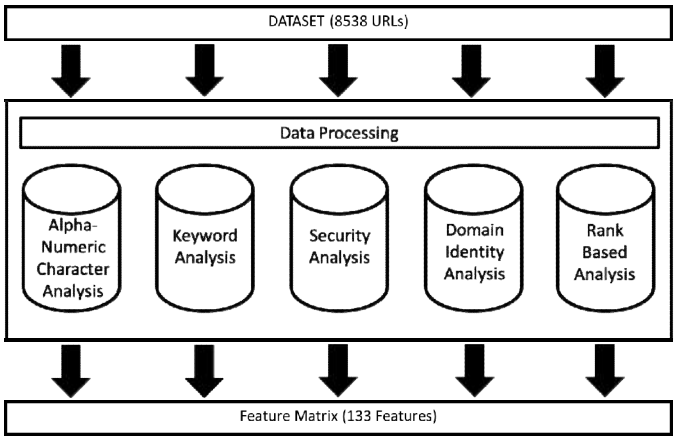

Fig. 1. Feature Extraction Categories.

C. Feature Selection

In this study, the second process uses attribute-based feature selection techniques to specify the most prominent properties. These techniques are used to minimize the feature matrix dimension by eliminating irrelevant and unnecessary properties.

We have comparatively evaluated Gain Ratio and ReliefF attribute-based feature selection techniques with their performance contribution to machine learning algorithms.

The two attribute-based feature selection techniques aforesaid above were individually run on the feature data set, which includes different attributes of 8,538 URLs. After using these techniques, we got new two feature matrixes with different numbers of properties. In this study, this step is analyzed by WEKA data mining and classification software tool.

D. Classification

After selecting prominent feature sets by using attribute-based techniques, the new two data sets were used as an input data to the machine learning algorithms to analyze the website`s if they are legitimate ones. In this study, we concentrate on three types of machine learning algorithms as a classification method, J48 (Decision Tree), Naïve Bayes and SMO (Sequential Minimal Optimization) as shown in Fig. 2. For each data set that we have these algorithms were run and then compared. To evaluate the classification algorithms the Precision, False Positive (FP) Rate, True Positive (TP) Rate and Overall Accuracy (ACC) were used. These three algorithms, as in our previous study [14], were evaluated by WEKA with default settings. As a same manner 10-fold cross-validation was used to divide the training and the test data set.

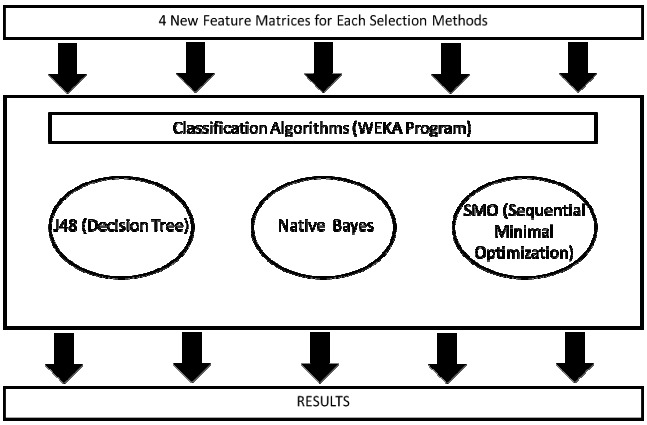

Fig. 2. Classification Algorithms

V. Results and Conclusions

The number of characteristics in the two feature selection techniques was mentioned as 36 for Gain Ratio Attribute and 58 for ReliefF Attribute, respectively. The feature selection techniques used in this study provided different results depending on the classification algorithm used. For example, SMO and J48 machine learning algorithms showed their best accuracy output values when they were used with the ReliefF attribute-based selection technique. On the other hand, the Naïve Bayes algorithm performed its' best result with the Gain Ratio Attribute feature selection technique in terms of accuracy.

The detailed outcomes of our analysis results are shown in Table 1. Naïve Bayes algorithm revealed its' best result with the 87.08% overall accuracy. The SMO method showed the best result as 96.42% and the J48 algorithm as 98.47%. This result is the highest ACC value obtained in the complete set of analysis.

TABLE I. The Results of the Analysis

Feature Selection Methods	Classification Algorithm	Overall Accuracy	TP Rate	FP Rate	Precision
Gain Ratio Attribute (36)	Naive Bayes	87,08%	0,871	0,100	0,894
	J48	97,18%	0,972	0,030	0,972
	SMO	95,95%	0,960	0,049	0,960
ReliefF Attribute (58)	Naive Bayes	81,99%	0,820	0,136	0,868
	J48	98,47%	0,985	0,015	0,985
	SMO	96,42%	0,964	0,043	0,965

The ReliefF attribute-based technique showed the weakest result in Naïve Bayes with the lowest ACC. There against, this technique shows its best result in J48. The J48 algorithm exhibited the best result in both techniques. Moreover, the SMO algorithm was the second-best performing classification algorithm. The Naïve Bayes algorithm has found as being the worse when it is compared to the others, as can be seen in Fig. 3.

The results obtained by the evaluation of the three algorithms showed that the J48 and SMO algorithms might be used to detect fraudulent websites based on URL features. On the other hand, the results point that the Naïve Bayes performs poorly and should not be employed in the classification analysis shown this study.

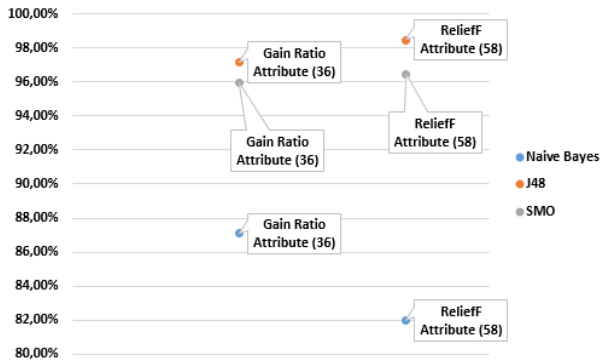

Fig. 3. Graph with 3 classification algorithms

We anticipate that the results of this study might ensure the different point of view to future studies in detection of fraudulent websites. As a future work, we will continue working on other machine learning algorithms with attribute-based and subset-based feature selection techniques. And, we are going to test and compare results in order to find the best performing one in terms of our evaluation criteria. In the end, we are going to test the validation process with unknown data sets to ensure the success of the proposed algorithms.

APPENDIX

The list of abbreviations used in this manuscript is as follows:

LIST OF ABBREVIATIONS

ACC : Overall Accuracy
CAR : Cumulative Abnormal Return
CCH : Contrast Context Histogram
DOM: Document Object Model
DM : Data Mining
DT : Decision Tree
FP : False Positive
LR : Logistic Regression
PII : Personal Identification Information
MLP : Multi-Layer Perceptron
NB : Naïve Bayes
NN : Neural Network
SVM : Support Vector Machines
TP : True Positive
TSVM: Transductive SVM
WEKA: Waikato Environment for Knowledge Analysis

ACKNOWLEDGEMENTS

This research has been partially supported by the Swedish Civil Contingencies Agency (MSB) through the projects RICS, by the EU Horizon 2020 Framework Programme under grant agreement 773717, and by the STINT grant IB2019-8185.

REFERENCES

[1] Butun, Ismail. "Privacy and trust relations in internet of things from the user point of view." 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2017.

[2] P. Fortuna, N. Pereira, and I. Butun, "A framework for web application integrity," 4th International Conference on Information Systems Security and Privacy, ICISSP, Madeira, Portugal, 22-24 January 2018, SciTePress, 2018.

[3] Aburrous, Maher, M. Alamgir Hossain, Keshav Dahal, and Fadi Thabtah. "Intelligent phishing detection system for e-banking using fuzzy data mining." Expert systems with applications 37, no. 12 (2010): 7913-7921.

[4] Xiang, Guang, Jason Hong, Carolyn P. Rose, and Lorrie Cranor. "Cantina+: A feature-rich machine learning framework for detecting phishing web sites." ACM Transactions on Information and System Security (TISSEC) 14, no. 2 (2011): 21.

[5] M. G. Alkhozae, and O. A. Batarfi, (2011). Phishing websites detection based on phishing characteristics in the webpage source code. International Journal of Information and Communication Technology Research, 1(6).

[6] Martin, A., Anutthamaa, N., Sathyavathy, M., Francois, M. M. S., & Venkatesan, D. V. P. (2011). A framework for predicting phishing websites using neural networks. arXiv:1109.1074.

[7] Basnet, Ram B., Andrew H. Sung, and Quingzhong Liu. "Rule-based phishing attack detection." In Proceedings of the International Conference on Security and Management (SAM), WorldComp, 2011.

[8] Shahriar, Hossain, and Mohammad Zulkernine. "Information source-based classification of automatic phishing website detectors." In 2011 IEEE/IPSJ International Symposium on Applications and the Internet, pp. 190-195. IEEE, 2011.

[9] Maurer, M. E., and Höfer, L. (2012). Sophisticated phishers make more spelling mistakes: using URL similarity against phishing. In Cyberspace Safety and Security (pp. 414-426). Springer, Berlin, Heidelberg.

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

[11] Balamuralikrishna T., Balamuralikrishna, N. Raghavendrasai and M. Satya Sukumar, (2012) "Mitigating online fraud by ant phishing model with URL and image-based webpage matching," International Journal of Scientific and Engineering Research (IJSER), March, 2012.

[12] Barraclough, Phoebe A., M. Alamgir Hossain, M. A. Tahir, Graham Sexton, and Nauman Aslam. "Intelligent phishing detection and protection scheme for online transactions." Expert Systems with Applications 40, no. 11 (2013): 4697-4706.

[13] Aburrous, Maher, and Adel Khelifi. "Phishing detection plug-in toolbar using intelligent Fuzzy-classification mining techniques." in the international conference on SCSE, San Francisco, California, USA. 2013.

[14] M. Aydin, and N. Baykal, "Feature Extraction and Classification Phishing Websites Based on URL", IEEE Conference on Communications and Network Security (CNS), 2015.