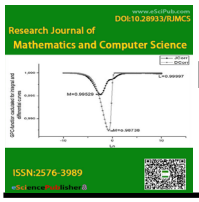
Research Article **RJMCS (2020) 4:20**



Research Journal of Mathematics and Computer Science (ISSN:2576-3989)

Phishing Detection Model for Emails Using Classification Algorithm

Adekunle Yinka, *Olu-Oshadare Olumide

Department of Computer Science, Babcock University, Nigeria

ABSTRACT

Anti-Phishing Working Group (APWG) is a contributing member that report, and study the ever-evolving nature and techniques of cybercrime. The APWG tracks the number of unique phishing emails and web sites, a primary measure of phishing across the globe. A single phishing site may be advertised as thousands of customized features, all leading to basically the same attack destination.

This work aims to design a machine learning model using a hybrid of two classification algorithms which include Random Forests and Support Vector Machine (SVM). Also perform feature selection on the obtained phishing dataset to select a subset of highly predictive features and evaluate the model against other classification algorithms and existing solutions with the following metrics: False Positive Rate (FPR), Accuracy, Area Under the Receiver Operating Characteristic Curve (AUCROC) and Weighted Averages. It is expected that upon evaluation of this model much improved efficiency would be recorded as against other existing models.

Keywords: Phishing Detection Model, Emails, Classification Algorithm

***Correspondence to Author:**
Olu-Oshadare Olumide
Department of Computer Science, Babcock University, Nigeria

How to cite this article:
Adekunle Yinka, Olu-Oshadare Olumide. Phishing Detection Model for Emails Using Classification Algorithm. Research Journal of Mathematics and Computer Science, 2020; 4:20 .

eSciPub LLC, Houston, TX USA.
Website: https://escipub.com/

Background

The APWG report showed that the number of URLs which were used to host phishing attacks has increased from 164,023 in the first quarter of 2012 to 175,229 in the second quarter of the same year. The first quarter through the third quarter of 2015, the APWG received reports of 630,494 unique phishing emails detected from the first quarter through the third quarter of 2015. The worldwide infection rate was 36.511% in the first quarter, 32.211% in the second quarter, and 32.122% third quarter of 2015 respectively (Tahir, et al., 2016). Pujara and Chaudhari (2017) noted that the total number of phishes detected in the second quarter (2Q) of 2018 according to the APWG was 233,040, compared to 263,538 in the first quarter (1Q) of 2018. These totals exceed the 180,577 observed in the fourth quarter (4Q) of 2017 and the 190,942 seen in the third quarter (3Q) of 2017. There were increases in webmail targeted sector with 21% of overall phishing attacks. Payment sector is continuing as the most attractive target for phishing. The total sum of phish identified in the 1Q of 2018 was 263,538. This was up 46 per cent from the 180,577 which was observed in the 4Q of 2017. It was also significantly more than the 190,942 seen in 3Q 2017. The number of unique phishing reports submitted to APWG during 1Q 2018 was 262,704, contrasted with 233,613 in 4Q 2017 and 296,208 in 3Q 2017 (Pujara & Chaudhari, 2017).

Table 1.1 Statistical Highlights for 3rd Quarter of 2019 Source: (APWG, 2019)

	July	August	September
Number of unique phishing Web sites detected	93,194	86,908	86,276
Number of unique phishing e-mail reports (campaigns) received by APWG from consumers	35,530	40,457	42,273
Number of brands targeted by phishing campaigns	444	414	425

The total number of phishing emails detected by APWG in the third quarter of 2019 was 266,387. This was up 46 percent from the 182,465 seen in Q2, and almost double the 138,328 seen in Q4 2018. “This is the worst period for phishing that the APWG has seen in three years, since the fourth quarter of 2016,” said Greg Aaron, APWG Senior Research Fellow and President of Illumintel Inc. The APWG recorded 277,693 attacks in the fourth quarter of 2016. In addition to the increase in phishing volume, the number of brands that were attacked by phishers in Q3 was also up notably. APWG contributor MarkMonitor saw attacks against more than 400 different brands (companies) per month in Q3, versus an average of 313 per month in Q2.

High false-positive rates are problems that occur in phishing detection which reduces the usability of a proposed solution. To solve the above problem, researchers have posited Machine learning and image comparison models in recent times. Machine learning employs robust classification algorithms to reduce high false-positive rates. However, many of these solutions present a new problem;

- time complexity and

- Implementation complexity.

Many machine learning approaches are very difficult to deploy, thus maintenance becomes a problem and cannot function efficiently in time-bound systems, because they use up to 50% more time in classifying a mail. This work consequently aims at proposing a phishing detection model based on feature selection and hybrid machine learning methods that achieves a low false-positive rate and lower time and space complexity as compared with existing solutions. The model is also designed to lessen complex relative to existing models which will aid its deployment and maintainability.

Jain and Gupta (2017) noted that detection and prevention from phishing attacks is a big challenge to scientists and researchers because attackers perform these attacks in such a way that existing anti-phishing techniques are bypassed and even an educated and experienced user may fall victim of these attacks. The following are detection methods through which phishing has been curbed:

- blacklist,
- heuristic,
- visual similarity and
- Machine learning

Literature Review

Over the past few years, given the rapid pace of embracing technological advancements, computer security attacks have significantly increased. Phishing is one of the attacks which is of a big concern to internet users. Phishing is a type of Social Engineering (SE) attack that targets at abusing or exploiting the weaknesses or shortcoming found in the system at the client's side. For instance, a system might be secure enough for password robbery however the uninformed client may release his/her password when the attacker sends a bogus update on his/her password soliciting that an update is made through forged (phished) email. The false update may come via an email which prompts the user to click a Uniform Resource Locator (URL) that leads the user to an illegitimate site created to gather information from the unaware user or through a pop-up while browsing, calling the user's action to what may seem important like a system/application update.

Tahir et al. (2016) proposed a supervised machine learning approach using seven algorithms individually and combining them to form a hybrid model in other to detect phishes. The best performing experiments were two hybrid models of RF + KNN (IBk on WEKA) and J48+KNN with accuracy of 97.75% each. The model consumed time during training as feature selection algorithms weren't used to remove redundant and duplicate features.

Subasi et al. (2017) proposed an intelligent phishing website detection system using the RF algorithm and compared with the performance of other algorithms. The intelligent system had an accuracy of 97.36%. Existing systems performed better than the intelligent system in terms of accuracy.

Ali (2017) proposed "a method for phishing website detection based on machine learning classifiers (back-propagation neural network (BPNN), naïve Bayes, c4.5 decision tree, support vector machine, random forest and k-nearest neighbour) with a wrapper features selection. The author compared it to the results of the selected classifiers using information gain and principal component analysis feature selection algorithms". The best performing classifier with wrapper feature selection according to the author is a random forest. However, the wrapper feature section consumed more time and computational power.

Sönmez et al. (2018) proposed a phishing website feature classification based on extreme learning machine. The system had an accuracy of 95.34% compared to SVM and NB which had 93.80% and 92.98% respectively. However, existing systems performed better in terms of accuracy.

Patil et al. (2018) proposed a system using three approaches where the features of the URL are first analyzed then the legitimacy of the websites are checked through the hosts and finally the genuineness of the site is checked using visual appearance-based analysis. Machine learning algorithms were used for the evaluation of different features of URL and websites. The system faulted in detecting false positive and false negatives.

Zhu, Chen, Ye, Li and Liu (2019) developed an effective phishing attack detection model based on optimal feature selection and neural network. The model achieved an accuracy of 99.3%. The authors used fewer features compared to existing frameworks.

Dataset Description

The dataset opted for this research is the phishing website dataset from the UCI machine learning repository. It contains 2456 instances with 30 attributes based on the address bar, hypertext mark-up language, JavaScript and domain features. The dataset was donated to the UCI library on the 26th of March 2015 based on the scarcity of a standard dataset for training models (Mohammad, McCluskey, & Thabtah, 2015).

This dataset will be used for training the model while the dataset donated by (Abdelhamid, 2016) will be used to test the trained model. The dataset has 1353 instances with 10 attributes. The phishing emails were collected from Phish tank data archive (www.phishtank.com), a free community site where users can submit, verify, track and share phishing data. The legitimate emails were collected from Yahoo and starting point directories using a web script developed in PHP. The PHP script was plugged with a browser and we collected 548 legitimate emails out of 1353 emails.

There are 702 phishing URLs and 103 suspicious URLs. When a website is considered SUSPICIOUS that means it can be either phishy or legitimate, meaning the website held some legit and phishy features (Abdelhamid, 2016).

The dataset for the study is obtained from the UCI Machine learning Repository, it has instances with attributes including the class attribute. The following steps will be taken to pre-process the dataset for analysis.

1. Find and deal with missing values discovered; either by dropping affected instances and/or features, inputting averages or median values or using null or zero value.
2. Identify and correct misformatted input values, for example, where strings are present in a numerical feature.
3. Identify features that may require normalization using the z-score, this applies to features that are larger numerically than others and can skew the results of the analysis.
4. Finding categorical values that are too granular and hold no real analytical value, for example, names of websites. These categorical features can either be removed or binned with specific criteria such as website names containing ‘sales.

Feature Selection: the phishing dataset contains features and class attributes, these features can be further categorized as Address Bar features, abnormal entries feature and Domain name

Features. Feature selection will be conducted with the following steps

1. Use the Pearson correlation to determine how much the features correlate to the class attribute.
2. Discover features that correlate with other features and remove them to avoid redundancy.
3. Rank the features in the data set using Information Gain, to discover the features which contain the most information or entropy.
4. Following the aim of reduced complexity (space, time) this work proposes, identify a high-performance subset of features for analysis using Gain ratio and Decision Tree.

Implement and Evaluate the Model: All analysis of the data set after feature selection will be carried out in WEKA, a data mining tool, which can also be used to evaluate several other models. This steps to be followed are:

1. Divide the data set into a test and train sets, with the ration 30% to 70%.
2. Fit the proposed model to the train set using 10 fold cross-validation, examine the performance of the model with evaluation metrics such as FP rate, Accuracy, Time spent training, Time spent testing, TP rate, precision, recall, f-measure, MCC, ROC Area (area under the curve of the receiver operating characteristic curve), PRC Area (Area under the Precision-Recall Curve) and the class (normal or anomaly).
3. Iteratively adjust the hyperparameters of the model to improve the performance of the model.
4. Fit other classification and hybrid algorithms to the train data set and compare performance.
5. Test proposed model on the aforementioned Test dataset, test existing work, classification and hybrid algorithms on the Test dataset, evaluate and report performance.
6. Draw Inferences and Conclusions from the reports

Framework

Several frameworks have been proposed and developed by different researchers for detecting phishing emails. Methods adopted are blacklist, heuristics, visual similarity and machine learning. Researchers who adopted the machine learning approach has proposed different types of models which are relatively efficient. Machine learning, through research, has proven to be efficient in detecting phishing emails. Many of these existing models were developed using single classifiers, a combination of classifiers and ensemble techniques. Some researchers opted for feature selection while others neglected it. Feature selection proved to be efficient as it selected only those feature that are key for the classification of instances. These models were evaluated using the classification accuracy as well as the true positive rate (detection rate) and false-positive rate. Few of the researchers noted the performance of their models in terms of classification time without expressly stating how long it took. Time taken for classification of instances is essential as implementation of models proposed will be done in real-time. A model that consumes a lot of time in identifying phishing emails will not be as efficient as the one that consumes less time. Lower times consumption helps in reducing

delays in the entire detection and prevention of phishing emails.

Experiment Design

The experiment will be conducted using the Weka environment. Pre-processing the training dataset involves the checking and cleaning of the dataset. Sequel to the pre-processing stage will be the feature selection stage where information gain and decision tree will be used to select features substantial enough to classify the instances in the dataset. At the classification stage, the training dataset will be imported to Weka for training the algorithms. Here, the algorithms classify the instances either as a legitimate site or a phishing site. The classifiers will work on the training dataset, identifying patterns and understanding the dataset in other to develop a model after which the tests dataset will be supplied in other to test the model built. At the end of the classification, Weka will evaluate the performance of the algorithms on the test dataset and output the evaluation using the following metrics: TP rate, FP rate, precision, recall, f-measure, MCC, ROC Area (area under the curve of the receiver operating characteristic curve), PRC Area (Area under the Precision-Recall Curve) and the class (normal or anomaly). This evaluation is carried out in relation to the classes in the dataset and a weighted average of the two classes is also outputted. A wide set of experiment will be conducted using single classifiers and ensemble methods in other to effectively compare the performance of the proposed model.

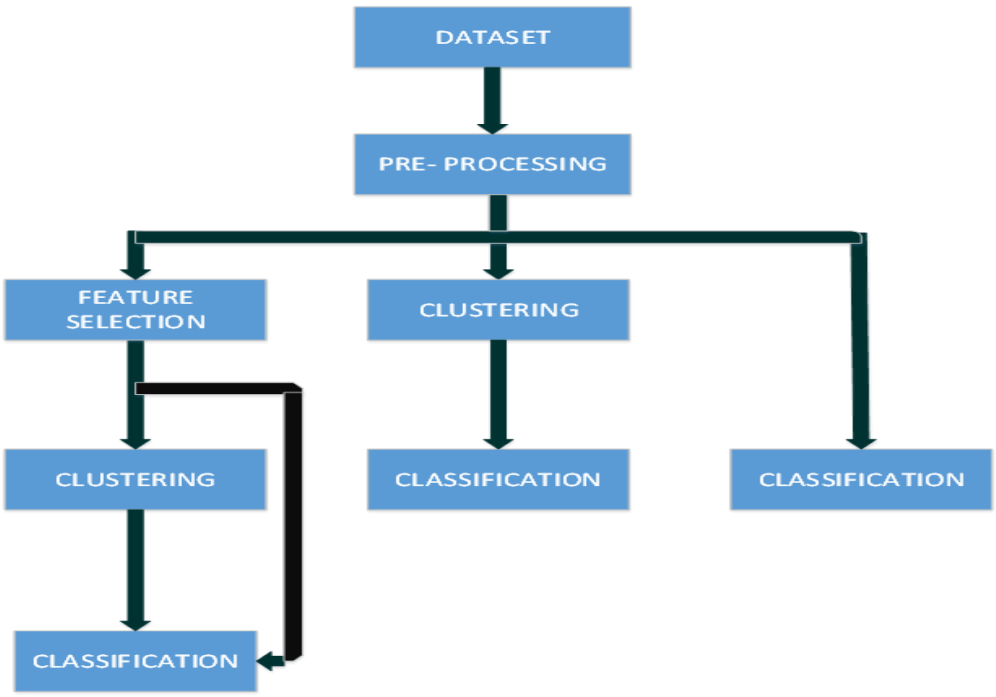


Figure1 Framework

The hardware specifications of the system used are:

- intel core i7-3610Q
- 10gb ram
- Nvidia K1000M
- 320gb hard disk

while the software specifications are:

- windows 10 operating system
- Weka version 3.8.3

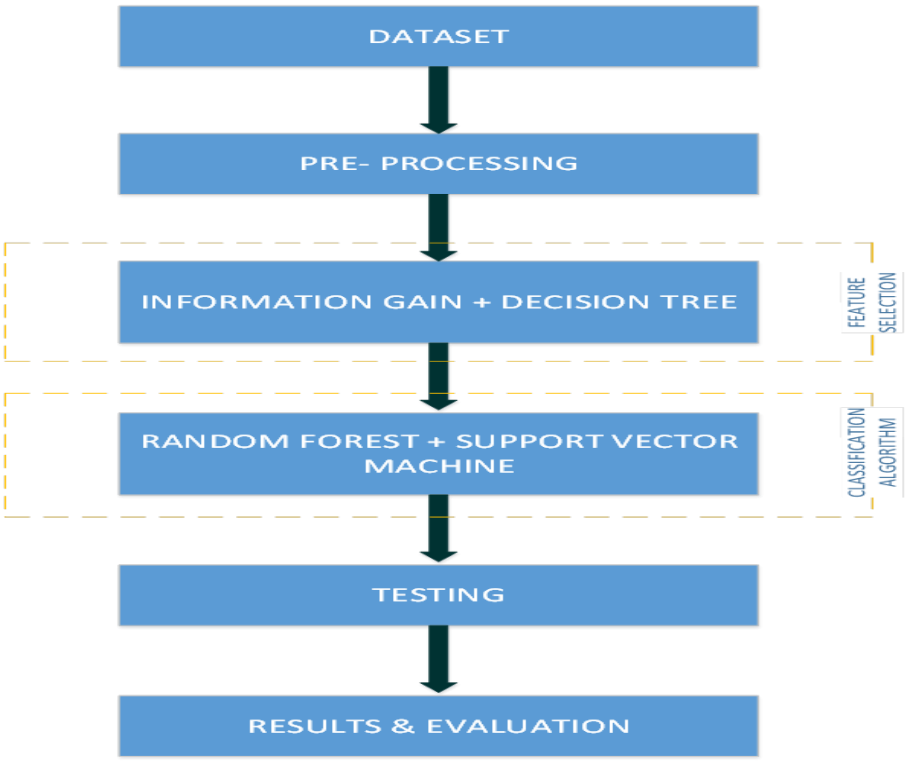


Figure 2 presents the proposed schematic model.

Justification of Tool and Methods

The Waikato environment for knowledge analysis (Weka) was developed by the University of Waikato in New Zealand. The goal was to identify hidden patterns which lead to information from raw data gathered in the agricultural domain. Weka performs machine learning processes such as pre-processing, feature selection, clustering, association rules, visualization and classification. It also infers valuable data as patterns and trends. The version adopted for this research is the 3.8.3 version which is the latest version of the application. Weka is user-friendly and is available for free under the GNU General Public License Agreement.

Support Vector Machine (SVM): is one of the most well-known and robust supervised machine learning techniques, which has been utilized effectively in many science and engineering applications. SVM is based on maximizing the margin and thereby creating the largest possible distance between the separating hyperplane and the instances to reduce an upper bound on the expected generalization error. Some instances of the training dataset called support vectors, which are close to the separating hyperplane and provide the most useful information for classification, are utilized in SVM training. In addition, an appropriate kernel function is used to transform the data into a high-dimension to use linear discriminate functions.

Random Forest (RF): is a tree predictor classifier, where every tree relies on upon the estimations of an irregular vector inspected autonomously can be used for both classification and regression. It is an ensemble of a number of decision trees independently trained on selected training datasets. The classification information is then determined by voting among all the trained decision trees. Random Forest achieves a better classification accuracy compared to a single tree. Based on the

literature, the random forest is an excellent classifier in the phishing domain.

Feature Selection (FS): is the process of selecting the highest number of features as a subset of the features found in the training datasets. Feature selection achieves two main goals. First, it makes the training datasets applied to a classifier more efficient by eliminating a number of features. Second, it often generates a classification accuracy rate that is comparable to that generated by a classifier without feature selection if the feature selection method eliminates only the irrelevant features (Hadi, Aburub, & Alhawari, 2016). Information gain and decision tree will be used in the elimination of redundant features and select the features substantial enough to classify the instances in the dataset.

Evaluation metrics

The following metrics would be used in evaluating the performance of the proposed framework:

- Accuracy
- Area Under the Curve Receiving Operating Characteristic
- Confusion matrix
    - True Positive
    - True Negative
    - False Positive
    - False Negative
- Classification time
    - Training
    - Testing
- Weighted average

These performance metrics were selected on the basis of evaluating the model's performance in solving the stated problems and also compare with existing frameworks. The accuracy metric is common in the works reviewed and it would be easy to compare this model with theirs using the accuracy metric

Conclusion

In this new age of technological advancement and the ever increasing usability of the internet as a whole, phishing attacks are eventually becoming inevitable. In this research work a machine learning model was developed using two classification algorithms which are Random Forests and Support Vector Machine (SVM). Also feature selection on the obtained phishing dataset was performed in order to obtain highly predictive features. Upon evaluation of these model, a significant increase in efficiency is expected to be recorded when compared with other existing models in terms of false positive rate (fpr), accuracy, area under the receiver operating characteristic curve (AUCROC) and weighted averages.

REFERENCES

1. Abdelhamid, N. (2016). *Website Phishing Data Set*. Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Website+Phishing#
2. Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). Phishing detection based Associative Classification data mining. *Expert Systems with Applications*, 5948–5959. doi:10.1016/j.eswa.2014.03.019
3. Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of Phishing Email Using Random Forest Machine Learning Technique. *Journal of Applied Mathematics*. doi:10.1155/2014/425731
4. Al-Daeef, M. M., Basir, N., & Saudi, M. M. (2014). A method to Measure the Efficiency of Phishing Emails Detection Features. *IEEE*.
5. Ali, W. (2017). Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection. *International Journal of*

Advanced Computer Science and Applications, 8(9), 72-78.

6. Anand, K. (2019). *Application Security: Social Engineering Attack.* Retrieved from Imperva: https://www.imperva.com/learn/application-security/social-engineering-attack/
7. Apuke, O. D. (2017). quantitative research methods a synopsis approach. *Arabian Journal of Business and Management Review (Kuwait Chapter), 6*(10), 40-47. doi:10.12816/0040336
8. Aroyo, A. M., Rea, F., Sandini, G., & Sciutti, A. (2018). Trust and Social Engineering in Human Robot Interaction: Will a Robot Make You Disclose Sensitive Information, Conform to Its Recommendations or Gamble? *IEEE, 3*(4), 3701-3708. doi:10.1109/LRA.2018.2856272
9. Barraclough, P., Hossain, M., Tahir, M., Sexton, G., & Aslam, N. (2013). Intelligent phishing detection and protection scheme for online transactions. *Expert Systems with Applications*, 4697–4706. doi:10.1016/j.eswa.2013.02.009
10. Buber, E., Demir, Ö., & Sahingoz, O. K. (2017). Feature Selections for the Machine Learning based Detection of Phishing Websites. *IEEE*.
11. Dakpa, T., & Augustine, P. (2017). Study of Phishing Attacks and Preventions. *International Journal of Computer Applications, 163*(2), 5-8.
12. Diana, K., & Seema, K. (2019). *Microsoft Security Intelligence Report.* Retrieved from Spear phishing campaigns—they're sharper than you think: https://www.microsoft.com/security/blog/2019/12/02/spear-phishing-campaigns-sharper-than-you-think/
13. Gupta, B. B., Arachchilage, N. A., & Psannis, K. E. (2017). Defending against phishing attacks: taxonomy of methods, current issues and future directions. *Springer*. doi:10.1007/s11235-017-0334-z
14. Hadi, W., Aburub, F., & Alhawari, S. (2016). A new fast associative classification algorithm for detecting phishing websites. *Applied Soft Computing*, 729-734. doi:10.1016/j.asoc.2016.08.005
15. Hadi, W., Aburub, F., & Alhawari, S. (2016). A new fast associative classification algorithm for detecting phishing websites. *Applied Soft Computing*, 729-734. doi:10.1016/j.asoc.2016.08.005
16. Haggag, M. H., Mohammed, E. H., & El-Rahmany, M. S. (2017). Social Engineering Attacks Detection Techniques: Survey Study. *International Journal Of Engineering And Computer Science, 5*(12). doi:10.18535/ijecs/v5i12.84
17. Islam, M., & Chowdhury, N. K. (2016). Phishing Websites Detection Using Machine Learning Based Classification Techniques. *international conference on advanced information and communication technologY.* Chittagong.
18. Islam, R., & Abawajy, J. (2013). A multi-tierphishingdetectionandfilteringapproach. *Journal ofNetworkandComputerApplications*, 324–335.
19. Jain, A. K., & Gupta, B. B. (2017). Comparative Analysis of Features Based Machine Learning Approaches for Phishing Detection. *IEEE*, 2125-2130.
20. Jain, A. K., & Gupta, B. B. (2017). Phishing Detection: Analysis of Visual Similarity Based Approaches. *Security and Communication Networks*, 1-21. doi:10.1155/2017/5421046
21. James, J., Sandhya, L., & Thomas, C. (2013). Detection of Phishing URLs Using Machine Learning Techniques. *International Conference on Control Communication and Computing (ICCC).* Thiruvananthapuram, India: IEEE. doi:10.1109/ICCC.2013.6731669

22. Kadam, A. S., & Pawar, S. S. (2013). comparison of association rule mining with pruning and adaptive technique for classification of phishing dataset. *IET*.

23. Kalnins, R., Purins, J., & Alksnis, G. (2017). Security evaluation of wireless network access points. *Applied Computer Systems*, 38-45. doi:https://doi.org/10.1515/acss-2017-0005

24. Khonji, M., Iraqi, Y., & Jones, A. (2013). Phishing Detection: A Literature Survey. *IEEE*. doi:10.1109/SURV.2013.032213.00009

25. Koyun, A., & Janabi, E. A. (2017). Social Engineering Attacks. *Journal of Multidisciplinary Engineering Science and Technology, 4*(6), 7533-7538.

26. Mohammad, R. M., McCluskey, L., & Thabtah, F. (2015, 03 26). *Phishing Websites Data Set*. Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites

27. Nguyen, H. H., & Nguyen, D. T. (2016). Machine Learning Based Phishing Web Site Detection. *Springer*.

28. Patil, P., & Devale, P. (2016). A Literature Survey of Phishing Attack Technique. *International Journal of Advanced Research in Computer and Communication Engineering, 5*(4), 198-200. doi:10.17148/IJARCCE.2016.5450

29. Patil, V., Thakkar, P., Shah, C., Bhat, T., & Godse, S. P. (2018). Detection and Prevention of Phishing Websites using Machine Learning Approach. *Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).* IEEE.

30. Patil, V., Thakkar, P., Shah, C., Bhat, T., & Godse, S. P. (2018). Detection and Prevention of Phishing Websites using Machine Learning Approach. *IEEE*.

31. Pokrovskaia, N. N., & Snisarenko, S. O. (2017). Social engineering and digital technologies for the security of the social capital' development. *International Conference of Quality Management, Transport and Information Security* (pp. 16-19). Petersburg, Russia: IEEE. doi:10.1109/ITMQIS.2017.8085750

32. Pujara, P., & Chaudhari, M. (2017). Phishing Website Detection using Machine Learning : A Review. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 3*(7), 395-399.

33. Pujara, P., & Chaudhari, M. B. (2018). Phishing Website Detection using Machine Learning : A Review. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 3*(7), 395-399.

34. Rathod, S. B., & Pattewar, T. M. (2015). A Comparative Performance Evaluation of Content Based Spam and Malicious URL Detection in E-mail. *International Conference on Computer Graphics, Vision and Information Security (CGVIS).* IEEE.

35. Salahdine, F., & Kaabouch, N. (2019). Social Engineering Attacks: A Survey. *Future Internet*, 1-17. doi:10.3390/fi11040089

36. Shaikh, A. N., Shabut, A. M., & Hossain, M. (2016). A Literature Review on Phishing Crime, Prevention Review and Investigation of Gaps. *International Conference on Software, Knowledge, Information Management & Applications* (pp. 9-15). IEEE.

37. Sönmez, Y., Tuncer, T., Gökal, H., & Avci, E. (2018). Phishing Web Sites Features Classification Based on Extreme Learning Machine. *IEEE*.

38. Tahir, Amaad, Haq, Asghar, Sohail, Zafar, . . . Saira. (2016). A Hybrid Model to Detect Phishing-Sites using Supervised Learning Algorithms. *International Conference on Computational*

Science and Computational Intelligence (pp. 1126-1133). IEEE.

39. Tahir, M. A., Asghar, S., Zafar, A., & Gillani, S. (2016). A Hybrid Model to Detect Phishing-Sites using Supervised Learning Algorithms. *International Conference on Computational Science and Computational Intelligence.* IEEE. doi:10.1109/CSCI.2016.213
40. Varshney, G., Misra, M., & Atrey, P. K. (2016). A survey and classification of web phishing detection schemes. *SECURITY AND COMMUNICATION NETWORKS*, 6266–6284. doi:10.1002/sec.1674
41. Zhu, E., Chen, Y., Ye, C., Li, X., & Liu, F. (2019). OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network. *IEEE*, 73271 - 73284. doi:10.1109/ACCESS.2019.2920655