DEVELOPMENT OF A RULE-BASED PHISHING WEBSITE CLASSIFICATION SYSTEM

Kudirat O. Jimoh[1], Adepeju A. Adigun[2], Ibrahim K. Ogundoyin[3], Babatunde M. Eniola[4]
Department of Information and Communication Technology, Osun State University, Nigeria
[1]kudirat.jimoh@uniosun.edu.ng, [2]fempej2013@gmail.com, [3]ibraheem.ogundoyin@uniosun.edu.ng, [4]babshenny@gmail.com

Keywords: Rule-based, Phishing, Uniform Resource Locator, Classification, Random Forest

Abstract: *Phishing websites are a kind of internet security problem that focuses on human vulnerabilities in contrast with software vulnerabilities. They are poisonous websites that look like legitimate websites to steal user's identities like passwords and financial information. The main objective of this study is to develop a rule-based phishing website classification system to detect and classify the website into phishing and non-phishing. The specific objectives are to determine and examine the specific features for classification, design and implement the model, and to evaluate the performance of the model. Samples of phishing data were collected by documenting and evaluating different behaviors of Phishing Website to train the system from URLhaus, Openphish, and Moz Trusted URL, the model was implemented using Random Forest in Python Programming Language environment, and the performance evaluation was done using sensitivity, specificity, and accuracy as metrics. The developed system has successfully identified and analyzed different URL features using a rule-based model with an accuracy of 81.6%, sensitivity of 78.4%, and specificity of 84.4%. Thereby, reducing cybercrime and improving the security level on the internet.*

1. INTRODUCTION

Due to the developing utilization of smart cell phones, large numerous individuals are relying upon online services for the payment of bills, to carry out various financial transactions, and even interact with family and friends, (both known and unknown). Most commercial and government establishments have also introduced more internet services to customers [1][2]. While such exercises importantly affected the worldwide economy, such expansive reliance on internet services builds security risks concerning customers alongside financial institutions. Phishing is a falsified practice that utilizes social engineering with specialized deception tricks to obtain a customer's account credentials and identity [3]. Phishing is a newly introduced method of stealing personal identity and most of the people involved in cybercrime are youths whose wants are insatiable [4]. Phishing is among the most widely perpetrated forms of cyber-attack used to collect susceptive details such as bank account and credit card numbers, user logins details and passwords, including other details entered through the website. [5]. Also, these Phishing crimes also include computer intrusions designed to obtain information of the most sensitive sort-such as credit cards, companies, trade secrets, or individual's private information. [6]. These phishing crimes not only affect our financial well-being and our privacy, they also threaten every nation's critical infrastructure, banking system, the stock market, the electricity and water supply, telecommunication networks, and critical government services all rely on computer networks [7].

The main objective of phishing websites is to fraudulently have access to users' data via surfing and visiting a forged website page that resembles a genuine site of a true bank or organization and requests that the victim enter individual data [8][9]. For example, their financial account number, personal identification number, credit card data and, so on. This results in the accessibility of data safety through the tradeoff of confidential information and the internet customers may end up suffering the loss of cash or some of his or her very valuable assets. More recently, phishers take advantage of the Coronavirus pandemic (COVID-19) to fool their prey [10]. In this study, a rule-based phishing website classification system was developed. This is to identify phishing webpage to curtail this criminal act. The rule-based system uses different features of phishing URLs such as the presence of @, URL containing IP address, and so on. However, Phishing on Websites has to be controlled in other to save guard users from online crimes and to protect individuals and organizations from cybercriminals, hackers, and other bad actors who want to disrupt and steal personal or classified data for personal gain.

However, the literature on the analysis of URL features for classification using rule-based has not been fully considered. Hence, the development of a rule-based phishing website classification system was considered. The system identified and analyzed sixteen features for the classification process.

The rest of the paper is organized as follows: related works is discussed in Section 2, materials and methodology are discussed in Section 3. In section 4 the results and discussion were presented and finally, section 5 concludes the paper.

2. Related Works

Several detection models for the malicious website have been developed and well documented. [11] developed an artificial neural network phishing detection model based on a Decision Tree and Optimal Features, the phishing detection method was created using two feature sets, for webpage identification. A support vector machine (SVM) algorithm was employed and the hidden knowledge from the proposed SVM model was 19 adopted to make the proposed method more functional and easier to use web page resource elements. The features used lack dependency on 3rd-party services such as search engines result and web browser history. [12] developed an intelligent rule-based phishing websites classification system based on extreme learning machine and a specific web page features set to prevent phishing attacks was created. The model used machine learning techniques to detect phishing web pages. This approach was quite effective; however, it only dealt with specific features only and the machine was not fully trained.

[13] improved Spoofed website detection using the machine learning method. It brings out a diverse set of robust features categorized into three categories, i.e., web page, URL and HTML-based features. These features are used independently to classify web pages, and then, a method that integrated the features was adopted for the classification purpose was proposed. The traditional solutions for detecting spoofed or phishing websites were based on the signature method. But this approach was not able to detect the newly created spoofed websites or web pages, which is a deficiency of the model. [14] examined the challenges exhibited by the phishing detection model based on decision tree and optimal features to tackle the shortcoming which is a neural-network phishing detection model based on decision tree and optimal feature selection. This research work has shown that this approach improves the clustering algorithm, with the designs of an optimal feature selection. However, the features were not improved from the existing feature.

The work of [15] considers the use of entropy as a means of selecting features for the classification and detecting phishing websites. In this approach, machine learning (ML) techniques were adopted to classify phishing website patterns non-phishing. The results proved that the integration of the entropy FS method with the ML classifiers helps to improve their efficiency and accuracy. This was a very good approach to Phishing detection looking at the results provided. However, the complexity of the system makes it difficult to adopt. A related work of [16] worked on the detection of phishing websites using a novel two-fold ensemble model. The two-fold ensemble learner was implemented as a user-friendly, interactive decision support system for classifying websites as phishing or legitimate ones and it was effective in stopping phishing on web pages. However, the research data set used in the study is publicly available, not encrypted, and easy to analyze by fraudsters compared to other analyses with real-time data set. Also, different variants of phishing threats must be first detected rather than focusing particularly toward phishing website detection which makes the process long and tedious [17].

In another work of [18] examines phishing detection using a search engine system called Jail-Phish. The work considers the efficiency of the search engine with the capability of detecting Phishing Sites. [19] and also detection of newly registered legitimate sites. This research work is related and the method adopted was fairly effective. The limitation of this method lies in its incapability to identify any newly uploaded phishing website on the server.

[20] presented an Adaptive Neuro-Fuzzy Inference System (ANFIS) based robust scheme using the integrated features of text, images, and frames for web-phishing detection and protection. But this model was not fully explored. [21] also worked on a phishing detection model with multi-filter approach. The model incorporates five layers: Auto upgrade white list layer, URL features layer, Lexical signature layer, String matching layer, and Accessibility Score comparison layer. It was established that single filter methods are incapable to detect different group of phishing attempts. In the related work of [22], a different data mining technique was adopted to decide whether a website is legitimate or phishing and random forest as its classifier. However, the number of rules adopted was not enough for the effective classification of the system into phishing and non-phishing. [23] focused on improving the performance of URL-based detection and [24] also examined the detection accuracy of phishing website considering selected phishing detector algorithm.

In the above-reviewed work, it was observed that various machine learning techniques have been adopted with the limited number of features extracted for the classification process. This study employed sixteen rules for efficient classification of phishing and non-phishing website.

3. Materials and methods

This section discussed various materials and methods employed for the implementation of the design of the proposed rule-based classification system.

3.1. Model Design

The model design consists of three stages namely, the data acquisition stage, feature extraction stage, and classification stage. From the data acquisition stage, phishing and non-phishing URLs were collected from PhishTank, URLhaus, OpenPhish, and Moz Trusted URLs. The feature extraction stage extracts the features used for the classification using a set of rules. A total of sixteen (16) rules were extracted from the URLs. The extracted rules were then saved. The classification stage classifies the input URL to either phishing or non-phishing using Random Forest. Figure 1 shows the block diagram of the designed model which consists of three stages as shown in the diagram. Data acquisition is the first stage followed by feature extraction and classification which is the last stage of the process.

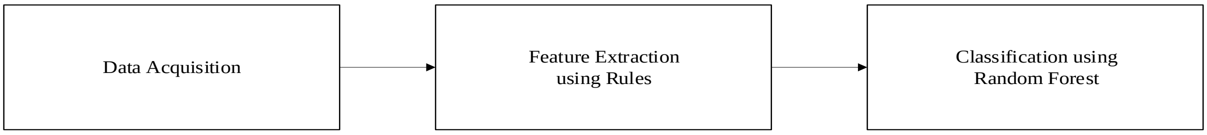

Data Acquisition
Feature Extraction using Rules
Classification using Random Forest

Figure 1. Block diagram of the designed model

3.2. Data Collection

Two classes of data which are phishing and non-phishing URLs were collected online from PhishTank, URLhaus, OpenPhish and Moz Trusted URL with a total of 2,015 URLs. The dataset of 1,612 which is 80% were used for training while 403 which is 20% were used to test the model. The training dataset obtained was used to train the model while the testing dataset was used to evaluate the performance of the developed model using Random Forest. The sample of the collected data both non-phishing and phishing URL is shown in Table 1 and 2 respectively.

Table 1: Sample of Non-phishing data

http://www.emnck.com:3000/archive/egan.tml
http://danoday.com/summit.shtml
http://www.livinglegendltd.com
http://voicechaser.com/forum/viewforum.php?f=8
http://www.hollywoodcollectorshow.com
http://www./geocities.com/hollywood/hills/8944/
http://asifa.proboards61.com/index.cgi?action=calenderviewall
http://us.imdb.com/name/nm0267724
http://www.pamelynferdin.com/
http://google.com

Table 2: Sample of Phishing data

http://asesoresvelfit.com/media/datacredit.co
http://hissoulreason.com/js/homepage/
http://unauthorized.newpage.com/webapps/66fbf/
http://133.130.103.10/23/
http://133.130.103.10/21/logar/
http://httpssicredi.esy.es/servico/sicredi/validarclientes/mobi/index.php
http://gamesaty.ga/wp-content///hy/en/?i=314amp;i=31416
http://luxuryupgradepro.com/ymailNew/ymailNew/
http://caixa.com.br.fgtsagendesaqueconta.com/consulta8523211/principal.php?
http://131,5,05,000000,12,11

3.3. Rule Based Feature Extraction

Feature extraction is mostly adopted to find the significant features for a given set of input data. The feature extractor first analyses the URL of the website to find the URL based features. Sixteen features were identified and examined for the extraction process. These include (the length of URL, having @ symbol, double slash redirecting, prefix suffix, having sub domain, having IP address, shortening service, HTTPS token, web traffic, domain registration length, age of domain, DNS record, statistical reports, get protocol, get domain and get path). Rules were generated for each of these features for extraction process. Two of the rules generated for the extracted features are discussed as follows:

1. **Length of URL:** On idea situation, length of the should not exceed or not up to 54 characters and should not exceed. If it exceeded 54 characters then the URL becomes phishing URL. The doubtful part of the URL is hidden by the phishers. 75% of the Phishing URLs contain the average length, and the URL length varies based on different websites. Phishing websites misrepresents the order of the fully qualified host name of the URL, and they always hide the registered domain. The rule for the length of URL is given in Eq. (1).

$$Rule_{leng_{URL}} = \\ = \begin{cases} URL_{length} < 54, Non-phishing \\ URL_{length} \geq 54 \leq 75, Suspicious \\ \quad Otherwise \quad , Phishing \end{cases} \tag{1}$$

2. **Having IP Address:** The legitimate website addresses are associated with domain names. The dataset checks for the IP address present in the hostname of the URL or not. If IP address is used as an alternative of a domain name in the URL e.g., 125.98.3.123 or it can be transformed to hexadecimal representation e.g., http://0x58.0xCC. 0xCA.0x62, then it is a phishing URL. The proposed rule is shown in Eq. (2).

$$Rule_{IP\_address} =$$

$$= if \begin{cases} IP\ exists\ in\ URL\ , Phishing \\ Otherwise\ , \qquad Non-phishing \end{cases} \tag{2}$$

3. **Contains@Symbol**: refers to the URL which contains the "@" symbol. Most of the legitimate websites don't use "@" symbols. If the URL contains the "@" symbol that URL is considered a phishing URL. The web browser ignores the content before @ symbol in the URL, and the probability of redirecting to the phishing website is high. The proposed rule is shown in Eq. (3).

$$Rule_{@symbol} =$$

$$= if \begin{cases} URL\ has\ @\ symbol, \qquad Phishing \\ Otherwise, \qquad Non-phishing \end{cases} \tag{3}$$

4. **Prefix Suffix**: The dash (-) symbol is used for separating the prefix and suffix in phishing websites. Legitimate URL's don't use the dash (-) symbol. The proposed rule is shown in Eq. 4.

$$Rule_{prefix\_suffix} =$$

$$= if \begin{cases} Domain\ part\ include\ \text{-}symbol, Phishing \\ Otherwise, \qquad Non-phishing \end{cases} \tag{4}$$

3.4 Classification using Random Forest

After the features have been extracted and the dataset has been trained, the testing dataset was used to classify the URLs to evaluate the performance of the model. A total of 403 images were tested. Random Forests (RF) is a combined approach for regression and classification. RF classifier builds number of decision trees during the training and generate a class that is the mode of the classification classes of the individual trees. RF classification performs better than any other decision tree algorithms as it uses a forest of classification trees to take a decision. Using Random Forests, to classify a new object from an input vector, the input vector is given each of the trees in the forest. Each tree gives a classification, and the tree 'votes' for that class. The forest chooses that classification which has more votes over all the trees in the forest. Figure 2 shows the flowchart of the developed model for the website phishing model. The collected URL data was loaded from the PC directory and later partitioned into training and testing. The training takes 80% while testing takes 20% of the acquired URL data. Sixteen (16) features were extracted from the URL data and training was done on the extracted features, then the training parameters were saved. The testing set was used to test the proposed model using random forest to estimate the performance of the metrics.

4. Results and Analysis

For the implementation purpose, a total of 1,612 data were used to train the model, and 403 data were used for the testing. The model was processed and implemented on a Core i5-5200 CPU of 2.2GHz with 8GB RAM machine. The confusion matrix table revealed the performance of a model on the dataset collected. Table 1 shows the confusion matrix with a total of 184 URLs correctly classified as phishing while 145 URLs were correctly classified as non-phishing respectively. A total of 34 phishing URLs were classified as non-phishing while 40 non-phishing URLs were classified as phishing respectively. Figures 3 and 4 show the output of the model correctly classifying websites into phishing and non-phishing respectively. Figure 3 shows a situation where the URL was correctly classified as phishing website and Figure 4 shows where the URL was classified as non-phishing.

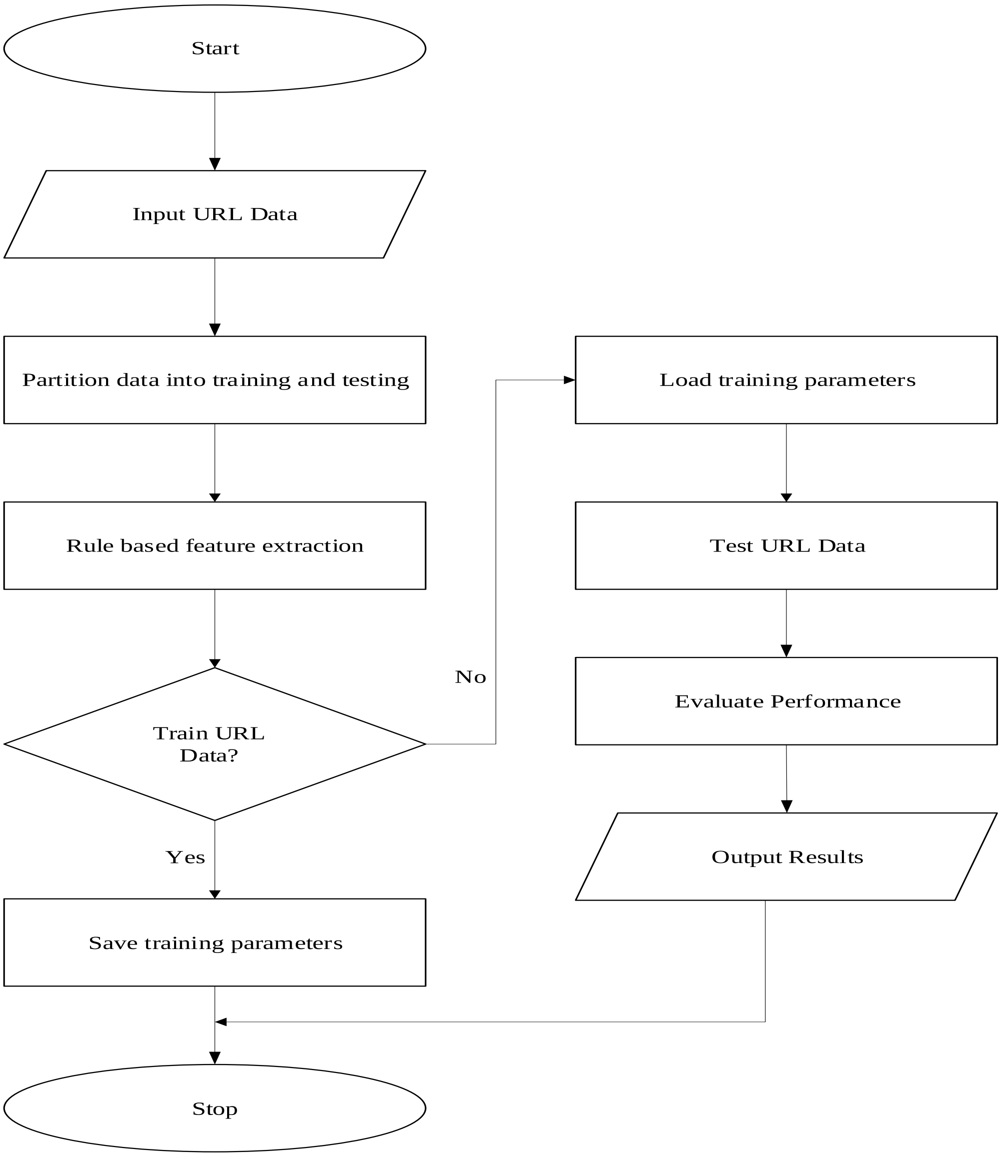
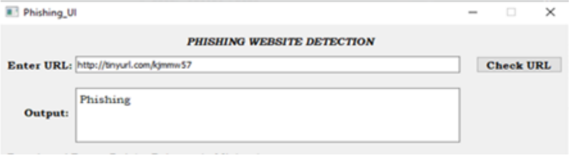


Figure 2. Flowchart of the developed model

Table 3- Confusion Matrix

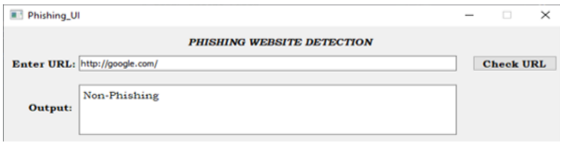
	Predicted Phishing	Predicted Non-Phishing
Actual Phishing	184	34
Actual Non-Phishing	40	145



Enter URL: http://tinyurl.com/kjmmw57

Output: Phishing

Figure 3. Phishing correctly classified as Phishing



Enter URL: http://google.com/

Output: Non-Phishing

Figure 4. Non-Phishing correctly classified as Non-Phishing

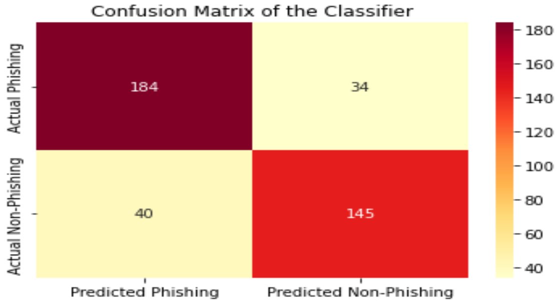


Figure 5. Confusion Matrix of the Classifier

The model was evaluated based on the following metrics, accuracy, sensitivity, specificity. Accuracy is the total number of correctly classified URLs by the total number of tested samples. Sensitivity is the proportion of true positives tests out of all URLs that is a phishing URL. Specificity is the percentage of true negatives out of all URLs that is non-phishing. These metrics were evaluated using Eqs. (5), (6), and (7). The accuracy of the model was recorded for each classification into phishing and non-phishing and the overall selected sample was retrieved using the confusion matrix. The analysis of the result obtained was represented in Figure 5. The developed system has successfully identified and analyzed different URL features using rule-based model with accuracy of 81.6%, sensitivity of 78.4%, and specificity of 84.4% obtained from the system.

$$Accuracy = \frac{Total\ Number\ of\ correctly\ classified\ URL}{Total\ Number\ of\ URLs} \qquad (5)$$

$$Sensitivity = \frac{True\ Positive(TP)}{True\ Positive(TP)+False\ Negative(FN)} \qquad (6)$$

$$Specificity = \frac{True\ Negative(TN)}{True\ Negative(TN)+False\ Positive(FP)} \qquad (7)$$

4. CONCLUSION

In this research work, the characteristics of the URLs were examined and analyzed. The rules which are generated are translated by highlighting some of the features which are more important for the classification process. It was revealed that the classification of the website into phishing and non-phishing using a rule-based algorithm is viable. The use of rules shows the helpful features which are present in phished URLs. This study has investigated problems presented by phishing and proposed a model, which describes the complete life cycle of phishing attacks. The adoption of Random Forests as an ensemble approach for classification and regression has tremendously enhanced the classification process. This approach has also demonstrated sufficient capability in handling phishing URLs. This approach provides a wider outlook for phishing attacks and provides an accurate definition covering end-to-end encryption.

5. REFERENCES

[1] Kumar, M. V. and Lalitha, T. Soft Computing: Fuzzy Logic Approach in Wireless Sensors Networks. Circuits and Systems, vol. 7 no. 8, pp. 1242–1249, 2016.

[2] Patil, S. and Dhage, S. (2019). A methodical overview on phishing detection along with an organized way to construct an anti-phishing framework. In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pp. 588–593. IEEE.

[3] Catherine, O. D. An Intelligent Rule Based Phishing Website Detection Model. International Journal of Computer Science and Mathematical Theory, vol. 5 no 2 pp. 10-19, 2019.

[4] Ojolo, T. Youth perception on yahoo-yahoo (cybercrime): A case study of Ado-Ekiti, Ekiti State Nigeria. Unpublished PhD thesis, 2017.

[5] Krombholz, K., Hobel, H., Huber, M., and Weippl, E. Advanced social engineering attacks. Journal of Information Security and applications, vol. 22 pp.113–122, 2015.

[6] Huang, H., Qian, L., and Wang, Y. A SVM-based technique to detect phishing URLs. Information Technology Journal, vol. 11 no. 7 pp. 921-925, 2012.

[7] Simon, T. Chapter seven: Critical infrastructure and the internet of things. Cyber Security in a Volatile World, 93, 2017.

[8] Buch, R., Ganda, D., Kalola, P., and Borad, N. World of Cyber Security and Cybercrime. Recent Trends in Programming Language, vol. 4 no. 2, pp. 18-23, 2017.

[9] Kahyaoglu, S. B. and Caliyurt, K. Cyber security assurance process from the internal audit perspective. Managerial Auditing Journal, vol. 33 no 4, pp. 360-376, 2018.

[10] Hewage, C. Coronavirus pandemic has unleashed a wave of cyber attacks–here's how to protect yourself. Technical Report, 2020.

[11] Moghimi, M. and Varjani, A. Y. New rule-based phishing detection method. Expert systems with applications, vol. 53 pp. 231–242, 2016.

[12] Kaytan, M. and Hanbay, D. Effective classification of phishing web pages based on new rules by using extreme learning machines. Computer Science, vol. 2 no. 1, pp. 15–36, 2017.

[13] Gandotra, E. and Gupta, D. Improving Spoofed Website Detection Using Machine Learning. Cybernetics and Systems, vol. 52 no 2 pp. 169– 190, 2021.

[14] Zhu, E., Ju, Y., Chen, Z., Liu, F., and Fang, X. DTOF-ANN: An artificial neural network phishing detection model based on decision tree and optimal features. Applied Soft Computing, 95:106505, 2020.

[15] Ali, S., Shahbaz, M., and Jamil, K. Entropy-Based Feature Selection Classification Approach for Detecting Phishing Websites. In 2019 13th International Conference on Open Source Systems and Technologies (ICOSST), pp. 1–6. IEEE, 2019.

[16] Nagaraj, K., Bhattacharjee, B., Sridhar, A., and Sharvani, G. Detection of phishing websites using a novel twofold ensemble model. Journal of Systems and Information Technology, vol. 20 no 3, pp. 321-357, 2018.

[17] Saleem, H. and Naveed, M. SoK: Anatomy of Data Breaches. Proceeding of Privacy Enhancing Technologies, vol. 2020 no. 4, pp. 153–174, 2020.

[18] Rao, R. S. and Pais, A. R. (2019). Detection of phishing websites using an efficient feature-based machine learning framework. Neural Computing and Applications, 31(8):3851–3873.

[19] Fischer, E. A. The Impact of Cybersecurity. Cyberwarfare, pp 10, 2017.

[20] Adebowale, M. A., Lwin, K. T., Sanchez, E., and Hossain, M. A. Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text. Expert Systems with Applications, 115:300-313 2019.

[21] Sonowal, G. and Kuppusamy, K. PhiDMA–A phishing detection model with multi-filter approach. Journal of King Saud University-Computer and Information Sciences, vol. 32 no. 1, pp. 99–112, 2020.

[22] Abdulrahman, M. D., Alhassan, J. K., Adebayo, O. S., Ojeniyi, J. A., and Olalere, M. Phishing attack detection based on random forest with wrapper feature selection method., 2019.

[23] Tupsamudre, H., Singh, A. K., and Lodha, S. Everything is in the name – A URL based approach for phishing detection. In International symposium on cyber security cryptography and machine learning, pp. 231–248. Springer, 2019.

[24] Abidoye, A. P. and Kabao, B. (2020). Hybrid Machine Learning: A Tool to Detect Phishing in Communication Network. International Journal of Advanced Computer Science and Application, vol. 11 No 6 pp. 559-569.