International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 780-788 www.ijaem.net ISSN: 2395-5252

Phishing Detection Using ML Based URL Classification

Sonam Malviya

Submitted: 10-03-2022 Revised: 21-03-2022 Accepted: 23-03-2022

ABSTRACT— The cyber security is one of the essential domains of research due to increasing use of internet infrastructure and wide level of increasing web based services and applications. These applications are not much secure due to tricky cyber attacks such as phishing. In this presented work, we are investigating the techniques of phishing attack deployment additionally proposed a new technique of phishing URL classification. The phishing URL classification is a task of cyber security which is previously performed by using the concept of only considering the set of phishing URLs. In this presented work we are considering the phishing URLs as well as legitimate URLs for designing the data model for classification task. In this context, we have collected phishing URLs from the phish tank database and for legitimate URLs we have collected some common URLs. Further the 14 features have been calculated using the collected URLs and the ANN will be trained with the feature data. The trained model has been used for recognizing the unknown URLs. The implementation of the proposed model has been carried out using the JAVA technology and with the help of WEKA machine learning library. Additionally a comparison with the available phishing detection model has been carried out. The performance of the proposed model in terms of accuracy and other performance parameters indicates the proposed model is able to recognize the 76.2% accurate classification with minimal resource consumption. Thus the proposed model is a promising model and can be used for future work extension with deep learning and large scale data analysis.

Keywords— machine learning, data mining, phishing URL, classification, binary classification problem, phishing detection.

I. INTRODUCTION

Increasing use of internet is increasing the profitability of human life in terms of building social, as well as professional task. The internet based applications has been used in various aspects such as banking, online shopping, communication, email and many more. That offers a number of services at our door step [1]. On the other side the internet has also been used by malicious intuitions by the hackers and attackers. These malicious users are trying to capture the various users' sensitive and private information by which they are conducting the frauds. Such kinds of attacks are much crucial and hard to detect by the normal security software. Among them the phishing attack is one of the critical and harmful attacks [2]. Thus in this presented work we are motivated to study and design a phishing detection approach using machine learning technique.

The proposed technique is utilizes the phishing URLs for extracting the valuable properties form the URLs. These properties of URLs are useful for identifying the URLs in terms of phishing or legitimate. Additionally for make use of these extracted features for developing a data model we have proposed to use the popular machine learning algorithm namely artificial neural network (ANN). The ANN is utilizing the extracted features of URL to learn and identify the similar URL patterns when an unknown URL has found. This approach is based on a binary classification technique. The study has provided the detailed understanding about the proposed machine learning model for classifying the phishing URLs.

II. PROPOSED WORK

The cyber security is one of the most popular research domains classically; additionally its scope has increasing day by day as the different cyber infrastructure has growing. Therefore, we need to study the cyber security and threats that impact on human life. In this work we are studying the one of the most crucial cyber attack namely phishing attack. This section offers describes the basics of the proposed technique which is designed for dealing with the phishing attack.

A. System Overview

The machine learning is a technique which is used for analyzing the data for recovering the knowledge from the raw data. The ability of ML techniques for prediction, classification, clustering make it profitable various application point of view such as security, smart city, medical science, engineering and many more. In this presented work we are investigating the employment of machine learning technique in cyber security, more specifically in detection of phishing attack. The phishing attack is one of the serious attacks, which is deployed with the social engineering. The attack tries to capture the victim's personal and confidential information. This stolen information can be used for conducting the frauds. In this context in most of the cases the attacker prepares a clone of an authentic website and the link of this forge website has distributed by using the communication channels such as email, social media message, or in other communication technique.

The victim has visited this link and put their details on the page which is captured by the attack and utilized in various malicious tasks. Therefore, there is a need of method which will analyze the message URLs to classify them into malicious and legitimate. However, there are a number of techniques are recently proposed and developed for preventing such kind of attack but most of them are not much effective. Thus in this work we are proposed an artificial neural network based technique for dealing with the phishing URL classification. The next section will discuss the proposed method for URL classification.

B. Proposed Methodology

The proposed phishing URL classification model based on ANN has demonstrated in figure 1. In this model the components of the model have been described additionally their functional aspects are described in this section.

Phish tank database: the phish tank database is a security database which maintains the historical phishing records. The dataset includes different attributes such as phish ID, URL, phish detail URL, submission date, verification time, online status, target. The dataset is available in two different formats we can use the dataset as Comma separated Vector (CSV) format or we can also use this database directly in our program using web service based Application Programming Interface (API). In this work we are using the CSV format of the data.

Self obtained URLs: the neural network is a type of classifier which will need to be train on minimum two classes. Thus we obtain the phishing class data using the phish tank and the second class of data which is legitimate has collected by own. Thus we collect some legitimate URLs from web directly and use as the legitimate class.

Combined data: the dataset which we required is required to have two classes thus we have combined the phish tank URLs with the class label "Phishing" and the self collected URLs with the class label "Legitimate". The combined data has been used with the next process for preparing the required data model.

Data preprocessing: the data preprocessing is an essential task in most of the machine learning based applications. The data pre-processing will help to refine the contents and improve the quality of learning data. However, in this presented work we have removing the different additional attributes of the obtained dataset and only keep phishing URL which will be used for deploying the attack.

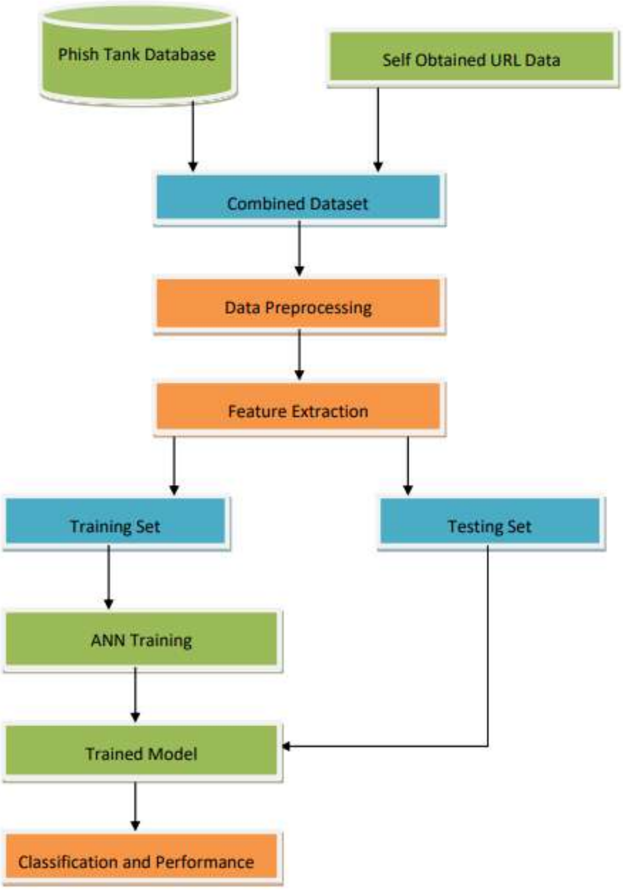


Figure 1 proposed phishing URL classification model

Feature extraction: the machine learning algorithms are cannot be directly learn with the URL data for recognizing the patterns. Thus we need to transform the URL data into a vectored form for utilizing with the learning algorithm. Thus we have extracting some essential attributes form the URL data. In this context we make use the article [3] technique for identifying the features form the URL to learn with the proposed algorithm. In this article we have found the following features:

1. Host URL length
2. Slashes in URL
3. dots in host name of the URL
4. In host name of the URL number of terms
5. special characters
6. IP address
7. Unicode in URL
8. transport layer security
9. Subdomain
10. URL with certain keyword
11. top level domain
12. In the path of the URL number of dots
13. Host name of URL with hyphen
14. URL length

The article [3] provides the 14 different properties to be compute from each URL. Using these properties we have computing a numerical value for each individual property. Thus each URL returns a set of 14 values as the attribute and each URL has a class label associated. Finally we have encoding the features into a binary string using the threshold value defined with each property. In order to encode the computed features we compare each value with the defined threshold if the value is higher than the threshold then the feature is recognize it as 1 otherwise it is 0.

Training set: after transforming the URLs into a 2D vector the data can be used for learning with the machine learning algorithms. Thus we have split the entire data into two parts first part of the data has used for the training of the ANN and second part is being used for validation of the model. Here we have used 80% of randomly selected samples as training set.

Testing set: the testing set will be used for validation of trained model. Thus the 20% of randomly selected data will be used here for testing of the trained model.

ANN training: An Artificial Neural Network (ANN) is data processing technique based on the concept of biological nervous systems. The ANN is structured as the network, which is composed by interconnected elements known as neurons. The configuration is depends on specific application, such as recognition, classification or prediction. Learning of ANN has involves update to the synaptic connections. The ANN is a complex, nonlinear, and parallel system. The neural network is made with small computational units which are known as neuron. The neural network which accepting N no of inputs can be denoted as $x(n)$ and each input values are multiplied with a weight which will be in between 0-1. That weight is defined by $w(n)$. The sum of multiplication of input values and connection weight are used with a transfer function which is also known as activation function for producing the final consequences of the network.

Trained model: the trained model is prepared after adjusting the weights of the neural network. This trained model will be used to accept the test data in same format as the model would train and then classify the URL features in terms of legitimate and phishing URLs.

Classification and performance: based on the classified test data consequences we are measuring the performance of the proposed model. Thus here we will measure the accuracy and error rate. Additionally we have measured the training time and memory usage during the model training phase.

C. Proposed Algorithm

This section provides the summarized steps of the proposed data model for processing the URL data. The table 1 demonstrates the steps of the proposed algorithm.

Table 1 steps of proposed URL classification model

```
Input: URLs database D
Output: URL class labels C
Process:
1.  $D_n = readDataset(D)$
2.  $P_n = preprocessDataset(D_n)$
3.  $for(i = 1; i \leq n; i++)$
a.      $F_i = ExtractFeatures(P_i)$
4.  end for
5.  $[Train, Test] = F.split(80,20, random)$
6.  $T_{model} = ANN.Train(Train)$
7.  $C = T_{model}.Classify(Test)$
8.  Return C
```

III. RESULTS ANALYSIS

The proposed work has evaluated in this chapter in order to describe the performance and also we include the comparative performance study in order to justify the performance of the proposed model based on ANN technique.

The accuracy of the ML model demonstrates how effectively a model will trained to recognize the objects. That is a ratio of correctly recognized and total samples of the model, thus using the following equation we can calculate the performance.

$$\text{accuracy} = \frac{\text{correctly classified}}{\text{total samples}} X100$$

According to the obtained results of both the models we have prepared a bar graph and a table for demonstrating the consequences of the classification accuracy. The figure 2(A) and table 1 shows the accuracy of the apriori and BPN algorithms performance in terms of percentage (%). The X axis show the amount of URLs has used for experimentation and Y axis shows the obtained accuracy. According to the results we have found the proposed model based BPN classifier will provide higher accuracy as compared to apriori based technique.

The error rate of a classifier demonstrates the misclassification rate of a ML model. Thus it can be measured as the ratio of misclassified samples and total samples for classification using the following formula:

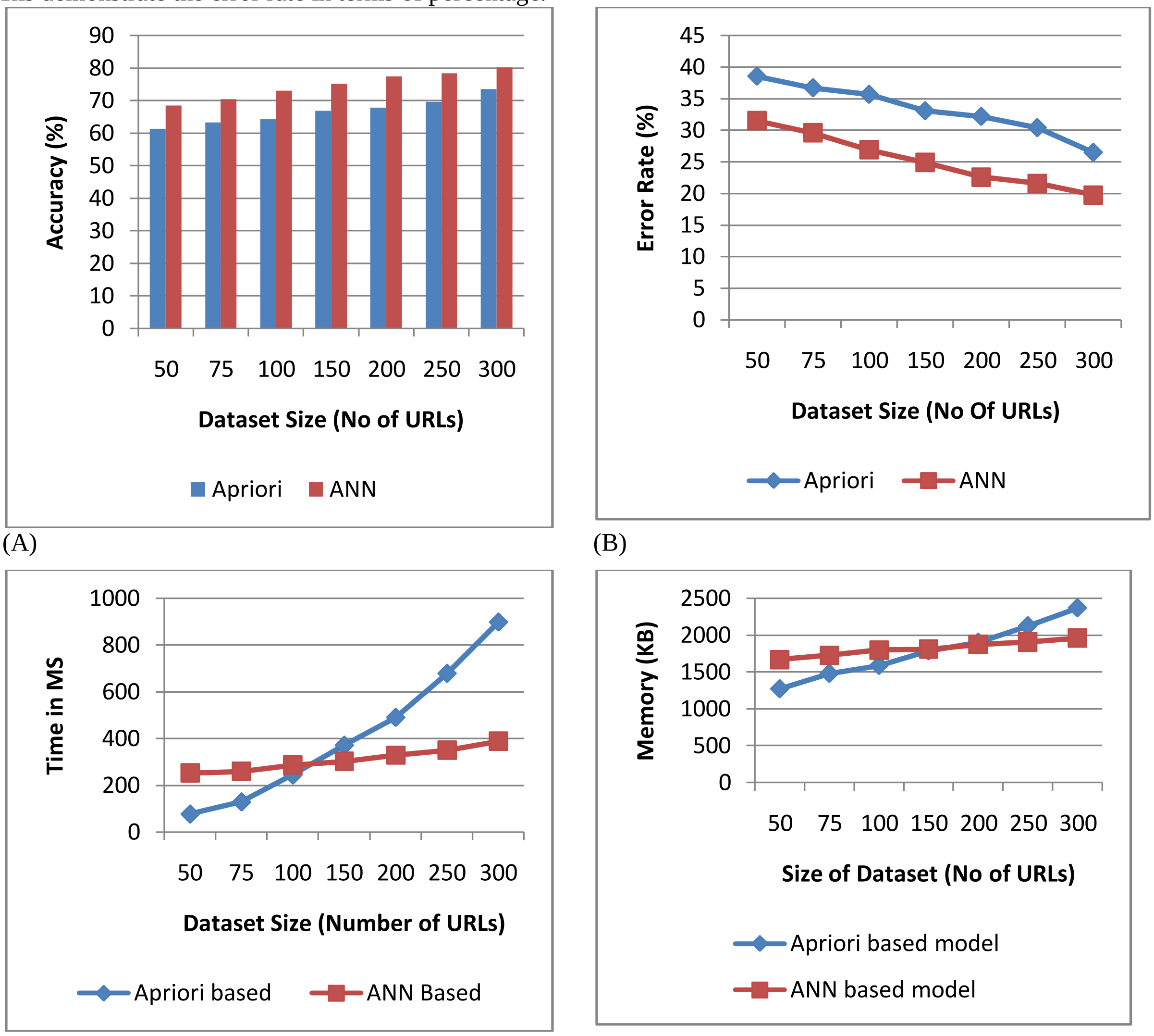
$$\text{error rate}(\%) = \frac{\text{misclassified samples}}{\text{total samples to classify}} * 100$$

The comparative performance of apriori based phishing URL classification and BPN based URL classification in terms of error rate has been demonstrated in figure 2(B) and table 1.

Table 1 performance of implemented techniques for phishing URL classification

S. no.	Dataset	Accuracy (%)		Error Rate (%)		Time in MS		Memory (KB)	
	size	Apriori	ANN	Apriori	ANN	Apriori	ANN	Apriori	ANN
1	50	61.43	68.48	38.57	31.52	78	253	1274	1672
2	75	63.28	70.44	36.72	29.56	130	260	1483	1729
3	100	64.31	73.09	35.69	26.91	245	287	1591	1799
4	150	66.92	75.11	33.08	24.89	372	302	1788	1811
5	200	67.81	77.39	32.19	22.61	491	329	1905	1875
6	250	69.59	78.41	30.41	21.59	679	351	2129	1909
7	300	73.5	80.27	26.5	19.73	898	389	2372	1962

The experiments have been carried out with the different size of samples and the performance for both the models has been measured. In this diagram the X axis shows the size of dataset used for experiment and Y axis demonstrate the error rate in terms of percentage.



(A) (B)

(C) (D)

Figure 2 shows the performance analysis of proposed ML based phishing URL classification model in terms of (A) Accuracy (B) error Rate (%) (C) Training Time (MS) (D) Memory Usage (KB)

According to the observed results the proposed BPN based URL classification technique provides more accurate results as compared to the traditional apriori based technique. Additionally as the amount of data size increases the error rate of the proposed model has reducing more as compared to apriori based technique. The training time requirements are also an essential parameter for performance analysis of a ML algorithm. In this context we have measured the training time using the following formula using the proposed system.

$$\text{Training time} = \text{end time} - \text{start time}$$

The comparative training time for both the ML models based on apriori and ANN has been reported using the figure 2(C) and table 1. In this diagram the X axis demonstrate the size of dataset used and Y axis shows the time consumed during the experiments in training of the model. According to the given results we can found the apriori algorithm time consumption has been increasing much faster as compared to ANN. Initially with the fewer amounts of attributes and transactions apriori provide efficient results but as we increasing the size of dataset the apriori algorithm time consumption has increased more. Thus according to the time consumption we found the ANN is suitable algorithm for URL classification.

The memory usages is also known as the space complexity, here we have measured the memory usage of the system by using the process based consumed memory. Thus to calculate the memory usage of the proposed model we utilize the following equation:

$$\text{memory used} = \text{total assigned} - \text{free space}$$

The table 1 and figure 2(D) shows the comparative memory consumption of the algorithms implemented for classifying the malicious URLs. The memory usages have measured here in terms of kilobytes (KB). According to the figure X axis consist of the different size of dataset for experiments and Y axis demonstrate the measured memory usages of the algorithms. According to the obtained performance the proposed ANN based URL classification model provides efficient results as compared to apriori based model. Finally based on the obtained performance by both the algorithms we can say the apriori algorithm is accurate but consumes significant amount of time and memory for classifying the URLs. On the other hand the BPN algorithm demonstrates higher accuracy with limited memory and time requirements. Thus the proposed malicious URL classification approach using ANN is acceptable for future work extension.

IV. CONCLUSION AND FUTURE WORK

This chapter provides the summary of entire study performed in order to investigate the phishing URL classification using ML approach. Thus we have reported conclusion and the feasible future extension of the proposed work in this chapter.

CONCLUSION

The phishing is one of the most crucial attacks which will be deployed by tricking the victim by clicking a malicious link. These links have the duplicate web page which is look alike the original web page. The victim fills the details on the duplicated web page and passes their valuable and confidential information to the attacker. The attacker utilizes this information for conducting the financial fraud. Therefore this attack can be prevented using the awareness about the phishing attack, but there are also security companies are making software which will identify the phishing attack. These techniques can be a list based approach or can be implemented using the machine learning techniques. However the list based techniques are accurate but the maintaining the list and searching from the list is a complex and resource consuming task. Therefore ML based techniques are much popular for the phishing detection.

In this presented work we are motivated to study the ML based phishing detection technique. That technique makes use the URLs for recognizing the URL is malicious or legitimate. Therefore, in order to train the proposed ML based model we are utilizing the phish tank database. That database consists of previously reported phishing URLs using these recovered URLs we are generating a set of features based on URLs. Further we have utilized the features with the back propagation neural (BPN) network for training the algorithm. After the training of artificial neural network (ANN) it is prepared for accepting the test URLs and classify the URLs into legitimate and phishing URL. The implemented approach is a binary classification technique. That approach has also been compared with the available Apriori based phishing URL classification technique.

The proposed phishing URL classification system using ANN technique has been implemented using the JAVA and WEKA technology. Additionally for providing the performance analysis the MySQL has been used for preserving the experimental observations. The table 6.1 demonstrates the mean performance of the experiments with the proposed and the traditional apriori algorithm based model.

Table 6.1 Mean performance of the models

S. No.	Parameters	Apriori Based	ANN based
1	Accuracy	66.70 %	74.89 %
2	Error rate	33.30 %	25.11 %
3	Training Time	413.28 MS	310.41 MS
4	Memory usage	1791.71 KB	1822.42 KB

According to the performance demonstrate in the table 6.1 the proposed model of ANN based URL classification provides the efficient and accurate results. Thus the ANN based security models for phishing attack classifications are much effective then the classical ML approaches.

FUTURE WORK

The established key objectives of the proposed work have been accomplished successful. In order to improve this model more the following suggestions has been made.

1. The ANN based model will provide great strength and efficiency for classifying the phishing URLs, thus in near future the proposed model need to be used with bulk amount of phishing and legitimate URLs
2. The deep learning techniques are able to directly work on different formats of data thus in near future the deep learning models has been suggested to be used for more effective phishing attack detection..

REFERENCES

[1] Y. K. Dwivedi, et al., “Setting the future of digital and social media marketing research: Perspectives and research propositions”, Inter. Jou. of Infor. Mana. 59 (2021) 102168

[2] J. J. Jaccard, S. Nepal, “A survey of emerging threats in cybersecurity”, Jour. of Comp. and Sys. Sci. 80 (2014) 973–993

[3] S. C. Jeeva, E. B. Rajsingh, “Intelligent phishing url detection using association rule mining”, Hum. Cent. Comput. Inf. Sci. (2016) 6:10, DOI 10.1186/s13673-016-0064-3

[4] D. Watson, T. Holz, S. Mueller, “Know your enemy: Phishing, behind the scenes of Phishing attacks”, The Hone. Proj. & Rese. Allia. (2005)

[5] T. Jagatic, N. Johnson, M. Jakobsson, F. Menczer, “Social Phishing, Community”, ACM 2007, Vol. 50, No. 10, pp. 94-100

[6] R. Basnet, S. Mukkamala, A. H. Sung, "Detection of phishing attacks: A machine learning approach", So. Comp. Appl. in Ind., Spri. Ber. Heid., PP. 373-383, 2008.

[7] H. Tout, W. Hafner, “Phishpin: An identity-based anti-phishing approach”, in proce. of inte. Conf. on comp. sci. and engg, Vanc., BC, pp 347-352, 2009

[8] APWG, “Phishing Activity Trends Report”, http://antiphishing.org/reports/apwg_report_DEC2005_FINAL.pdf. Access date [4/1/2018]

[9] T. Gundel, “Phishing and Internet Banking Security, Technical Security report”, IBM Cr. Comp. Cen., 2005

[10] I. R. A. Hamid, J. Abawajy, T. Kim, “Using Feature Selection and Classification Scheme for Automating Phishing Email Detection”, Stu. in Info. and Con. 22(1): pp. 61-70, March 2013

[11] V. Suganya, “A Review on Phishing Attacks and Various Anti Phishing Techniques”, Inter. Jour. of Comp. Appl., Vol. 139 – No.1, Apr. 2016.

[12] “Data Mining: What is Data Mining?”, http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/dataminin.htm

[13] “Data Mining - Applications & Trends”, http://www.tutorialspoint.com/data_mining/dm_applications_trends.htm

[14] M. Chowdhary, S. Suri, M. Bhutani, “Comparative Study of Intrusion Detection System”, 2014, IJCSE All Rights Reserved, Volume-2, Issue-4

[15] Mrs. P. Muley, Dr. A. Joshi, “Application of Data Mining Techniques for Customer Segmentation in Real Time Business Intelligence”, International Journal of Innovative Research in Advanced Engineering, Issue 4, Volume 2 (April 2015)

[16] Q. Zhao, S. S. Bhowmick, “Association Rule Mining: A Survey”, Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116, 2003

[17] P. Asthana, A. Singh, D. Singh, “A Survey on Association Rule Mining Using Apriori Based Algorithm and Hash Based Methods”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, July 2013

[18] X. Wu, X. Zhu, G. Q. Wu, W. Ding, “Data Mining with Big Data”, IEEE Transactions on Knowledge and Data Engineering Volume:26 , Issue: 1

[19] F. Cao, J. Liang, D. Li, L. Bai, C. Dang, “A dissimilarity measure for the k-Modes clustering algorithm”, Knowledge-Based Systems, 2011 Elsevier B.V. All rights reserved

[20] S. Dzeroski, “Multi-Relational Data Mining: An Introduction”, ACM SIGKDD Explorations Newsletter Homepage archiveVolume 5 Issue 1, July 2003

[21] Q. Wan, A. An, “Compact Transaction Database for Efficient Frequent Pattern Mining”, 2005 IEEE International Conference on Granular ComputingVolume:2

[22] S. H. Liao, P. H. Chu, P. Y. Hsiao, “Data mining techniques and applications – A decade review from 2000 to 2011”, Expert Systems with Applications, 2012 Elsevier Ltd. All rights reserved

[23] W. Lin, M. A. Orgun, G. J. Williams, “An Overview of Temporal Data Mining”, http://research.microsoft.com/apps/pubs/default.aspx?id=71389

[24] B. Lalithadevi, A. Merry Ida, W. A. Breen, “A New Approach for Improving World Wide Web Techniques in Data Mining”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 1, January 2013

[25] V. Shreeram, M Suban, P Shanthi, K Manjula, “Antiphishing detection of phishing attacks using genetic algorithm”, Proceedings of the International Conference on Communication Control and Computing Technology, pp. 447-450, 2010

[26] J. Chen, C. X. Guo, “Online Detection and Prevention of Phishing Attacks”, Proceeding of the First International Conference on Communication and Networking in China, Beijing, pp. 1-7, 2007.

[27] M. Dunlop, S. Groat, D. Shelly, “ Goldpolish: Using Images for Content-based Phishing Analysis, In Proceedings of the Fifth International Conference on Internet Monitoring and Protection, Barcelona, pp. 123-128, 2010.

[28] S. Shah, “Measuring Operational Risks using Fuzzy Logic Modeling”, Article, Towers Perrin, JULY 2003.

[29] H. Tout, W. Hafner, “Phishpin: An identity based anti-phishing approach”, in proceedings of international conference on computational science and engineering, Vancouver, BC, pp. 347-352, 2009.

[30] S. Afroz, R. Greenstadt, "Phishzoo: Detecting phishing websites by looking at them", In Semantic Computing (ICSC), 2011 5th IEEE International Conference on, pp. 368-375, IEEE, 2011.

[31] D. Sahoo, C. Liu, S. C. Hoi, "Malicious URL detection using machine learning: A survey", arXiv preprint arXiv: 1701.07179 (2017).

[32] S. Kotsiantis, D. Kanellopoulos, “Association Rules Mining: A Recent Overview”, GESTS International Transactions on Computer Science and Engineering, Volume 32 (1), 2006, pp. 71-82.

[33] J. Usharani, Dr. K. Iyakutti, “Mining Association Rules for Web Crawling using Genetic Algorithm”, International Journal Of Engineering And Computer Science, Volume 2 Issue 8 August, 2013, pp. 2635-2640

[34] U. Fayyad, G. P. Shapiro, P. Smyth, "The KDD process for extracting useful knowledge from volumes of data." Communications of the ACM 39.11 (1996): pp. 27-34.

[35] A. A. Orunsolu, A. S. Sodiya, A. T. Akinwale, “A predictive model for phishing detection”, Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

[36] A. Odeh, I. Keshta, E. Abdelfattah, “Machine Learning Techniques for Detection of Website Phishing: A Review for Promises and Challenges”, 978-0-7381-4394-1/21/$31.00 ©2021 IEEE

[37] H. Shirazi, B. Bezawada, I. Ray, “Know Thy Domain Name: Unbiased Phishing Detection Using Domain Name Based Features”, SACMAT’18, June 13-15, 2018, Indianapolis, IN, USA

[38] L. S. Songare, Dr. D. Verma, “The Survey and Proposal on Machine Learning Based Phishing Detection Techniques”, Turkish Journal of Computer and Mathematics Education Vol.11 No.01 (2020), 311– 321

[39] M. Zabihimayvan, D. Doran, “Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection”, IEEE International Conference on Fuzzy Systems 2019

[40] M. Das, S. Saraswathi, R. Panda, A. K. Mishra, A. K. Tripathy, “Exquisite Analysis of Popular Machine Learning–Based Phishing Detection Techniques for Cyber Systems”, Journal of Applied Security Research, https://doi.org/10.1080/19361610.2020.1816440

[41] O. K. Sahingoz, E. Buber, O. Demir, B. Diri, “Machine learning based phishing detection from URLs”, Expert Systems With Applications 117 (2019) 345–357

[42] Y. A. Alsariera, V. E. Adeyemo, A. O. Balogun, A. K. Alazzawi, “AI Meta-Learners and Extra-

Trees Algorithm for the Detection of Phishing Websites", IEEE access VOLUME 8, 2020

[43] M. Vijayalakshmi, S. M. Shalinie, M. H. Yang, U. R. Meenakshi, "Web phishing detection techniques: a survey on the state-of-the-art, taxonomy and future directions", IET Netw., 2020, Vol. 9 Iss. 5, pp. 235-246, © The Institution of Engineering and Technology 2020

[44] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques", Telecommunication Systems (2021) 76:139–154