International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Antiphishing Model Based on Similarity Index and Neural Networks

Bhawna Sharma, Parvinder Singh, Jasvinder Kaur, Pablo García Bringas

Abstract: ***Phishing is a negative technique that is used to steel private and confidential information over the web. In the present work author proposed a hybrid similarity of Cosine and Soft Cosine to calculate the similarity between the user query and repository as an anti-phishing approach. The proposed work model uses a multiclass learning method called Feed Forward Back Propagation Neural Network. The model evaluation results with 100 to 3000 test files shows that the hybrid model is able to detect the phishing attack with an average precision of 71% and is highly effective.***

Keywords: Phishing, Similarity, Neural Network.

I. INTRODUCTION

The world is moving quite fast in terms of technology and advancements. Modern day processing architectures like big data and Cloud computing are live examples of the advancements which not only possess hardware sophistication but are also capable of complex instruction set processing. Social Media like Facebook, Twitter and Instagram are also recent advancements of the modern frameworks of information technology [1, 13, 15]. It is concluded from the records that more than 95% e-users, uses social media platforms. Interest based recommendations over the webservers are also becoming an architectural methodology from the last couple of years. Information sharing over internet connects people from around the world but also raises risk of security and privacy leakage. Phishing is a negative art in order to mis-lead a user or to misuse someone private information [2, 16, 17]. Obviously a common user gets attracted towards fake illuminations of internet. It is found from the past records that more than 18% users, uses internet, to overcome their depressive life elements and they get stuck with internet phishing [3, 4]. Password hacks, malware, ip-spoofing are kind of phishing attacks which is commonly observed over internet protocol. A list of phishing observations are listed in the literature survey. Different authors from around the world have proposed different protocol suites and architecture for the prevention of phishing attacks. Rule based prevention mechanism is one of the tried and testified methods of phishing attack prevention [5,19]. This method analyses the definition from a set of rule and if the testing definition matches with the list definition then the data is said to be phishing as shown in Fig 1.

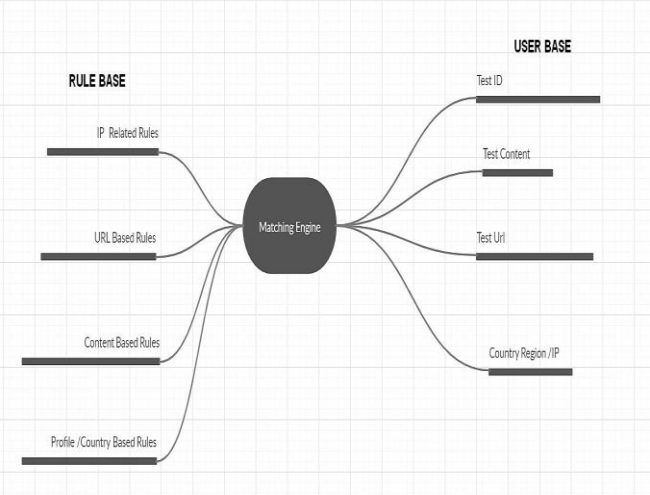

Fig 1: Rule Base with User Base

Most of the antivirus systems uses the architecture of rule base and user base matching pattern in order to identify threat [6-7]. The issue with this architecture is that, for each new type of phishing architecture, it becomes mandatory to add data to rule set. This means that the rule engine is not adaptive. To overcome this problem, researchers aimed to select modern frame adaptive algorithms like Swarm Intelligence and Machine Learning [8,20]. The main contribution of this paper is to identify the phishing architecture based on relative similarity by the integration of Machine learning [9-10]. The rest of the paper is organized in the following manner. Section 2 illustrates the literature survey which describes different phishing attacks and prevention methods used by the researchers around the world. Section 3 illustrates the proposed methodology in the same contrast and Section 4 describes the results for the same. Section 5 concludes the paper.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Bhawna Sharma*, Research Scholar, Deenbandhu Chhotu Ram University of Science & Technology, Murthal, Sonepat-131039, Haryana, India, Email: bhawnash024@gmail.com
Dr. Parvinder Singh, Professor, Department of Computer Science & Engineering, Deenbandhu Chhotu Ram University of Science & Technology, Murthal, Sonepat -131039, Haryana, India, Email: parvinder23@rediffmail.com
Dr. Jasvinder Kaur, Assistant Professor , Department of Computer Science & Engineering, PDM University, Bahadurgarh-124507, Haryana, India, Email:jasvinder.kaur@pdm.ac.in
Dr. Pablo García Bringas, Professor, University of Deusto. Spain, Email: pablo.garcia.bringas@deusto.es

II. LITERATURE SURVEY

In 2007, Steve et al. developed an online anti-phishing game to predict the effectiveness of the Anti-Phishing Phil. The game aimed to educate users to identify phishing websites, phishing elements in web browsers and search engines.
The whole work methodology revolved around the science principles aimed at the refinement of the game architecture. The results of the proposed game design had shown that this approach offered an interesting method to educated masses against phishing attacks. The game offered the spread of phishing education to masses in real terms [1]. In 2008, Cao et al. presented an Automated Individual White-List (AIWL) as an anti-phishing approach that routinely retained white-list of user's all familiar Login User Interfaces (LUIs) of web sites utilizing the strengths of Naïve Bayesian classifier. AIWL was bestowed with an alert whenever the legitimate IP got maliciously attacked because of the fact that the websites nearly had a stable IP addresses. The authors concluded AIWL as a competent programmed tool specialized for identifying pharming and phishing websites [2]. In the same year, Medvet et al proposed a new technique that could visually identify a phishing webpage. The approach was inspired from the browser plugin-AntiPhish and DOMAntiPhish. To achieve this, authors designed a 3-page features that played an important role in the phishing of the website. These features involved images, text and styles that form the visual components of a web page. The experimental evaluation involved the dataset of 41 real phishing pages and their corresponding legitimate pages. The study showed optimal results for missed detection and false alarms with no false positives [3]. In 2011, Arachchilage et al. used the theoretical model inspired with the Technology Threat Avoidance Theory (TTAT) to deal with the game design problems. The sample of this design could be reached on Google App Inventor Emulator. The authors had designed an anti-phishing mobile game design to educated masses so that they could identify phishing websites by analyzing the URLs to protect them from phishing attacks. For future research authors proposed its application as a modified model to fight phishing attacks dealing with sign, jargons and lock icons of the webpage and email messages [4]. In 2012, Ramanathan et al. proposed a multi-layered anti-phishing technique called phish GILLNET. The phishGILLNET3 was evaluated with a data of 400,000 emails and it proved to be the best among the prevailing antiphishing methods. The technique required only 10% of the annotated data that in turn lowers the error rates to save the time and labour. The application had shown encouraging results for detecting phishing attempts at chat, Facebook, blogs, Twitter, instant messages with a condition that the webpage content must be present in HTML and MIME formats [5]. In 2013, Almoman et al. had presented a comprehensive surveillance on anti-phishing approaches. The study focused the various methods for phishing protection. The study involved the phishing threats, its current status and possible solutions in near future. The authors summarized that the most of the past studies involved the merits of machine learning techniques (supervised, unsupervised and hybrid) that had a questionable aspect of time and cost. The numerous algorithms and methods had been developed to detect the phishing attack but nothing had been found to prevent the phishing attach as such [6]. In 2015, Solanki et al. proposed a phishing site identification model based on the features that could distinguish a phishing website from a legitimate website. In the process, SVM was employed as a machine learning algorithm for the purpose of classification. The designed model demonstrated the ability to successfully detect even temporary and new phishing websites. The system showed very low false-positive rate with 96% accuracy of detection. The model had employed the strengths of generated tree and decision tree algorithms [7]. In 2015, Mensah et al conducted a study that was aimed to evaluate the significance of the cipher-suite, SSL/TLS protocol version and certificate information in distinguishing the legitimate and detecting the phishing sites. The author had designed an intelligent system that had showed an instrumental ability to detect phishing websites with accuracy found between 87% and 92% [8]. In 2016, Nguyen et al. proposed a novel neuro-fuzzy ideal for detecting phishing attacks. The technique had employed a dataset of legitimate (10,000) and phishing websites (11,660) that was trained using neural network with adaptive learning rates. In the process Heuristics value was also computed. The results of the study showed its effectiveness in identifying the phishing websites [9]. In 2016, Arachchilage et al. reported an educational tool that was developed as a mobile game model to educate the users against features of phishing websites while encouraging them to avoidance circumstances that made them victim of phishing attacks. The results summarized that that learning of users had taken place demonstrating that the participants have showed 56% success rate in the pre-test and 84% success rate in the post-test. It has also raised the avoidance behaviour by 28% implying the self-efficacy and understanding the phishing threats. Although the study had numerous merits, it had a negative impact adjoining the protection cost [10]. Zhang et al in 2017 constructed a vigorous model to detect phishing attack. In this they had mined out some semantic features while utilizing word2vec and combined them with multiscale statistical features. The work offered an enhanced system capable of identifying the phishing attempts in real world environment. The results demonstrated that sementic features alone are enough to achieve a high detection accuracy and efficiency of the worked model [11]. In 2018, Sönmez et al. performed a study to conduct an Extreme Learning Machine (ELM) based classification on the data available in UC Irvine Machine Learning Repository database corresponding to phishing websites considered in the analysed 30 features. The efficiency of ELM was evaluated with Support Vector Machine and Naïve Bayes machine learning algorithms. The results showed the system achieved a phishing attack detection accuracy of 95.34% [12]. Jain et al. in 2018 had presented an innovative approach that was based on outstanding hyperlink features that aids in the recognitions of phishing attacks. In the proposed methodology, the authors trained the model on 12 different categories comprising of the important hyperlink features. It offered an effective client side solution and on logistic regression classifiers it exhibited an accuracy of even higher than 98.4% [13].

In the same year Chin et al. presented a new application named PhishLimiter that had the ability to identify and reduce the phishing attacks through emails and web based applications.

In this application deep packet inspection (DPI) and software-defined networking (SDN) techniques were used, where it further consisted of phishing key-identification classification and real time DPI as the two important part of DPI. The authors had tested the application with a dataset of real emails embedded with phishing links and the real world testbed environment [14]. In 2018, Li et al here used the features of URL and HTML webpages to design and deploy a stacking model. This application was proposed as a combination of GBDT, XGBoost and LightGBM in multiple layer architecture in order to enhance power of the application to the identify phishing pages. The proposed approach outperformed the traditional machine learning models that exhibited an accuracy of 94.30%, missing alarm rate of 4.46% and false alarm rate of 1.61% with 50K-PD dataset. When the same dataset is evaluated using proposed model, it demonstrated an improved accuracy of 98.60%, a missing alarm rate of 1.28% and a false alarm rate of 1.54%. Overall, the model showed an improvement in identifying phishing attacks [15].

In 2019, Aksu et al. study was based on deep learning approaches. It utilized the merits of classification models, namely, Support Vector Machine (SVM), neural networks, stacked autoencoders and decision tree to distinguish original websites from phishing websites. The study exemplified that the implication of stacked autoencoders showed a success rate of 86%. It was also concluded that the proposed work had an accuracy of 80% [16]. Lam and Kettani in 2019 proposed the judicial use of PhAttApp, which was an antiphishing application developed against ransomware to minimise the risk of ransomware attack that usually spreads through phishing emails and even by mistakenly visiting an infected website. The PhAttApp was based on machine learning approach that has the ability to detect ransomware from emails and web resources. The application practiced the strengths of psychological-analysis techniques for its functionality [17]. In 2019 Azeez proposed a PhishDetect method to find the reliability between the published web content and the URL (Uniform Resource Locator). The authors also recognized the phishing sites with an aim to provide a defence against the phishing attacks. The results showed the success of PhishDetect that had an ability to provide defence against various phishing attacks [18]. In same year Joshi et al. had proposed the combination of Random Forest algorithm and ReliefF algorithm for feature selection using forward selection approach. The features extracted were related to the web source of the web pages, hence the investigation of the web page was considered more important in phishing attack. The proposed combination has revealed that when the results were evaluated with 10 features, it demonstrated an accuracy of 93.63 % and with 48 features it demonstrated an accuracy of 94.13 % [19].

III. PROPSOED METHODOLOGY

The proposed methodology is divided in two blocks as Similarity Calculator (SC) and Cross Validator (CV) as shown in Fig 2.

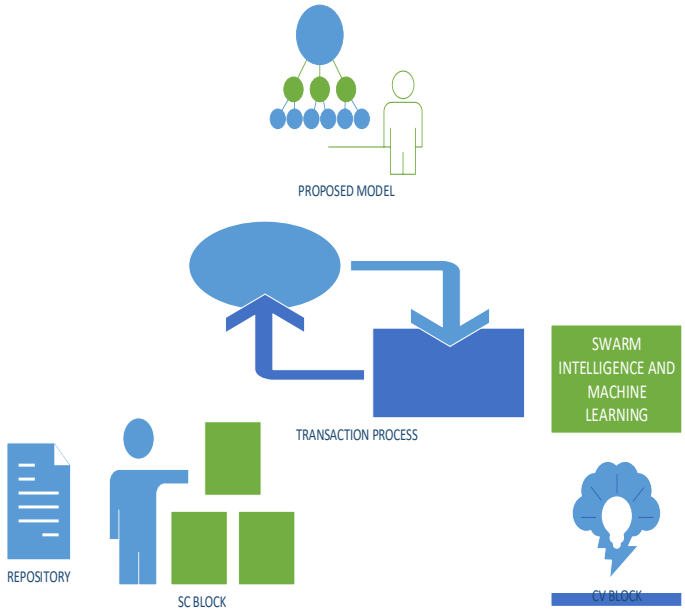

Fig 2: Proposed Block Model

A. SC Block

The SC block is applied for the similarity calculation of the test data and repository data. In order to calculate the similarity between the user query and repository, a hybrid similarity of Cosine and Soft Cosine angular co-relation is proposed. The cosine similarity is the angular relation between passed two vectors. The proposed work model uses the similarity by Pseudo Code 1 as follows.

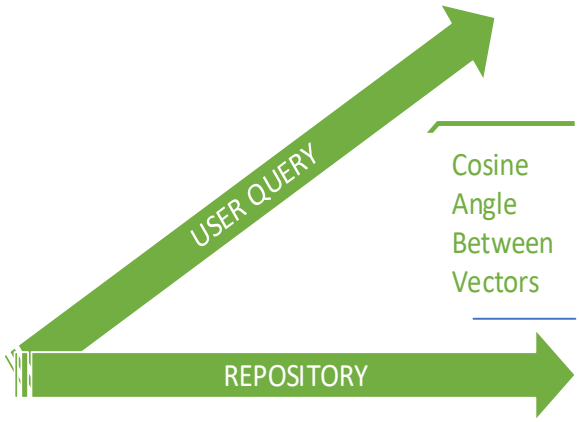

Fig 3: Similarity Calculation

The cosine similarity of two vectors are evaluated using Equation (1)

$$Cosine_{Similarity} = \sum_{i=1}^{n} \frac{User_{Query_i} Repository_{Value}}{\sum_{i=1}^{n} (User_{Query})^2 \sum_{j=1}^{n} (Repository_{Value})^2} \quad (1)$$

PSEUDO CODE 1: Similarity Calculation

1. $For_{each}\ dv\ in\ data_{values}$ // For every data value in the repository
2. $Convert\ dv\ into\ Vector$ // Convert the value after removing stop words from the list
3. $Segregate\ Word\ from\ Repository\ File$
4. $Remove\ Stop\ Words$
5. $Generate\ ASCII\ from\ Each\ Word\ Value$
6. $Calculate\ Similarity\ Using\ Equation\ (1)$ //
7. $Store\ to\ List$
8. End

The proposed algorithm uses the same pseudo code for the calculation of the soft cosine similarity whose mathematical formula is defined in equation (2)

$$SoftCosine_{Similarity} = \frac{\sum_{i,j}^{n} User_{Query_i}\ Repository_{Value_j}}{\sum_{i,j}^{n} User_{Query_i}\ User_{Query_j}\ \sum_{i,j}^{n} Repository_{Value_i} Repository_{Value_j}} \quad (2)$$

The hybrid similarity calculation is made by adding up both the cosine similarity value and the soft cosine similarity value.

$$Hybrid_{Similarity} = Cosine_{Similarity} + Soft\_Cosine_{Similarity} \quad (3)$$

B. CV Block

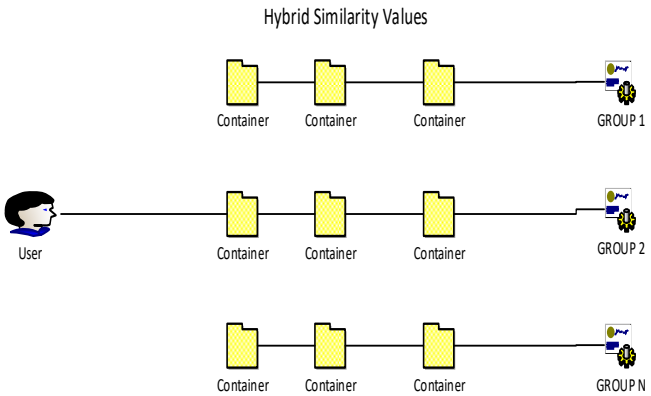

Fig 4: Arrangement of Similarity Structure

The similarity calculated with each repository value is arranged in group. The container contains the similarity values of user data and the repository values and the data is arranged in groups. Each repository container set is tagged as Group. Being the supervised learning mechanism, the entire set is passed to Neural Network which is a multiclass classifier. If the classified value is equal to the group-value, then it is classified as true value else false value. Fig 5 demonstrate the working process of Neural Network. Neural Network is a three layer architecture namely the input layer, hidden layer and the output layer. The output layer produces the classified result based on the weights and architecture processed at the hidden layer. The input layer takes the raw data and passes it to hidden layer through sigmoid function.

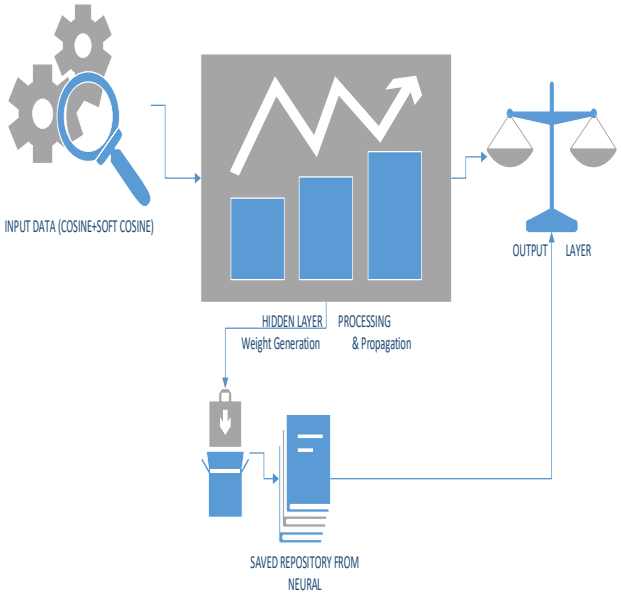

Fig 5: The Architecture of the Neural Networks

PSEUDO CODE NEURAL NETWORK

$Training\_Value =$

1. $Hybrid\ Similarity\ of\ Repository\ Value\ and\ User\ Value$
2. $Group_{Value} = Container\ Number$
3. $Distribution\ Ratio = .70$
4. $Cross_{Validation}\ Ratio = .15$
5. $Test\ Ratio = .15$
6. $Initialize\ Neural\ Network\ with\ Training_{Value}\ and\ Group_{Value}\ with\ 20\ neurons$
   $Train\ Neural\ Network$
7. $Classify\ with\ same\ Training_{Data}$
8. $Foreach\ classified\ Frame$
9. $If\ Classified_{Label}\ Matches\ with\ Training_{Label}$
10. $True_{Classified}\ ++$
11. $Else$
12. $False_{Classified}\ ++$

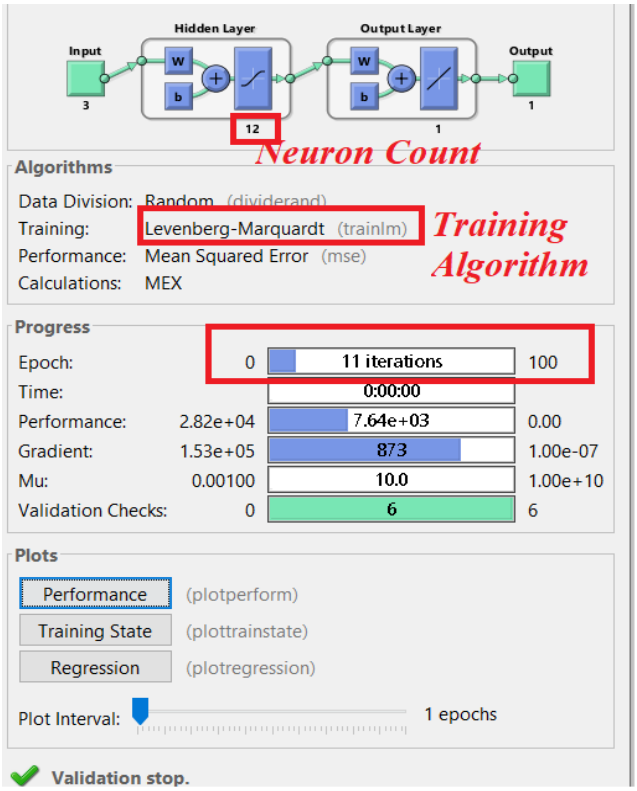

Fig 5: Propagation Structure of Neural Networks

The Neural Networks is organized in such a manner that a supervised classification method is applied over the provided data. If the similarity measure of the provided data, does not matches with the training data value, then the classified data value is said to be non-positive and is considered to be phishy.

C. DATASET:

The proposed work uses the phishtank dataset which is free available data online and is referred by a lot of researchers. The download Uniform Resource Locator (URL) is " https://www.phishtank.com/ " and has following attributes.

a) Phish Id
b) URL
c) Detailed URL
d) Online Status
e) Type of Target

IV. RESULTS

The results are evaluated based on the following evaluation parameters.

a) Precision: It is the ratio of the true detection to the total number of detection in the proposed system.

$$Precision = \frac{tp}{(tp+fp)} \quad (4)$$

b) Recall: It is the ratio of the total number of false detections to the total number of detections.

$$Recall = \frac{fp}{(tp+fp)} \quad (5)$$

c) F measure: It is ratio of the twice of the multiplicative product of precision and recall and the sum of precision and recall.

$$F_{measure} = \frac{2*Precision*Recall}{Precision+Recall} \quad (6)$$

Table-I summarize the results of average precision, recall and f-measure for the number of tested files. A close inspection of the values shows that as the number of tested file increase there is a slight increase in the precision of the results.

Table- I: Average Precision, Recall and F-measure values for Tested Files.

Number of Tested files	Average		
	Precision	Recall	F-measure
100	0.705	0.718	0.712
200	0.709	0.72	0.715
500	0.708	0.719	0.714
1000	0.711	0.724	0.718
2000	0.712	0.725	0.719
3000	0.715	0.726	0.721

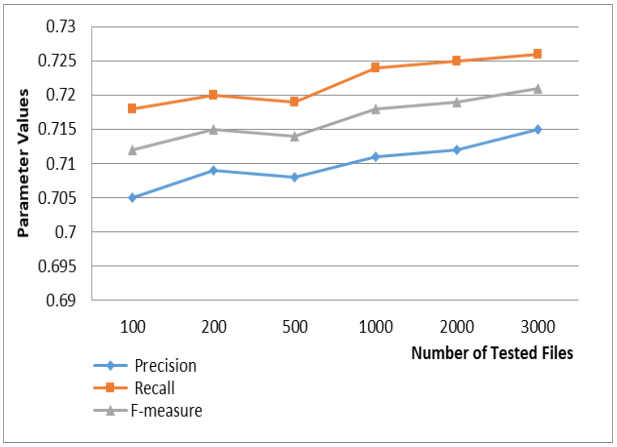

Fig 7: Average Precision, Recall and F-measure for Tested Files

V. CONCLUSION

This paper presents a preventive architecture of social data over internet. The paper introduced a new architecture which is different from traditional architecture of rule-base prevention mechanism. The paper introduced an adaptive learning which uses the forward and backward mechanism of Neural network and utilizes the supervised training mechanism which is based on the similarity values of acting and database similarity value of calculated angular cosine and soft cosine similarity. The evaluation parameters are precision, recall and f-measure. The average precision and recall rate is noted to be .71 to .72. The future aspects of this paper may include the usage of Swarm Intelligence and combination of other machine learning algorithm.

REFERENCES

1. Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L. F., Hong, J., & Nunge, E. (2007, July). Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In Proceedings of the 3rd symposium on Usable privacy and security (pp. 88-99). ACM.
2. Cao, Y., Han, W., & Le, Y. (2008, October). Anti-phishing based on automated individual white-list. In Proceedings of the 4th ACM workshop on Digital identity management (pp. 51-60). ACM.
3. Medvet, E., Kirda, E., & Kruegel, C. (2008, September). Visual-similarity-based phishing detection. In Proceedings of the 4th international conference on Security and privacy in communication netowrks (p. 22). ACM.
4. Arachchilage, N. A. G., & Cole, M. (2011, June). Design a mobile game for home computer users to prevent from "phishing attacks". In International Conference on Information Society (i-Society 2011) (pp. 485-489). IEEE.
5. Ramanathan, V., & Wechsler, H. (2012). phishGILLNET—phishing detection methodology using probabilistic latent semantic analysis, AdaBoost, and co-training. EURASIP Journal on Information Security, 2012(1), 1.
6. Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2013). A survey of phishing email filtering techniques. IEEE communications surveys & tutorials, 15(4), 2070-2090.
7. Solanki, J., & Vaishnav, R. G. (2015). Website phishing detection using heuristic based approach. In Proceedings of the third international conference on advances in computing, electronics and electrical technology
8. Mensah, P., Blanc, G., Okada, K., Miyamoto, D., & Kadobayashi, Y. (2015, November). AJNA: Anti-phishing JS-based Visual Analysis, to Mitigate Users' Excessive Trust in SSL/TLS. In 2015 4th International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS) (pp. 74-84). IEEE
9. Nguyen, L. A. T., Nguyen, H. K., & To, B. L. (2016). An efficient approach based on neuro-fuzzy for phishing detection. Journal of Automation and Control Engineering Vol, 4(2).
10. Arachchilage, N. A. G., Love, S., & Beznosov, K. (2016). Phishing threat avoidance behaviour: An empirical investigation. Computers in Human Behavior, 60, 185-197.
11. Zhang, X., Zeng, Y., Jin, X. B., Yan, Z. W., & Geng, G. G. (2017, December). Boosting the phishing detection performance by semantic analysis. In 2017 ieee international conference on big data (big data) (pp. 1063-1070). IEEE.
12. Sönmez, Y., Tuncer, T., Gökal, H., & Avcı, E. (2018, March). Phishing web sites features classification based on extreme learning machine. In 2018 6th International Symposium on Digital Forensic and Security (ISDFS) (pp. 1-5). IEEE.
13. Chin, T., Xiong, K., & Hu, C. (2018). Phishlimiter: A Phishing Detection and Mitigation Approach Using Software-Defined Networking. IEEE Access, 6, 42516-42531.
14. Jain, A. K., & Gupta, B. B. (2019). A machine learning based approach for phishing detection using hyperlinks information. Journal of Ambient Intelligence and Humanized Computing, 10(5), 2015-2028.
15. Li, Y., Yang, Z., Chen, X., Yuan, H., & Liu, W. (2019). A stacking model using URL and HTML features for phishing webpage detection. Future Generation Computer Systems, 94, 27-39.
16. Aksu, D., Turgut, Z., Üstebay, S., & Aydin, M. A. (2019). Phishing Analysis of Websites Using Classification Techniques. In International Telecommunications Conference (pp. 251-258). Springer, Singapore.
17. Lam, T., & Kettani, H. (2019). PhAttApp: A Phishing Attack Detection Application.
18. Azeez, N. (2019). Identifying phishing attacks in communication networks using URL consistency features.
19. Joshi, A., Pattanshetti, P., & Tanuja, R. (2019, May). Phishing Attack Detection using Feature Selection Techniques. In Nutan College of Engineering & Research, International Conference on Communication and Information Processing (ICCIP).

AUTHORS PROFILE

Bhawna Sharma is currently pursuing Ph.D from Deenbandhu Chhotu Ram University of Science & Technology, Murthal, Sonepat, Haryana,India.. She received her M.Tech. degree in Computer Science & Engineering from ITM University, Gurgaon in 2013. Also, she holds the degree of B.Tech. in Computer Science & Engineering from PDM College of Engineering, Bahadurgarh affiliated to Maharishi Dayanand University, Rohtak. She is working as Assistant Professor in JMIT College, Radaur. Her research interests include cyber security and machine learning.
Email:bhawnash024@gmail.com (Corresponding author)

Dr. Parvinder Singh is Dean and Chairperson in the department of Computer Science & Engineering at Deenbandhu Chhotu Ram University of Science & Technology, Murthal, Sonepat, Haryana,India. He holds Ph.D degree from Maharishi Dyanand University, Rohtak. He pursued M.Tech(CSE) from Guru Jambheshwar Univeristy, Hisar . He pursued B.E.(Electronics) from Baba Saheb Ambedkar Marathwara University, Aurangabad (STB College of Engg, Tuljapur). He has published many research papers in journals of reputed publishers and also associated with review and editing work with many journals. Email: parvinder23@rediffmail.com

Dr. Jasvinder Kaur is working as Assistant Professor in the department of Computer Science & Engineering at PDM University, Bahadurgarh, Haryana, India. She did Ph. D. from DCRUST, Murthal. Information Security and Information Hiding are her areas of specialization. Email: jasvinder.kaur@pdm.ac.in

Dr. Pablo García Bringas is Executive Master in Business Administration by Deusto Business School, and Doctor Computer Engineer, specialized in artificial intelligence application to the field of cyber security. He is a professor of engineering at the University of Deusto. He is currently deputy dean of external relations, training, and research. He was founder and director of the new Chair Deusto in Digital Industry. Email: pablo.garcia.bringas@deusto.es