International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Extreme Learning Model Based Phishing Classifier

Praveen Tumuluru, Radha Manohar Jonnalagadda, Divya Sai Sree Konatham, Vineetha Samineni, Burra Lakshmi Ramani

Abstract: ***Phishing Is An Act Of Attempting To Acquire The Users' Data Such As Usernames, Passwords And Credit Card Details As It Was The Trustworthy Entity In An Electronic Communication. Because Of The Quick Development Of The Internet, Clients Change Their Inclination From Customary Shopping To The Electronic Business. Rather Than Bank Or Shop Robbery, These Days Culprit Attempt To Discover Their Victims In The Internet With Some Particular Tricks. By Utilizing The Mysterious Structure Of The Internet, The Culprits Set Out New Strategies, For Example, Phishing, To Betray Victims With The Utilization Of Fake Webpages To Gather Their Delicate Data, For Example, Account Ids, Usernames, Passwords, And So On. Understanding Whether A Website Page Is Genuine Or Phishing Is A Very Testing Issue, Because Of Its Semantics-Based Assault Structure, Which Predominantly Misuses The Pc Users' Susceptibilities. Despite The Fact That Most Of The Software Companies Introduce Many Anti-Phishing Products, Which Use Blacklist Generator, Heuristic Approach And Ml-Based Methodologies, These Products Can't Stop All Of The Phishing Attacks.***

The Main Objective Of This Paper Is To Find An Efficient Approach For Distinguishing The Phishing Sites Which Depends On The Extreme Learning Model. Specifically, The Proposed Method Computes Some Features As Input And Checks Whether The Given Url Is Phishing Url Or The Legitimate Url. The Proposed Extreme Learning Model Attains 97% Accuracy Rate For Detection Of Phishing Urls And If The Hidden Layers Increases The Accuracy Is Also Discussed.

Keywords: Phishing Classification, Url, Neural Networks, Extreme Learning Model, Classification, Back Propagation Algorithm.

Manuscript published on November 30, 2019.
* Correspondence Author
Praveen Tumuluru*, Assistant professor, Department of CSE, Koneru Lakshmaiah College of Engineering, KLEF, Guntur, Andhra Pradesh, India.
Radha Manohar Jonnalagadda, Student, Department of CSE, Bachelor of Technology, Koneru Lakshmiah College of Engineering, KLEF, Andhra Pradesh, India.
Divya Sai Sree Konatham Student, Department of CSE, Bachelor of Technology, Koneru Lakshmiah College of Engineering, KLEF, Andhra Pradesh, India.
Vineetha Samineni Student, Department of CSE, Bachelor of Technology, Koneru Lakshmiah College of Engineering, KLEF, Andhra Pradesh, India.
Burra Lakshmi Ramani, Asst professor, Computer Science and Engineering Dept., PVP Siddhartha Institute of Technology, Kanuru, Vijayawada, India.

I. INTRODUCTION

Phishing indicates to the act that the attacker allure the users to visit fraud/fake webpage by sending them fraud e-mails/ any other means, and quietly gets the victim's personal data like usernames, passwords, and other secured information. This data at that point can be utilized for future objective commercials or even diagnose the fraud assaults like money transfer and so forth. The regularly used attack technique was to send e-mails to potential victims, which appeared to be sent by banks, online associations, or ISPs. In these e-mails, they will make up certain causes, such as the password of their credit card had been wrongly-entered for many number of times, or they are providing updating services, to ask urgently you to visit their webpage to affirm or alter their account number and password through the hyperlink which was given in the email. In that case, if the user inputs the account number and password, the assailants at that point effectively gather the data at the server side, and can play out their following stage activities with that data.

Phishing itself is definitely not a new idea, yet it's undeniably utilized by phishers to steal the users' data and perform business crimes in the present-day. Inside one to two years, the quantity of phishing assaults expanded drastically. As indicated by a study by Verizon about 30% of the phishing emails that are sent are opened by the clients and 12% of the clients did tapped the tainted attachment.

Phishing attacks utilize different methods; the most widely recognized mean is email. Attacks are on the ascent and are turning out to be increasingly refined making them harder to distinguish. The annual economic cost of phishing may exceed $3 billion dollars. In addition to direct financial loss, also consider the costs to restore damaged reputations, repair security breaches and regain customer confidence, and given the regulatory environment, no company wants to consider the potential audit findings for failing to address phishing attacks. Phishing is extremely powerful and utilizes a procedure called social engineering. This system misuses regular human trust. The criminal claims to be somebody or an element you may trust and takes the client certifications like usernames, passwords or other verified data. This kind of phishing URLs can be distinguished by utilizing different supervised and unsupervised methods like Artificial Neural Network, K-Nearest Neighbor, Random Forest, Naive Bayes, and Support Vector Machine etc.

The primary goal of this paper is to analyse the URL as safe or legitimate, and find if the hidden layers in the neural network increases then the accuracy increases or decreases? In this paper, we have used Extreme Learning Model to classify the phishing URLs. We have taken a dataset which has data of some features like Address Bar based characteristics, Abnormal based characteristics, HTML and JavaScript based characteristics and Domain based characteristics as input, processing will be done in the hidden layer and at last delivers the output whether the given URL is phishing URL or legitimate URL. And work is done in checking the changes in the accuracy rate changes by increasing the hidden layers in the Extreme Learning Model. There is a minimal raise and fall of accuracy rates with the change in the number of hidden layers.

II. MOTIVATION

The literature analysis of the work done in this section

Literature Survey:

Amani Alswailem et al. [1] utilized supervised learning technique for example Random Forest algorithm to distinguish phishing URLs. Because of its great execution in classification they have chosen this procedure. The primary target of the paper is to seek after a better classifier by examining the highlights of phishing site and pick the better mix of them to prepare the classifier. Subsequently, the paper accomplishes the exactness of 98.8% by utilizing the mix of 26 features. Yasin Sonmez et al. [2] utilized Extreme Learning Model based grouping with 30 features using Phishing Websites Data in UCI Glocbal Machine Learning Repository database. For results, ELM was compared with other AI techniques, Support Vector Machine (SVM), Naïve Bayes (NB) and identified to have the most elevated precision of 95.34%. The proposed arrangement model is utilized so as to order of the phishing features. This technique comprises of feature extraction from sites and grouping area. Here certain standards are characterized in order to use for phishing highlight extraction in acquiring highlights. So as to grouping of these highlights, SVM, NB and ELM were utilized. In the ELM, 6 diverse initiation capacities were utilized and ELM accomplished most exactness score.

Wenchuan Yang et al.[3] plans a Convolutional Gated-Recurrent-Unit (CGRU) neural framework for the acknowledgment of malevolent URLs revelation subject to characters as substance classification features. Considering that noxious catchphrases are stand-out to URLs, a component depiction methodology for URLs reliant on malignant watchwords is proposed, and a gated irregular unit (GRU) is used rather than the first pooling layer to perform highlight securing on the time estimation, achieving high-accuracy multicategory results. Test outcomes show that the proposed neural framework area model is altogether proper for high exactness classification assignments. Contrasted and other classification models, the model exactness rate are above 99.6%. The utilization of deep learning out how to order URLs to distinguish Web guests' aims significant hypothetical has and scientific esteem for Web security inquire about, giving new plans to shrewd ideas for security identification. Amirreza Niakanlahiji et al. [4] proposed PhishMon, another component rich AI structure to perceive phishing site pages. It relies upon a great deal of fifteen novel highlights that can be efficiently enlisted from a site page without requiring pariah organizations, for instance, web records, or WHOIS servers. These highlights get various properties of authentic web applications similarly as their essential web establishments. Replicating of these features is costly for phishers as it solicitations to contribute significantly more vitality and effort on their essential establishments and web applications; despite the endeavors required for rehashing the nearness of target locales. Through expansive appraisal on a dataset involving 4,800 undeniable phishing and 17,500 specific kind pages, it shows that PhishMon can perceive covered phishing from genuine site pages with an exceptionally elevated level of accuracy. The exploratory results show that PhishMon achieved 95.4% precision with a 1.3% counterfeit positive rate on a dataset containing wonderful phishing events.

Shraddha Parekh et al.[5] exhibited URL recognition technique utilizing Random Forest classification algorithm with the assistance of Rstudio. There are 3 significant stages, Heuristic-Classification, Parsing and Performance Analysis in this model and each stage utilizes an alternate system or calculation for preparing of information to give better outcomes. Here, a few highlights were exactly exhibited out of 31 of them are the most reasonable for identification of phishing sites. This model has utilized a wide scope of measurements, including the F-measure, ROC, precision, exactness, and affectability for examination purposes along these lines giving a reasonable view on the presentation and precision each time the identification happens. There is no single answer for phishing till now and with the up and coming innovation, the sort and number of phishing assaults are relied upon to increment. For these, the programs must be made proficient enough to arrangement techniques that identify and caution of potential phishing assaults.

Abdulhamit Subasi et al. [6] proposed different data mining procedures to perceive classes of locales whether they are phishing or valid destinations. Different classifiers were used to construct a precise sharp system for phishing site disclosure. Portrayal Classification exactness, the zone under recipient working trademark (ROC) bends (AUC) and F-measure is used to survey the display of the data mining frameworks. Results showed that Random Forest has defeated best among the portrayal methodologies by achieving the most raised precision 97.36%. Arbitrary woods runtimes are exceptionally snappy, and it can oversee different destinations for phishing identification.

Surbhi Gupta.et.al. [8] proposed a phishing destinations classifier using improved polynomial neural frameworks in hereditary calculation. Individuals by bringing blind data into copy sites. The fundamental objective of these sites is to assault our classified data, for example, passwords, banking subtleties or other individual data and usernames. Shraddha Parekh et al. [9] proposed a model as a response for perceive phishing destinations by using the URL recognizable proof methodology using Random Forest computation to show signs of improvement results and proposed to recognize phishing locales by utilizing

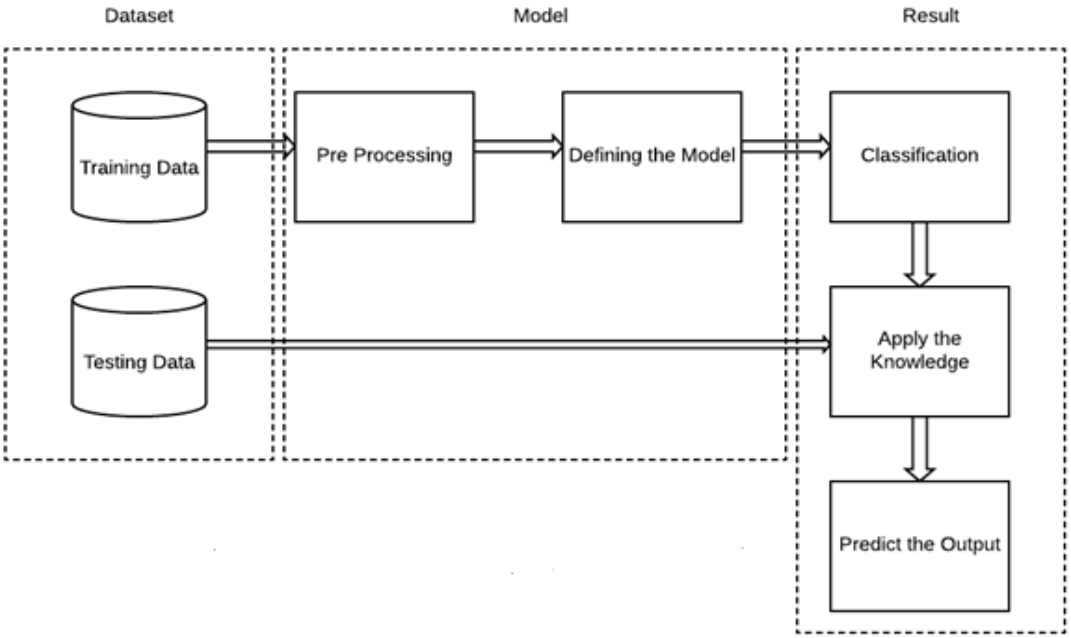

Fig 1: Schematic block diagram for the proposed system

random forest as the characterization calculation with the assistance of Rstudio.

III. PROPOSED MODEL FOR THE CLASSIFICATION

The goal of this paper is to develop a model which can classify the URL based on given characteristics and make it ready for deployment in any of the cloud server. The proposed classification model is based on Extreme Learning, which has an Input Layer, a Hidden Layer and an Output Layer. The problem is identified as a binary classification problem. In the pre-processing stage the classes in the dataset is mapped as per the binary classification, followed by the check for null values. The model has an input layer, a hidden layer and an output layer. For activation either linear or non-linear activation functions can be used.

Consider a URL, for the model to classify the URL, the features of the URL are to be identified and are to be given to the model. The model then classifies the features of the URL as phishing or legitimate. Given below is the schematic

block diagram of the proposed model. The dataset has been taken from UCI Machine Learning Repository. It has 30 features of the URL and 11,000 documents classified as phishing or legitimate.

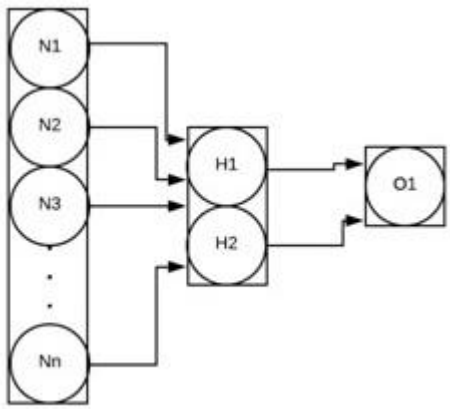

Fig 2 : Model Architecture

Dataset:

The dataset for this work has been taken from UCI Machine Learning Repository. It has 30 features and an output column.

[1] having_IP_Address
[2] URL_Length
[3] Shortining_Service
[4] having_At_Symbol
[5] double_slash_redirecting, Prefix_Suffix
[6] having_Sub_Domain
[7] SSLfi nal_State
[8] Domain_registeration_length
[9] Favicon
[10] port
[11] HTTPS_token

[12] Request_URL
[13] URL_of_Anchor
[14] Links_in_tags
[15] SFH
[16] Submitting_to_email
[17] Abnormal_URL
[18] Redirect
[19] on_mouseover
[20] RightClick
[21] popUpWidnow
[22] Iframe
[23] age_of_domain
[24] DNSRecord
[25] web_traffic
[26] Page_Rank
[27] Google_Index
[28] Links_pointing_to_page
[29] Statistical_report

Architecture of Extreme Learning Model:

The architecture of the learning model is a Neural Network with an Input layer, a hidden layer and an output layer. Preprocessed data is given as input to the input layer, where the layer fires the neurons with the given activation function and passes the data to the hidden layer and the data will be propagated to the output layers. The output layer classifies the given url,

Training Phase of Extreme Learning Model:

The training in the MLP is based on adam optimizer. Adam is an adaptive learning rate method. It uses square of the gradients to scale the learning rate. Adam optimizer uses first and second moments of gradient to adapt the learning rate for individual weights.

$$m_n = E[X^n]$$

To estimate the moments, adam optimizer uses exponentially moving average, computed using the gradient evaluated on minor batch

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^{\,2}$$

The less first values of the estimate's gradient contribute to the overall value. The formula for the moving average will be

$$m_t = (1-\beta_1)\sum_{i=0}^{t} \beta_1^{t-i} g_i$$

The estimator needs to be corrected, so that the expected value is reached. Bias correction is the key step in the training stage.

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

The final step is to scale the learning rate for each parameter by using the moving averages. To perform the update of the weight we will use,

$$w_t = w_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

API

For the model to be accessed with ease, an API has been developed, where the model is saved in the disk and the API locates the model when it is routed correctly.

Once the model is loaded, the inputs are taken from the API using POST method. The features of the url will be provided in JavaScript Object Notation(JSON), and that is converted into an array of 30 features and given to the model for prediction. The predictions are stored in an array and returned back in JSON format.

The use of this API is that, it can be hosted in any of the cloud servers and the model can be used in any of the platform, which can process JSON encryption.

To API has been tested using an open source platform called POSTMAN, which accepts all types of network requests and process them.

Flow Chart for the Classification

The figure given below is the flowchart of the training the model and predicting the output using the API developed. If the model is not located or the model has not been trained, then the model is trained and then stored in the disk. The server instance cannot be started until unless the saved trained model with the modified synaptic weights.

If the model is located, then the local server instance is started and the API is ready for predicting the output. If a POST request is sent, the API gets fired and return the result. Since the API is started in local host, postman is used to send and receive the data.

The API returns a JSON object which contains the predicted values as 0 or 1.

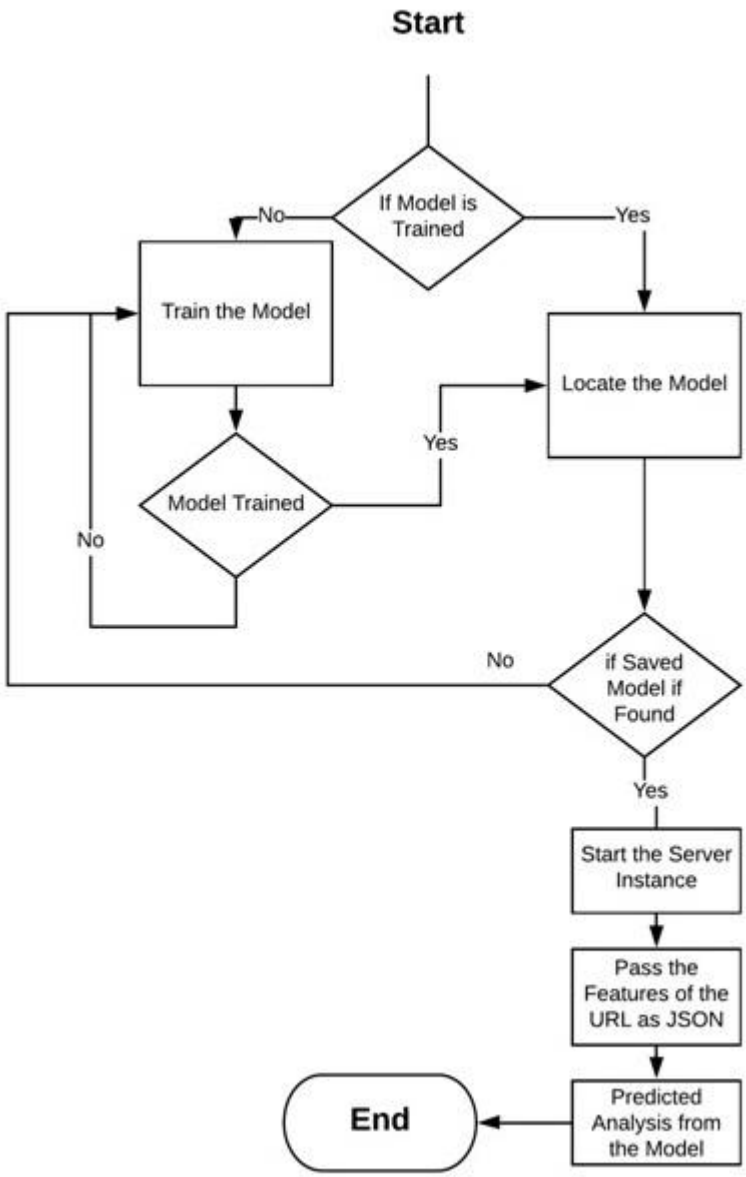

Fig 3: Flowchart for classification

IV. RESULTS AND DISCUSSIONS

This section deals with the results of the proposed model and the analysis with respect to the existing models and reveals the effectiveness of the proposed model.

Experimental Setup:

The proposed extreme learning model is implemented using Python and the application programming interface is implanted using Flask framework of Python, in a computer that operates in macOS Catalina Operating system with 8GB of memory.

Accuracy:

It is the degree of correctness of the classification as given as,

$$Accuracy = \frac{(TP) + (TN)}{(TP) + (TN) + (FP) + (FN)}$$

Where TP is true positive, TN is true negative, FP is false positive, FN is false negative.

Proposed model got at accuracy of around 97% in classifying the data

Competing methods:

The performance of the proposed model is compared with Naïve Bayes Classifier[NB], Random Forest Algorithm[RF], Support Vector Machine[SVM] and Extreme Learning Model[ELM].

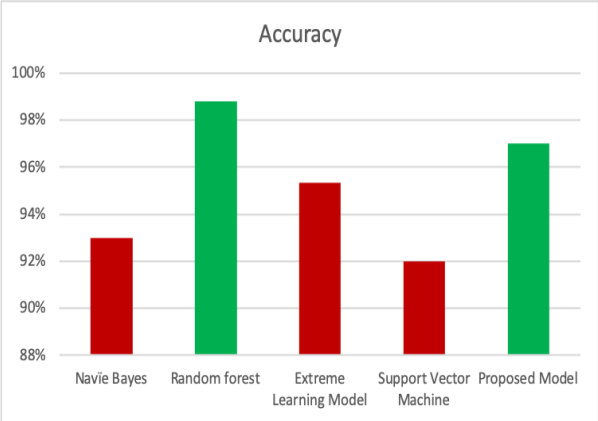

Fig 4: Comparative Analysis

Dataset Used:

The dataset considered for analysis is includes websites dataset which comprises of 11,000 unique rows of 30 features. It has 2 output classes.

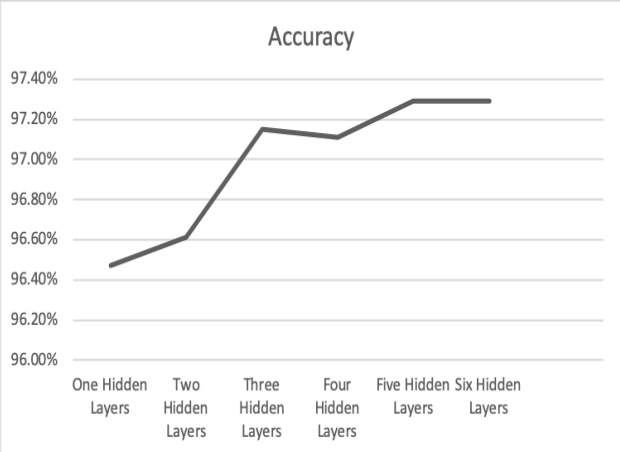

Fig 5: Accuracy change with increase in Hidden Layers

Performance Analysis using comparative methods:

In this investigation, we received a classifier model that is utilized for recognizing phishing sites in an astute and robotized way. An insightful framework to distinguish phishing assaults is exhibited. The proposed model chooses the classifications of sites: authentic or phishing. We have contrasted various classifiers and our proposed model to build an exact clever framework for phishing site identification. Results demonstrated that our proposed technique and Random Forest have outflanked best among the characterization strategies by accomplishing the most elevated precision 97% or more.

The proposed model runtimes are quite fast, and it can deal with different websites for phishing detection.

Method of URL Classification		Accuracy
Websites Dataset	Random Forest	98.80%
	Extreme Learning Model	95.34%
	Naïve Bayes	93.00%
	Support Vector Machine	92.00%
	Proposed Model	97.00%

Table 1: Comparative Discussion of the Methods

Being a Extreme Learning Model, it has varying number of hidden layers in the model. The above graph indicates the effect of the number of hidden on the accuracy of the prediction of the model. The change in the accuracy is slight and the change is neither increasing nor decreasing.

V. CONCLUSION

Phishing has turning into a genuine system security issue, causing finical misfortune of billions of dollars to the web based business organizations. What's more, maybe more in a general sense, phishing has made online business doubted and less appealing to typical purchasers. In the report, the qualities of the hyperlinks that were inserted in phishing messages are referenced. We at that point planned an enemy of phishing calculation, in view of the determined qualities. Since, Phishing Website is trademark based, it can identify known assaults, however additionally is successful to the obscure ones. We have executed Feed Forward Neural Network Algorithm, our test indicated that Feed Forward Neural Network is light-weighted and can identify up to 97% obscure phishing assaults progressively. We accept that Feed Forward Neural Network isn't helpful for recognizing phishing assaults, yet additionally can shield clients from malevolent or spontaneous connections in Web pages and Instant messages. The analysis using the websites dataset on the proposed model overcomes the existing Extreme Learning Model in terms of accuracy that is 97%. The proposed model achieves higher classification accuracy.

REFERENCES

1. Amani Alswailem, Bashayr Alabduillah, Norah Alrumayh and Dr.Aram Alsedrani, “Detecting Phishing Websites Using Machine Learning,” 2019 IEEE.
2. Yasin Sonmez,Turker Tuncer,Huseyin Gokal and Engin Avc, “Phishing Web Sites Features Classification Based on Extreme Learning Machine,” 2018 IEEE.
3. Wenchuan Yang,Wen Zuo and Boajiang CUI, “Detecting Malicious URLs via a Keyword-based Convolutional Gated-recurrent-unit Neural Network,” 2018 IEEE.
4. Amirreza Niakanlahiji, Bei-Tseng chu and Ehab Al-Shaer, “PhishMon: A Machine Learning Framework for Detecting Phishing Webpages,” 2018 IEEE.
5. Shraddha Parekh, Dhwanil Parikh, Srushti Kotak and Smita Sankhe, “A new method for Detection of Phishing Websites: URL Detection ,” Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - ISBN:978-1-5386-1974-2.
6. Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi and Touseef J.Chaudhery, “Intelligent Phishing Website Detection using Random Forest Classifier,” 2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA)
7. S.Gupta and A.Singhal, “Phishinsg URL detection by using Artificial Neural Network with PSO,” 2nd International Conference on Telecommunication and Networks 2017.
8. Gayathri. S , “Phishing websites classifier using polynomial using neural networks in genetic algorithm,” 2017 4 the International Conference on Signal Processing, Communications and Networking (ICSCN -2017), March 16 – 18, 2017, Chennai, INDIA
9. Alejandro C.B and E.C. Bohorquez, “Classifying Phishing URL Using Recurrent Neural Networks”, 2017.

10. Abdulghani Ali Ahmed and Nurul Amirah Abdullah, “Real Time Detection of Phishing Websites,” 2016 IEEE.
11. Priyanka Singh, Yogendra P.S. Maravi, and Sanjeev Sharma, “Phishing Websites Detection through Supervised Learning Networks,” 2015 IEEE.
12. Luong Anh Tuan Nguyen, Ba Lam To, Huu Khuong Nguyen, and Minh Hoang Nguyen, “An Efficient Approach for Phishing Detection Using Single-Layer Neural Network,” The 2014 International Conference on Advanced Technologies for Communications(ATC 14).
13. Dona Abraham and Nisha S Raj, “Approximate String Matching Algorithm for Phishing Detection,” 2014 IEEE.
14. Kannan,S, R.Rakesh,A, Muthuraj, L.SaiRamesh, and V.Pandiyaraju “Enhancing the Precision of Phishing Classification Accuracy using Reduced Feature Set and Boosting Algorithm,” 2014 Sixth International Conference on Advanced Computing.
15. Samuel Marchal, Jérôme François, Radu State, and Thomas Engel, “PhishStorm: Detecting Phishing With Streaming Analytics,” IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, VOL. 11, NO. 4, DECEMBER 2014.
16. McCluskey.L, Mohammad.R.M and Thabtah.F, “Intelligent rule-based phishing websites classification,” IET , vol. 8, no. 3, 2014.
17. R. M. Mohammad, F. Thabtah, and L. McCluskey, “An assessment of features related to phishing websites using an automated technique,” Internet Technol. ..., pp. 492–497, 2012.
18. Rungsawang.A and Sanglerdsinlapachai.N, “Using domain top-page similarity feature in machine learning-based web phishing detection,” International Conference on Knowledge Discovery and Data Mining, 2010, WKDD.
19. Varjani.A.Y, and Moghimi.M “New rule-based phishing detection method,” Expert Syst. Applications., volume 53, 2016.
20. Faruk Erturul.O and Kaya.Y, “Adetailed analysis on extreme learning machine and novel approaches based on ELM,” Am. J. Computer. Science. Engineering, volume. 1, no. 5, 2014.
21. Praveen Tumuluru , "A Framework for Identifying of Gene to Gene Mutation causing Lung Cancer using SPI - Network", vol. 152, no. 10, International Journal of Computer Applications, October 2016.
22. Praveen Tumuluru, "Credentials of Lung-Cancer Associated Genes Using Protein-Protein Interaction Network", Vol. 6, No. 3, International Journal of Advanced Research in Computer Science and Software Engineering, March 2016.
23. Lakshmi Ramani Burra and Padmaja Poosapati, A Study of Notations and Illustrations of Axiomatic Fuzzy Set Theory, International Journal of Computer Applications Volume 134 – No.11, (0975 – 8887) January 2016.
24. Praveen Tumuluru, "Dijkstra’s based Identification of Lung Cancer Related Genes using PPI Networks", Vol. 163, No. 10, pp. 1-10(0975 – 8887), International Journal of Computer Applications , April 2017.
25. Praveen Tumuluru, "A Survey on Gene Expression Classification Systems", International Journal of Scientific Research and Review, Volume 6, Issue 12, ISSN NO: 2279-543X, 2017.
26. Tumuluru Praveen, B. GOA-based DBN: Grasshopper Optimization Algorithm-based Deep_Belief_Neural Networks for Cancer Classification, 2017, International Journal of Applied Engineering Research, 2017.
27. Lakshmi Ramani Burra and Padmaja Poosapati Adaptive Lion_Fuzzy_System to Generate the Classification Rules using Membership_Functions based on Uniform-Distribution, , Volume 12, (24) ISSN 0973-4562, 2017, International Journal of Applied Engineering Research, 2017
28. Lakshmi Ramani Burra and Padmaja Poosapati, A literature study of fuzzy rule based classifiers, International journal of scientific research and review , ISSN NO: 2279-543X, 2017.
29. Lakshmi Ramani Burra and Padmaja Poosapati, Adaptive Fuzzy System with Robust GSCA-based Fuzzy-Rule-Extraction for Data Classification, Vol. 10, Issue-1, 2018, Jour of Adv Research in Dynamical & Control Systems, 2018
30. Tumuluru Praveen and Bhramaramba Ravi, Chronological Grasshopper_Optimization Algorithm-based Gene-Selection and Cancer-Classification, No. 3, Vol. 10, 2018, Journal of Advanced Research in Dynamical & Control Systems, 2018
31. Burra Lakshmi Ramani, Padmaja Poosapati, Praveen Tumuluru “Deep Learning and Fuzzy Rule-Based Hybrid Fusion Model for Data Classification”, Volume-8 Issue-2, ISSN: 2277-3878, International Journal of Recent Technology and Engineering (IJRTE), July 2019.
32. Praveen Tumuluru, Burra Lakshmi Ramani, Open CV Algorithms for facial recognition, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-8, ISSN: 2278-3075 June, 2019.
33. Bhargav, K., Vijay Kumar, G, Sreedevi, M & Mohan Krishna, P. (2018). Incremental mining of popular patterns from transactional databases. International Journal of Engineering and Technology(UAE-2018), 7, 636-641.
34. Vijayakumar & Vinothkanna, R, & T. (2019). Using contourlet transform based RBFN classifier for face detection and recognition-2019, doi:10.1007/978-3-030-00665-5_176