EXTREME LEARNING MACHINE ALGORITHM FOR PHISHING WEBSITES DETECTION

[1]Pediredla Swamy, [2]Dr. Kunjam Nageswara Rao, [3]Dr. G. Sita Ratnam & [4]Kalidindi Venkateswara Rao

[1]M. Tech Scholar, [2]Professor, [3]Associate Professor, [4]Research Scholar

[1,2, 4]Department of Computer Science & Systems Engineering (A), Andhra University, Visakhapatnam.

[3]Department of CSE, LENDI Institute of Engineering and Technology, Vizianagaram.

Abstract: Phishing is one in all the foremost common and dangerous attacks among cybercrimes. The victim's confidential knowledge is predicted by the phishing sites by derivation them to surf a phishing web site that resembles to legitimate web site, that is one in all the criminal attacks prevailing within the net. Phishing websites is comparable to cyber threat that's targeting to induce all the credential-based data like data accessed from the credit cards and Social Security numbers. The aim of this project is to perform Extreme Learning Machine (ELM) primarily based classification for thirty options as well as Phishing Websites knowledge in University of California (UC) Irvine Machine Learning Repository information. There are differing kinds of options supported sites. Hence, to forestall phishing attacks there is tendency to use a particular web content feature. The project model supports machine learning techniques like Naïve Bayes, Linear Discriminant Analysis (LDA) and Support Vector Machine to discover phishing sites. For results assessment, ELM was compared with different machine learning ways like Support Vector Machine (SVM), Naïve Bayes (NB), Linear Discriminant Analysis (LDA) and detected to possess the 96.74% accuracy.

Index Terms: Extreme Learning Machine, Classification, data Security, Phishing

I. INTRODUCTION

Internet use has become a vital a part of our daily activities as a result of rapid growing technology. The rapid growth of technology and intensive use of digital systems, data security of these systems has gained importance. The first objective of maintaining security in data technologies is to confirm that necessary precautions are taken against threats and dangers probably to be long-faced by users throughout the utilization of those technologies. Phishing is outlined as imitating reliable websites so as to get the proprietary data entered into websites on a daily basis for numerous functions, like usernames, passwords and citizenship numbers. Phishing internet sites contain numerous hints among their contents and web browser-based data. Individual(s) committing the fraud sends the pretend web site or e-

mail data to the target address as if it comes from a corporation, bank or the other reliable supply that performs

reliable transactions. Contents of the web site or the e-mail embrace requests targeting to lure the people to enter or update their personal data or to alter their passwords moreover as links to websites that seem like same copies of the websites of the organizations involved.

Phishing is a serious threat to both users and enterprises; numerous anti-phishing techniques have been developed we have a tendency to examined phishing websites and extracted options of those websites. Tips relating to the extracted options of this information are given below. Within the initial section we have a tendency to outlined rules and that we gave equations of internet options. These equations are required so as to clarify phishing attacks characterization.

II. LITERATURE SURVEY

Phishing may be a serious threat to each users and enterprises; various anti-phishing techniques are developed. In general, the techniques are classified as either List-based or Heuristics-based technologies. List-based techniques maintain a blacklist or whitelist or each. Tons of anti-phishing ways use a blacklist to forestall users from accessing phishing sites. These techniques follow a quantitative approach for evaluating the phishing chance of a given web site victimization the refined security risk parts for domain and web content. Style and implementation of the web site risk assessment system for anti-phishing also are enclosed (Young-Gab 2012). Heuristic-based mechanisms use many criteria to work out whether or not an internet {site a web site} may be a phishing site or not (Chun-Ying et al 2011). The CAPTCHA authentication application designed in a cost-effective mode protects the protection unconscious user by sanctionative safe on-line banking authentication, thereby addressing the web banking threats. finding the final cognitive content of security warning moreover as making certain safe on-line banking authentication even on a compromised host are the prime challenges of a secure on-line industry. The projected hardware solutions don't seem to be possible for the house users because of its usurious price (Leung 2013). Many industrial digital forensics code suites are obtainable for examining digital media associated with laptop crimes. Though these tools offer examiners with in depth capabilities for rhetorical examinations, they will have vital drawbacks in terms of coaching, initial prices of the tool, and yearly maintenance upgrades or else, there are Free and Open supply code (FOSS) tools with equivalent practicality that examiners will use to perform most of an equivalent task's potential by industrial applications (Philili et al 2007).

URL Verification**:**URL obfuscation has become a key trick among all the tricks utilized in phishing activities. Therefore, equipping the user by making awareness regarding the obfuscated URLs and the way to work out truth nature of the strange URLs is that they would like of the hour (Ed Skoudis 2012). Universal Resource Locator analysis involves all the formats of the universal resource locator being analyzed however largely the links associated with login info solely (Debra et al 2009). Whereas the distinction and similarities between the URLs are known with the assistance of the trustworthy website, usually the Link Guard rule

is applied for analyzing the common characteristics and hyperlinks (Juan and Guo 2006). Content matching techniques and DNS queries are accustomed determine the malicious URLs aside from mistreatment the regular expressions and hash maps for analytic the Symantec variations (Pawan et al 2010). Phish want isn't mistreatment any coaching any coaching and whitelist or blacklist (Debra et al 2008). Page Safe could be a tool that's supported user input to seek out the legitimacy of the universal resource locator (Sengar and Vijay 2010).

Registering an analogous domain to trick the user into a fallacious website is changing into common by mistreatment the @symbol for redirection. For e.g., within the case of web.paypal.com@123.123.123.123 the user should feel that they're visiting the positioning web.paypal.com, however truly being directed to a website with 123.123.123.123 because the information science address. Therefore, checking the URLs for any special characters gains the importance currently (Chun-Ying et al 2011).

The following points were thought-about within the Phish Market Model (Tyler and Tal 2010).

- Only specific URLs requested by the receiving party are shared.
- Providing party wasn't given info regarding the URLs, that was given to the receiving party.
- The range of URLs given to the receiving party is tallied in an exceedingly secured manner.
- URLs happiness to the receiving party don't seem to be counted.

In the anti-phishing toolbars, the programmed ought to be properly designed to provide the warnings and also the choices for customizing the settings by the user. the most parts within the anti-phishing shopper aspect applications include; Main programmed, essential Warnings & facilitate Systems (Linfeng and Marko 2007).

III. METHODOLOGY

Procedural steps for solving the classification problem presented is as follows:

Identification of the problem This study attempts to solve the problem as to how phishing analysis data will be classified datasetApproximately 11,000 data containing the 30 features extracted based on the features of websites in University of California Irvine Machine Learning Repository database.ModelingAfter the data is ready to be processed, modelling process for the learning algorithm is initiated. The model is the construction of the need for output identified in accordance with the task requirements.

A. Classification

Classification is to work out the category to that every information sample of the ways belongs, that ways are used once the outputs of computer file are qualitative. The aim is to

divide the whole downside house into an exact range of categories. A good vary of classification ways are gift. This can be thanks to the actual fact that completely different classification ways are made for various information as there's no excellent technique that works on each information set. As mentioned in literature studies, the aim of classification is to assign the new samples to categories by mistreatment the pre-labelled samples. The foremost normally used classification ways are delineated below.

•Support Vector Machine (SVM) are supervised learning models with associated learning algorithms that analyze information used for classification and multivariate analysis. It presents one amongst the foremost sturdy prediction ways, supported the applied math learning framework. SVM coaching rule builds a model that assigns new examples to at least one class or the opposite, creating it a non-probabilistic binary linear classifier.

•Naive Bayes (NB) classifiers are a family of easy "probabilistic classifiers" supported applying theorem with study independence assumptions between the options. they're among the only theorem network models, however plus Kernel density estimation, they will win most accuracy levels. Naïve Bayes classifiers are extremely ascendible, requiring {a range variety} of parameters linear within the number of variables in an exceedingly learning downside.

• Linear Discriminant analysis (LDA)is discriminant operate analysis could be a generalization of Fisher's linear discriminant, a way utilized in statistics, pattern recognition, and machine learning to seek out a linear combination of options that characterizes or separates or additional categories of objects or events. It's used as a spatial property reduction technique. conjointly referred to as a normally utilized in the pre-processing step in machine learning and pattern classification.

Extreme Learning Machine (ELM) could be a feed-forward artificial neural network (ANN) model with one hidden layer For the ANN to confirm a high-performing learning, parameters like threshold worth, weight and activation operate should have the suitable values for the information system to be modelled. In gradient-based learning approaches, all of those parameters are modified iteratively for acceptable values. Thus, they will be slow and manufacture low-performing results thanks to the probability of obtaining stuck in native minima. In ELM Learning Processes, otherwise from ANN that renews its parameters as gradient-based, input weights are indiscriminately chosen whereas output weights are analytically calculated. As associate degree analytical learning method considerably reduces each the answer time and also the probability of error worth obtaining stuck in native minima, it will increase the performance quantitative relation. So as to activate the cells within the hidden layer of ELM, a linear operate still as non-linear (sigmoid, Gaussian), non-derivable or separate activation functions will be used.

Extreme learning machines are feedforward neural networks for classification, regression, clustering, distributed approximation, compression and have learning with one layer or multiple layers of hidden nodes, wherever the parameters of hidden nodes needn't be tuned.

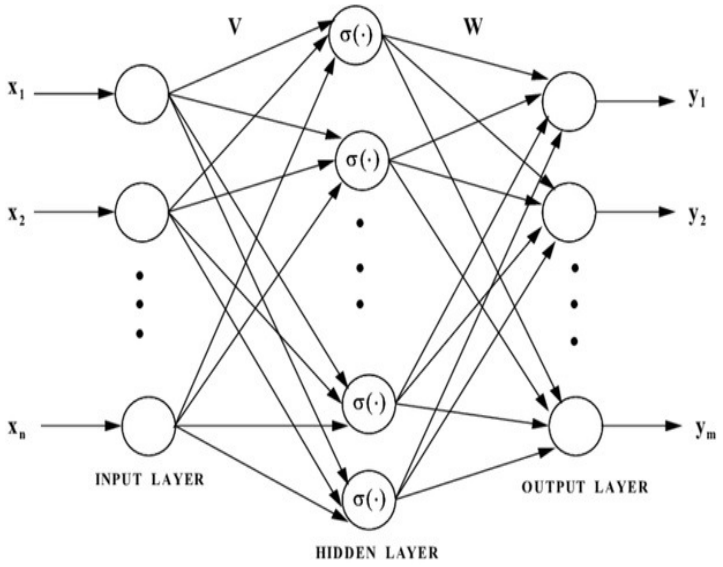

Fig.1. An artificial neural network model with a single hidden layer with forward feed.

As shown in Fig.1. x_1,x_2, …. xn are selected attributes (or) features after feature selection are given as inputs to hidden layer, where hidden layer consists of function which is indicated by σ (.) from which the predicted outputs are given as y_1, y_2…..y_n.

Model performance evaluation

The topics addressed in this section are the two measures that affect the performance of the model and the algorithm used, the first one being the division of data set into training and test data set and the second one being the definition of expressions measuring the performance. In the first measure, the data set is divided into three parts as training, validation and test data by three-phase division in K-Fold method, and model selection and performance status are simultaneously performed. In the second measure, performance assessment of classifier models generally uses a validation value. Validation value can be measured as the ratio of data count detected or estimated correctly by the algorithm into all data in the data set.

IV.RESULTS

Achieved performance of Extreme Learning Machine(ELM), Support Vector Machine (SVM), Naive Bayes (NB), Linear Discriminant Analysis (LDA) are presented in Table1. As deduced from these data, ELM achieved higher performance compared to other methods in terms of performance and speed.

$$\boldsymbol{y(p) = \sum_{j=1}^{m} \beta j \; g(\sum_{i=1}^{n} \mathbf{Wi, j \;\; Xi + bj})}$$

In the above equation, $\mathbf{X_i}$ refers to input vector and y(p) refers to output vector (m and n neuron count), $\mathbf{W_{i,j}}$ indicates input layer to hidden layer weights and j indicates output layer to hidden layer weights, bj represents the threshold value of neurons in the hidden layer and

g(.) represents activation function. Input layer weights (w) and bias (bj) values in the equation are randomly assigned. Activation function (g(.)), input layer neuron count (n) and hidden layer neuron count (m).

Algorithm	Accuracy (%)
ELM	96.74
SVM	94.47
NB	60.70
LDA	91.46

Table.1 Accuracy Comparison of Algorithms.

In this study, features in the database created for phishing websites are classified by determining the input and output parameters for the ELM classifier. Results obtained by ELM show that ELM has higher achievement compared to other classifier (SVM, LDA and NB) methods. This study is considered to be an applicable design in automated systems with high performing classification against the phishing activity of websites. Furthermore, in literature comparisons, this study is observed to high performance of 96.74% that is also the highest test performance (i.e., both precision and accuracy).

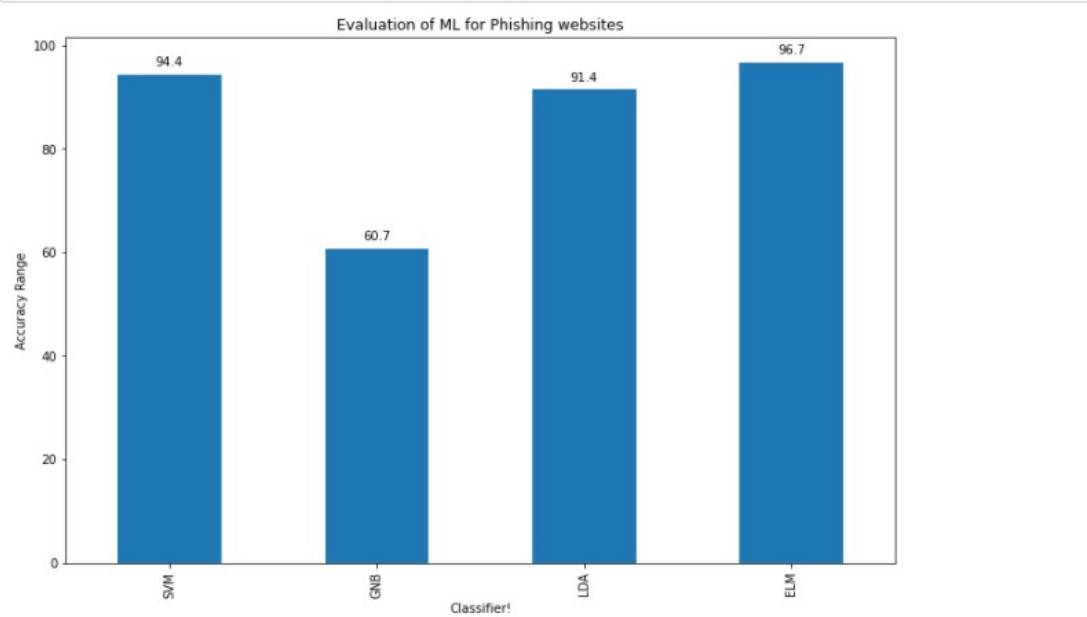

Fig.2. Accuracycomparison graph for SVM,NB, LDA AND ELM algorithms.

The above graph represents the performance of Extreme Learning Machine(ELM), Support Vector Machine (SVM), Naïve Bayes (NB), Linear Discriminant Algorithm (LDA). ELM achieved higher performance compared to other algorithms.

V.CONCLUSION

In this paper, features of phishing attack are defined and proposed a classification model in order to classification of the phishing attacks. This method consists of feature extraction from websites and classification section. In the feature extraction, we have clearly defined rules of phishing feature extraction and these rules have been used for obtaining features. In order to classification of these features Support Vector Machine (SVM), Naïve Bayes (NB), Linear Discriminant Analysis (LDA) and Extreme Learning Machine (ELM) were used. In the ELM, 6 different activation functions were used and Extreme Learning Machine algorithm achieved highest accuracy score.

VI. REFERENCES

[1]. P. Ying and D. Xuhua, "Anomaly based web phishing page detection, "in Proceedings - Annual ComputerSecurity Applications Conference, ACSAC, 2006, pp. 381–390.

[2]. M. Moghimi and A. Y. Varjani, "New rule-based phishing detection method," Expert Syst. Appl., vol. 53, pp. 231–242, 2016.

[3]. DATASET: Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

[4]. G.-B. Huang et al., "Extreme learning machine: Theory and applications," Neurocomputing, vol. 70, no. 1–3, pp. 489–501, 2006.

[5]. C. S. Guang-bin Huang, Qin-yu Zhu, "Extreme learning machine: A new learning scheme of feedforward neural networks," Neurocomputing, vol. 70, pp. 489–501, 2006.

[6]. T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to Spam filtering," Expert Systems with Applications, vol. 36, no. 7. pp. 10206–10222, 2009.

[7]. Ö. F. Erturul, ArÖrenmeMakineleriilebiyolojiksinyalleringizlikaynaklarnaayrtrlmas. D.Ü. MühendislikDergisiCilt: 7, 1, 3-9-2016

[8]. W. D. Yu, S. Nargundkar, and N. Tiruthani, "A phishing vulnerability analysis of web-based systems," IEEE Symp. Comput. Commun. (ISCC 2008), pp. 326–331, 2008.

[9]. P. Ying and D. Xuhua, “Anomaly based web phishing page detection,” in Proceedings - Annual Computer Security Applications Conference, ACSAC, 2006, pp. 381–390.

[10]. M. Moghimi and A. Y. Varjani, “New rule-based phishing detection method,” Expert Syst. Appl., vol. 53, pp. 231–242, 2016.

[11]. DATASET: Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

[12]. G.-B. Huang et al., “Extreme learning machine: Theory and applications,” Neurocomputing, vol. 70, no. 1–3, pp. 489–501, 2006.

[13]. C. S. Guang-bin Huang, Qin-yu Zhu, “Extreme learning machine: A new learning scheme of feedforward neural networks,” Neurocomputing, vol. 70, pp. 489–501, 2006.

[14]. T. S. Guzella and W. M. Caminhas, “A review of machine learning approaches to Spam filtering,” Expert Systems with Applications, vol. 36, no. 7. pp. 10206–10222, 2009.

[15]. Ö. F..Ertu÷rul, AúÕrÕÖ÷renmeMakineleriilebiyolojiksinyalleringizlikaynaklarÕnaayrÕútÕrÕlmasÕ . D.Ü. MühendislikDergisiCilt: 7, 1, 3-9- 2016 [16] M. E. Tagluk, M. S. Mamiú, M. Arkan, and Ö. F. Ertugrul, “AúiriÖgrenmeMakineleriileEnerjiIletimHatlari Ariza Tipi veYerininTespiti,” in 2015 23rd Signal Processing and Communications Applications Conference, SIU 2015 - Proceedings, 2015, pp. 1090– 1093.

[17]. Ö. Faruk Erturul and Y. Kaya, “A detailed analysis on extreme learning machine and novel approaches based on ELM,” Am. J. Comput. Sci. Eng., vol. 1, no. 5, pp. 43–50, 2014.

[18]. Ö. F. Ertugrul, “Forecasting electricity load by a novel recurrent extreme learning machines approach,” Int. J. Electr. Power Energy Syst., vol. 78, pp. 429–435, 2016.

[19]. G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: Theory and applications,”*Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[20]. N. Sanglerdsinlapachai and A. Rungsawang, “Using domain top-page similarity feature in machine learning-based web phishing detection,” in 3rd International Conference on Knowledge Discovery and Data Mining, WKDD 2010, 2010, pp. 187–190.

[21]. N. Sanglerdsinlapachai and A. Rungsawang, “Using domain top-page similarity feature in machine learning-based web phishing detection,” in 3rd International Conference on Knowledge Discovery and Data Mining, WKDD 2010, 2010, pp. 187–190.

[22]. S. V. N. Santhosh Kumar, Yogesh Palanichamy, Energy efficient and secured distributed data dissemination using hop by hop authentication in WSN, Wireless Networks, vol.24, pp.1343-1360, 2018.

[23]. Patrick Lawson, Carl J. Pearson, Aaron Crowson, Christopher B. Mayhorn, Email phishing and signal detection: How persuasion principles and personality influence response patterns and accuracy, Applied Ergonomics, Elsevier, vol. 86, pp. 1-10, 2020.