International Journal of Recent Research and Review, Vol. XII, Issue 3, September 2019
ISSN 2277 – 8322

Detection of Phishing Hybridization of Naive Bayes Classifier with Artificial Intelligence

Gyan Kamal[1], Krishna Gupta[2]
[1]*Scholar, Deptt.of Computer Science Engg., Yagyavalkya Inst. of Technology, Jaipur, India*
[2]*Asst. Professor, Deptt. of Computer Science Engg., Yagyavalkya Inst. of Technology, Jaipur, India*

***Abstract* - Internet has become an indispensible avenue of modern life style. Most of the work has transferred in the digital realm. Online banking, e-mail, social media & messaging apps are the norm of the day. With these applications, people are becoming more & more dependent on the internet & private, confidential & financial data is stored & transacted vide various internet servers multiples times daily. This situation, opens on opportunity for hacker's to steal & misuse, personal information, e-mail accounts, banking passwords. Hackers use multiple tools & attack techniques to achieve the same, one of the prominently used techniques is phishing attacks, in which user is lured to a malicious web page as app & is fooled to enter his/her login & password. There has-been a lot of work done previously in phishing attack detection & prevention especially detection of phishing websites. The author has proposed a unique method of combining Naive Bayes Classifier with Artificial Intelligence. The work is aimed at classification of phishing website using Naive Bayes classifier, artificial neural network & its comparison to the hybrid method. The hybrid method employs the Weighted Average of Artificial Intelligence & Naive Bayes Classifier to achieve at a more accurate result.**

***Keywords*–Detection of Phishing, Naïve Bayes Classifier, Artificial Intelligence, Artificial Neural Network.**

I. INTRODUCTION

The Internet (Internet port) is a general system that connects computer networks and connects to public devices using the Internet Protocol Packet (TCP / IP). Private, public, academic, commercial and government networks are interconnected with a variety of local and global electronic, wireless and fiber optic network technologies. The Internet provides a wide variety of information sources and services, including interconnected hypertext documents, World Wide Web (WWW) applications, email, phone and file sharing. This originates from the Internet to the 1960s, when the federal government to US conducted research on computer networks to build powerful communications and unstable. This is the first claim of the needs of academic as the backbone of the ARPANET of the war in the 1980s.. The National Science Foundation Network, a new private fund backbone in the 1980s and other private funds, allowed the world to develop new network technologies and to participate in the integration of many networks. In the early 1990s, the link between business networks and businesses marked the beginning of the transition to modern Internet, and several generations of organizations, individuals and mobile computers were connected to the network and continued to grow exponentially. The Internet is widely used in academia since the 1980s, but also services and Technology, which is a trade integrated into all aspects of modern life. The Internet (Internet port) is a general system that connects computer networks and connects to public devices using the Internet Protocol Packet (TCP / IP). Private, public, academic, commercial and government networks are interconnected with a variety of local and global electronic, wireless and fiber optic network technologies. The Internet provides a wide variety of information sources and facility, including interconnected hypertext documents, World Wide Web (WWW) applications, email, phone and file sharing.

II. PROPOSED WORK

1. Design & Development of a Phishing Attack Website or App Detection System Using Hybridization of Modern Classifiers & Artificial Intelligence.

2. Use of advanced classification techniques such as Random Forest, Naive Bayes to adapt to growing anti-detection methods & techniques.
3. Use of Function Fitting Neural Network to estimate Phishing Risk based on URL features extracted using Levenberg Marquardt Algorithm.
4. Improvement In Accuracy of Prediction of phishing Risk Score by Combination of Neural Network & Classifiers by employing averaging, mean, weighted average, standard deviation etc.
5 Optionally, the system can be provided with access to Internet, to enable it to extract Page Rank or Google index automatically from a Input URL.
6 Optionally, the developed system can be modified to perform lexical analysis of the URL entered & compute the URL parameters itself instead of manually entering the same.

III. LITERATURE REVIEW

In this paper, a different approach is to use a random forest as a classification algorithm to detect phishing sites with the help of Rstudio. Here, our experience confirms that 31 of the proposed features are best suited for removing phishing sites. Performance indicators and our literature surveys also show that random forests have a 95% accuracy level, so random forests are selected for classification [10]. The model uses a wide range of indicators including true positive, true negative, false negative, F, ROC, accuracy and analytical sensitivity to clearly understand performance and accuracy in each test. . So far, there is no single phishing solution, and with the upcoming technology, the type and number of phishing attacks is expected to increase. For this, the browser must be powerful enough to set up a way to detect and warn of potential phishing attacks. Future work is aimed at developing a system that can detect new phishing attacks on its own by adding more enhancements to the detection process [1].
In this paper, we combine the prediction results generated by various feature classifications and the hierarchical clustering algorithm to classify phishing, and propose a smart phishing website detection framework. Empirical studies of the large and actual daily data sets collected by Kingsoft Internet Security Lab show that our approach performs well in phishing detection and classification. Our IPDCM is integrated with the SIAT Internet Security software product [2].
Phishing has become a serious cybersecurity issue that costs consumers and e-commerce companies billions of dollars. Perhaps more fundamentally, it makes e-commerce unreliable and unattractive to normal consumers. In this research report, we examine the properties of hyperlinks embedded in phishing emails. Then, we designed a Link-Guard anti-phishing algorithm based on derived features. Because phishing protection is feature-based, it not only detects known attacks, it also protects against unknown attacks. We LinkGuard implemented in Windows XP. Our experiments show that it is able to detect lightweight and LinkGuard to 96% of phishing attacks in real time is unknown. We believe he feels LinkGuard phishing attacks, but also protects users from malicious links or unsolicited web pages for instant messages. Our future work contains an algorithm LinkGuard further expanding, to be able to handle CSS (cross-site scripting) attacks [3]. The main security issue for banks and financial institutions is phishing. Phishing is a cyber attack that uses a strategy and imitation from an unauthorized individual or organization to impersonate a customer's web service. It is not allowed to steal personal information (such as bank details, social security numbers and credit card details) through the mouth to provide it to the public network. Hand, to provide users with confidential information, you are not ignorant of the website and to use it is a phishing website. This article describes a method for removing phishing attacks and detection of phishing sites by integrating source code and URL into web pages [4]. Deceptive websites are designed to mimic the look of the company's web pages. Phishing attackers test users by using various social engineering techniques, for example, if they do not complete the account update process, threatening to suspend user accounts, providing additional information to verify their account or other reasons for users to access their fraudulent web pages. Why is it important to solve the problem phishing? According to data from the Anti-Phishing Royal Society of London, in March 2006, 18,480, and the attacks of unique phishing sites (9) 666 unique phishing related. Phishing attacks affect millions of Internet users and are

a huge cost burden for victims and victims phishing (phishing 2006). A study conducted in April 2004 Gartner found that false information from a site that is in direct losses to banks and credit card issuers US $ 1.2 billion (Litan 2004). Phishing has become a major threat to user's developer [5]. Recently, limited anti-phishing activities have given phishers the opportunity to prevent their advanced fraud. In addition, failure to design appropriate classification methods to effectively identify these deceptive behaviors has reduced the visibility of phishing sites. Therefore, the use of new; ethnic minorities; predictions; the most effective functionality has been a major challenge in maintaining test flexibility. Therefore, some of the previous work has been completed to develop and apply some of the selected methods to develop their own classification methods. However, there is no research in general that considers which feature selection method can be used as the best assistant to improve classification performance [6]. We suggest a machine learning method that uses the functionality of the X.509 public key certificate to detect phishing sites. We have proven that this is better than sites that support HTTPS. Our solution immediately identifies phishing sites, which is an important addition to the existing anti-phishing-based mechanism that blacklisting primarily uses. Blacklists have some inherent shortcomings in terms of accuracy, timeliness and integrity. Because there may be significant lags before the site's blacklist, the attacker has a window of opportunity. The advantage of public key certificate classification is that it relies less on the central server than on the blacklist [7]. In this paper, we introduce a new method of data analysis is based on the wrong phishing site is detected as legitimate servers. According to this, idea is the finding by reference. For this reason citations, and a time to every victim opens a phishing site to site identify a website asking for legal resources. From all this we find with phishing website. In this paper, author introduce a new method of phishing detection based on legitimate web server log information [8]. The use of technical skills and social technology to exploit the innocence of unconscious users to take advantage of phishing is a crime. The procedure typically involves trusted entities to influence the consumer to take actions when investigated by the simulated entity. In most cases, users will notice phishing attacks, but security is the main motivation for key users because they do not understand the situation. However, some methods are limited to focusing only on phishing attacks, and discovery is mandatory. In this article, we will focus on different strategies for eliminating phishing attacks. In other words, we can also say that other electronic transactions will also be part of the threat. Since then, it has been recommended to deal with these issues in good faith before the madman attacks. You must obtain an order to protect all important online banking activities [9]. A single solution does not solve the phishing problem. In this case, phishers are always trying to come up with a new model of consumer manipulation. Online shoppers should conduct periodic risk assessments to discover the latest technologies that can lead to thriving phishing attacks. To find a safer approach, users must be aware of the dangers of advanced malware currently occurring. In addition, the regulatory team will need to implement advanced techniques that may ultimately prevent advanced threats to their predictable resentment [10]. In this paper, author first extract vocabulary and host-based features from the phishing website example, and then we develop a system to apply sample categorization or phishing categorization, sharing some common features by kernel k-means. family. in the Clustering method. The large and standard data sets collected from our data center and Phish load of our system will work well for phishing site feature classification applications. With this technology, as opposed to aggregation-based technology, the accuracy of classification and error rate of phishing website samples has been improved. The accuracy and error rate will increase the classification of phishing sites by 10% to 20%. In future work, you can first capture three aspects of your work through different phishing methods, then there are many clustering algorithms that can be applied to malware and phishing sites, and the third may include of phishing detection anomalies [11].

Phishing is a fraudulent method used to trick users into obtaining personal information such as usernames, passwords, credit cards, and bank account information on the Internet. The key to phishing is deception. Phishing uses spoofing email as the initial medium for

fraudulent communication, and then fraudulent websites get the information they need from the victim. Phishing was discovered in 1996 and is now one of the most serious cybercrime cases facing Internet users. Researchers are studying the prevention, detection, and education of phishing attacks, but to date, there are no complete and accurate solutions to stop them. This paper reviews, analyzes and analyzes the most important and novel approaches proposed in the field of phishing website discovery, and outlines their advantages and disadvantages. In addition, a detailed review of the latest scenarios presented by various sub-categories researchers is provided [12].

IV. METHODOLOGY

(A) ANN Training Flow Chart

Initialize matrix test with values of 8 Nos URL parameters and matrix test1 with corresponding values for Phishing Score Predicted. Input = transpose of test and targets = transpose of test1. Initialize hidden layer size to 10. Define a function fitting neural network of hidden layer size. Initialize network input, Process remove constant rows, mapminmax network. Initialize network parameters, Evaluate output, error and performance, compute all parameters. View the network at display plot regression.

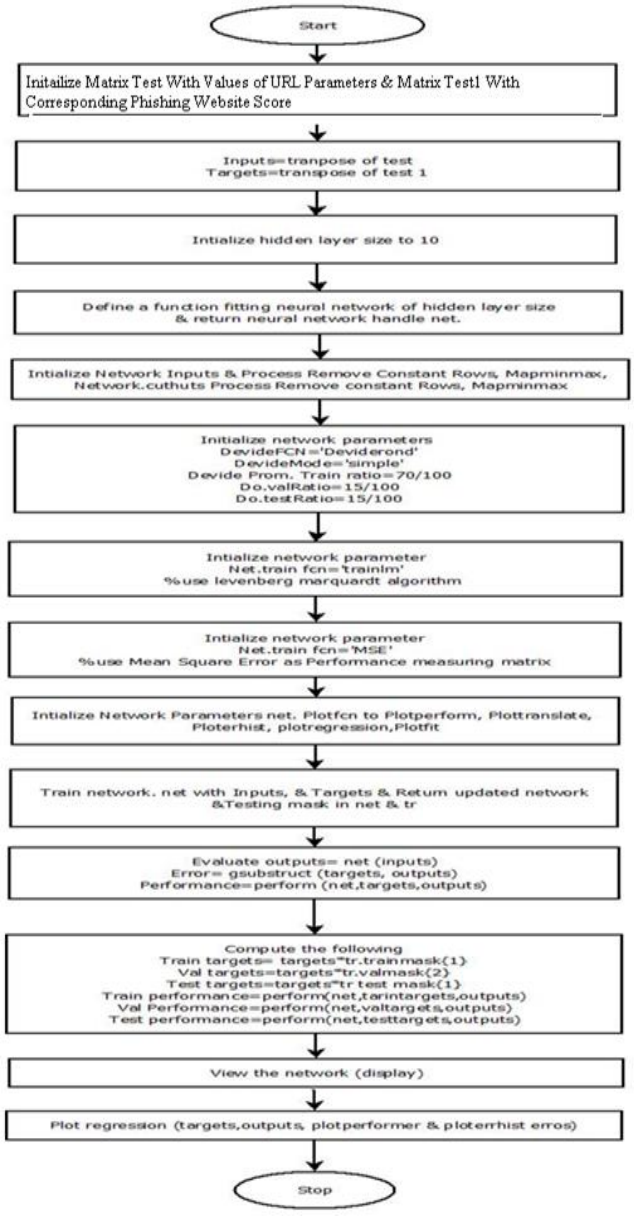

Fig. 1 ANN Training Flow Chart

(B) Naive Bayes Algorithm Training

In below flow chart we load ('Input_CNB.mat'); & ('ouput_CNB.mat'); mdl input and dummy = input (CNB Classifier initialized) and stop.

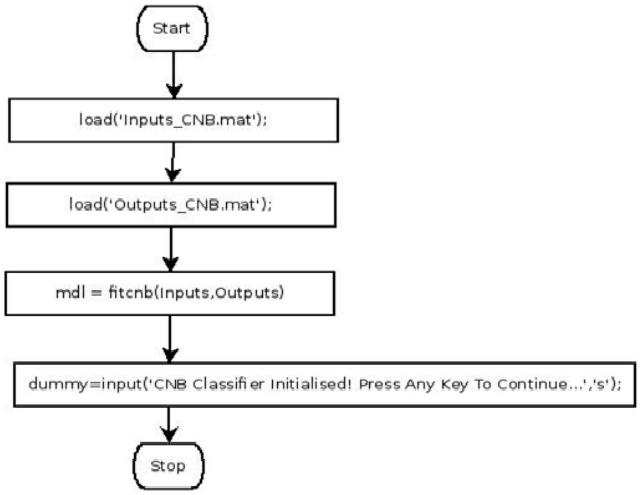

Fig. 2 Naive Bayes Algorithm Training

(C) ANN Phishing Score Prediction

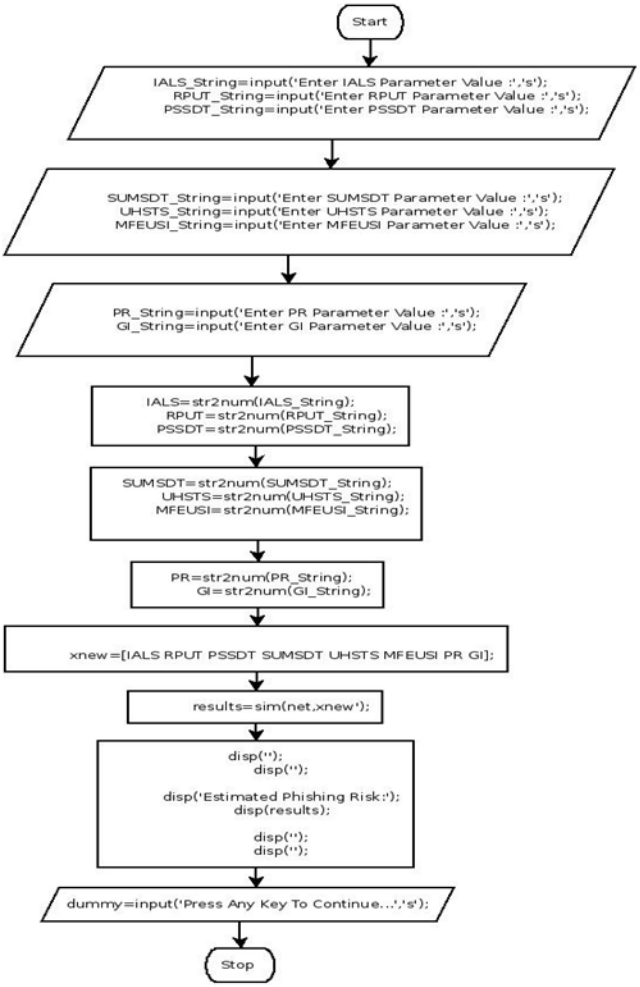

Fig. 3 ANN Phishing Score Prediction

(D) Naive Bayes Phishing Score Prediction

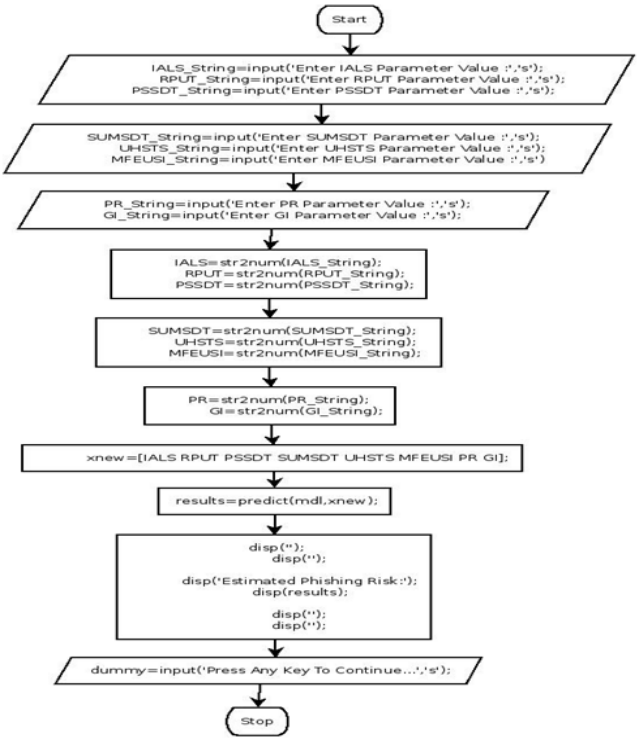

Fig. 4 Naive Bayes Phishing Score Prediction

V. RESULTS

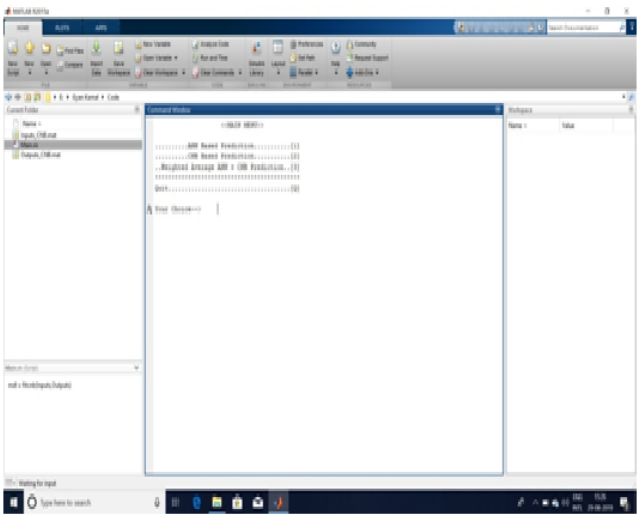

Fig. 5 Main Menu

This figure depicts the main menu of the software developed. All three operations supported are available on menu.

The menu organization is as follows:

1. ANN Based Prediction
2. CNB Based Prediction
3. Weighted Average ANN + CNB Prediction

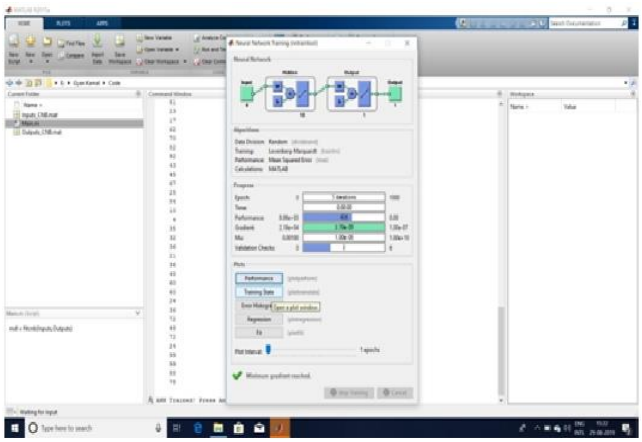

Fig.6 Neural network training tool

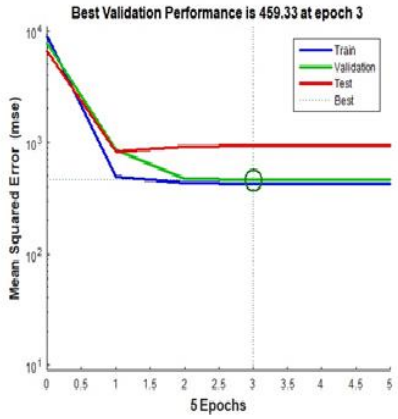

Fig. 7 Best Validation Performance

This figure depicts the best validation performance plotting of Mean square Error of Validation Data Vs Epochs & finding lowest MSE at an Epoch. The best validation performance is 459.33 at epoch 3.

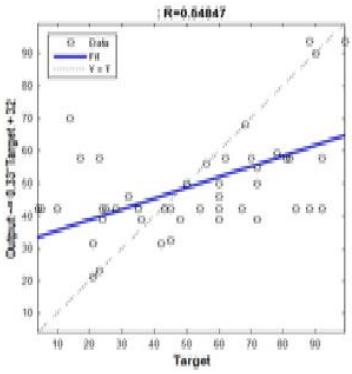

Fig. 8 Regression Analysis

This figure depicts the regression of Output + Target Vs Input as inform of I = Mx + C, it tries to approximate the prediction function out come as a straight line. The regression line that best approximate the prediction function is given as. R = 0.54847.

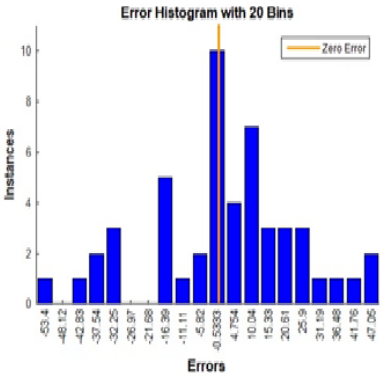

Fig. 9 Error Histogram

This figure depicts the Error Histrogram of the trained Neural Network by plotting graph between Error Instances Vs No.of Error

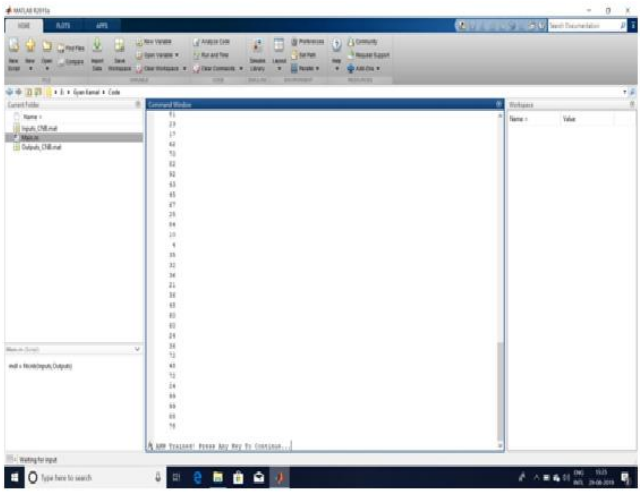

Fig. 10 Enter Value ANN Trained

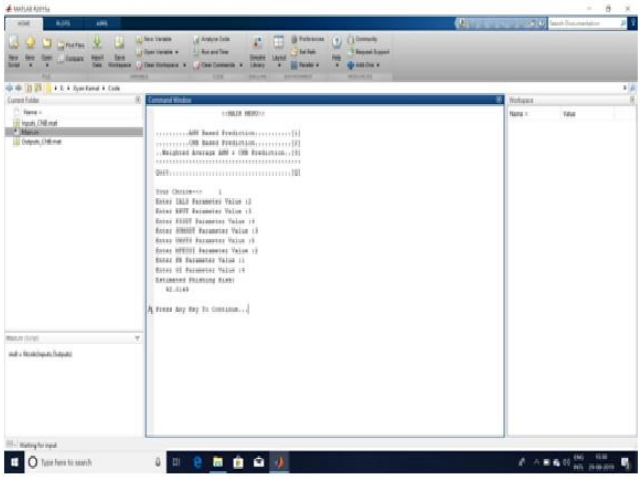

Fig. 11 Estimating Phishing Risk through ANN based Prediction

In above figure we input the parameter like, IALS, RPUT, PSSDT, SUMSDT, UHSTS, MFEUSI, PR & GI and their respective values. And get estimated Phishing Risk : 92.0149.

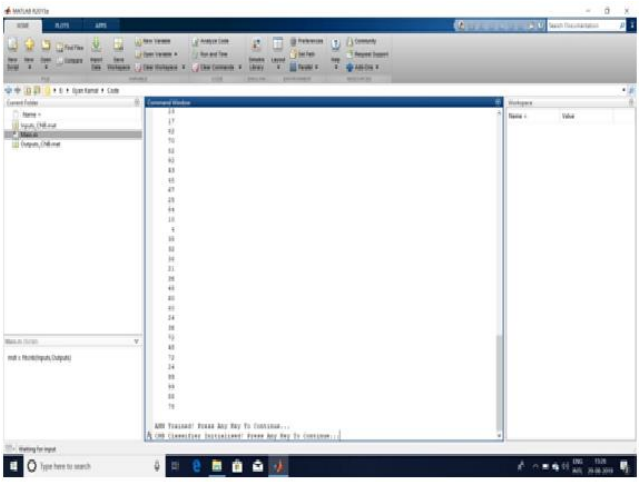

Fig. 12 CNB Classifier initialized

In above figure we see that after ANN Trained the CNB Classifier is initialized.

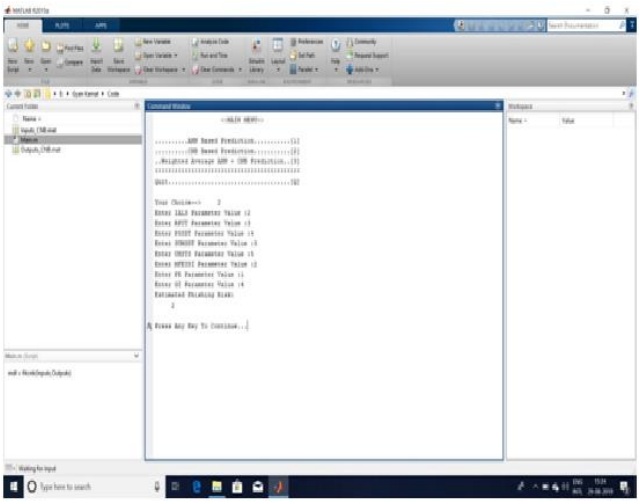

Fig. 13 Estimating Phishing Risk through CNB based Prediction

In above figure, we input the parameter like, IALS, RPUT, PSSDT, SUMSDT, UHSTS, MFEUSI, PR & GI and their respective values to get estimated Phishing Risk :2.

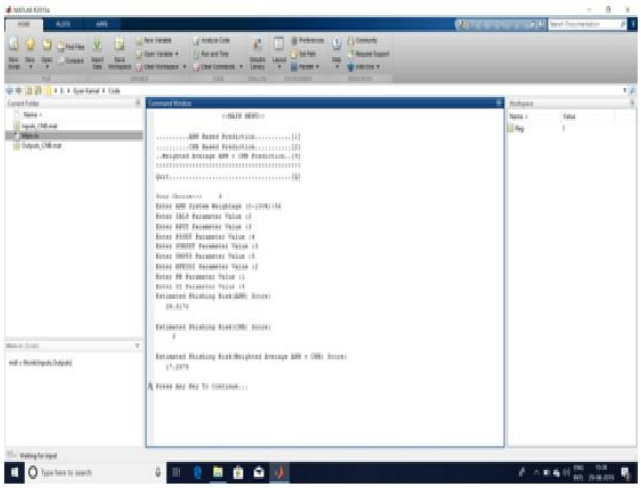

Fig. 14 Estimating phishing risk through Weighted Average ANN + CNB Prediction Results

In above figure we input the parameter like, IALS, RPUT, PSSDT, SUMSDT, UHSTS, MFEUSI, PR & GI and their respective values. And get estimated Phishing Risk: 29.3170 of ANN, Phishing Risk: 2 of CNB, Phishing Risk: 17.2975 of ANN + CNB.

VI. CONCLUSION

The proposed work is aimed at detection & classification of phishing website. Phishing website & apps lure unsuspecting users on malicious pages & apps which are look-alike to genuine web pages & apps, to foot them to enter the credentials such as email passwords, social media account details, internet banking passwords etc. These types of attacks are disastrous for security of crucial internet applications. This work is aimed at improvising the detection accuracy of the previous methods employed. Usage of Artificial Intelligence for classification of Phishing Websites is demonstrated. A widely renewed technique for classification, as Naive Bayes Classifier is also compared. The author has proposed a unique method by Augmenting Artificial Neural Network prediction with Naive Bayes classifier to achieve higher accuracy. The method of augmentation employed is Weighted Average, which gives the user flexibility to using trust/weight score to both techniques, so accuracy of results can be further enhanced.

VII. FUTURE SCOPE

The author has demonstrated an agile from work for detection of phishing websites & their segregation based on URL features. With the evolving technology, the phishing attackers are updating their tools & techniques to camouflage original websites & apps, & fool around phishing detection software with advanced techniques including redirection, random name modification etc. With the evolving technical scenario, there is a constant need to update the phishing detection systems to cope up with the growing challenges. Primarily sought after improvement in the proposed system may be the introduction of Self Learning Technique(s), to make the Neural Network learn itself, adopt & update according to its previous predictions & actual results. Also Bio Inspired algorithms can be combined with Machine Leaning to form a Cognitive Phishing Detection System.

VIII. REFERENCES

[1] Shraddha Parekh, Dhwanil Parikh, Srushti Kotak, Prof. SmitaSankhe, “A new method for Detection of Phishing Websites: URL Detection” IEEE 2018.

[2] Weiwei Zhuang, Qingshan Jiang, TengkeXiong, “An Intelligent Anti-phishing Strategy Model for Phishing Website Detection” 2012 IEEE.

[3] U.Naresh, U.VidyaSagar, C.V. Madhusudan Reddy, “Intelligent Phishing Website Detection and Prevention System by Using Link Guard Algorithm” JCE, Issue 3 Sep. - Oct. 2013.

[4] Satish.S, Suresh Babu.K, “Phishing Websites Detection Based on Web Source Code and URL in the Web page” International Journal of Computer Science and Engineering Communications. Vol.1 Issue.1, December 2013.

[5] Ram Basnet, Srinivas Mukkamala, and Andrew H. Sung, “Detection of Phishing Attacks: A Machine Learning Approach” Research gate may 2014.

[6] Hiba Zuhaira, Ali Selmat, MazleenaSalleh, “The Effect of Feature Selection on Phish Website Detection” International Journal of Advanced Computer Science and Applications, Vol. 6, No. 10, 2015.

[7] Zheng Dong, Apu Kapadia, Jim Blythe and L. Jean Camp, “Real-time Detection of Phishing Websites Using Public Key Certificates” 2015 IEEE.

[8] Jun Hu, XiangzhuZhang,Yuchun Ji, Hanbing Yan, Li Ding, Jia Li and HuimingMeng, "Detecting Phishing Websites Based on the Study of the Financial Industry Webserver Logs" ©2016 IEEE.

[9] Pratik Patil, Prof. P.R. Devale, "A Literature Survey of Phishing Attack Technique" IJARCCE, Vol. 5, Issue 4, April 2016.

[10] Pratik Patil, Prof. P.R. Devale, "PHISHSTORM: DETECTING PHISHING WITH STREAMING ANALYTICS" International Journal of Scientific & Engineering Research, Volume 7, Issue 5, May-2016.

[11] Amol C. Jadhav, A. M. Pawar, "Enhancement in Phishing Detection Using Features Clustering" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 6, June 2016.

[12] Gaurav Varshney, ManojMisra and Pradeep K. Atrey, "A survey and classification of web phishing detection schemes" 26 . October 2016 in Wiley Online Library.