Asian Journal of Information Technology 13 (3): 156-169, 2014
ISSN: 1682-3915

An Automatic Detection of e-Banking Phishing Web Pages with NEFCLASS Back Propagation

[1]P. Malathi and [2]P. Vivekanandan
[1]Dhanalakshmi Srinivasan College of Engineering and Technology,
Anna University, Chennai, India
[2]A.C.Tech, Anna University, Chennai, India

Abstract: Phishing webpage that mimic the webpage of legitimate, to steal information from users which become the fashionable practice and sophistical growing among the perpetrators of the Web. This phishing scams become a gigantic problem from e-Bankers and e-Commerce users. It is dynamic and very complex problem to classify phishing webpage because of alike absolute character of legitimate webpage. This study presents an approach to overcome the complicatedness for foretelling or classifying e-Banking phishing webpage. The classification of phishing webpage leads to the subjective consideration of various factors, the Neuro-Fuzzy Classification (NEFCLASS) Back Propagation algorithm can be an effectual analysis of classification model. The NEFCLASS Back Propagation algorithm analyzes the webpage in natural way in human intellectual manner. The various multi modal features are considered in this study for effectual classification with three phishing stratums. Thirty features are extracted are grouped in six different property that under three stratums, respectively.

Key words: e-Banking, webpage, phishers, phishing attack, fuzzy logic, neural networks

INTRODUCTION

Internet has become an inseparable and vital part of day to day life and it is always changing, evolving and living entity. With the development of computer networks, owing to the revolutionary changes throughout the network has bought increasingly paid attention. Beyond imagination, internet dependence as an information miner and knowledge bank is increasing exponentially. The advent of the telecommunication in internet, confidential information is also being transferred through internet. Internet's widespread pros tied with the incredible growth in e-commerce and m-commerce, it creates security challenges. However, this growth of internet leads intruders and crackers to play online con game over the victims and so it is a vital need for information security. If the information's security is erroneous and not ensured then tremendous development in internet can become void and null. It is, hence the information over internet is a challenge before the professional community in need to ensure high-level security.

e- Banking: e-Banking refers to electronic banking and it is like e-Business in banking industry. e-Banking is also called as "Online banking", "Internet banking", "Virtual banking", "Personal banking", "Home banking" and other names. e-Banking is a result of the growing expectations of bank's victims and the various names holds different banking activities can be transacted at anytime from anywhere. e-Banking is completely delivered by computer controlled system and so victims do not have to spend time in the bank's premises. The improvement in computer controlled system in telecommunication and computing shows great effect in retail banking. Due to the advent of technological innovations in e-Business, the financial sector is changing under the impact of technological advent as well as global competitive. In global banking sector, e-Banking is being considered as a strategic and competitive tool in the emerging internet technology (Al-Khatib, 2006).

Phishing: An online con game called phishing, played with confidential information of innocent internet users. Phishers are tech-gumption con artists and they use crime-ware, spam and fake webpage to trick people into unwrapping confidential data such as bank account details. Mostly, phishers phishes by sending a wave of spam email, sometimes the message may up to millions in count. Phisher use spam mail as a lure, the mail appears

Corresponding Author: P. Malathi, Dhanalakshmi Srinivasan College of Engineering and Technology, Anna University, Chennai, India

that come from trusted and well-known company and so the contains an emotional content to a false crisis with company's logo and name. The content of mail looks like business-like language, couched in urgent and mail urges the user to enter personal/confidential data and so, the mail redirect to a spoofed webpage (Wang *et al*., 2012). Like the mail content, the webpage instances have been masked so the webpage seem real and urge the victims to provide confidential information. The confidential information may be login username, passwords, account number, other bank confidential details and so on (Dhanalakshmi *et al*., 2011).

The phishers well-known that least a fraction of victims is bamboozled into surrendering their data, since the webpage seemed like legitimate and so only successful response rate of 5% (Fu *et al*., 2006). The juxtaposition with bogus and hostile webpage as crime-ware began by phishers that leverages the malware and other vulnerabilities in victim machine. By following the phony webpage, the phisher can steal the all details about victim and would no longer necessitate getting the victim details. The malware that capture all victim details when the victim entered into the legitimate webpage of e-Bank. Recent years, phony webpage crime-ware genus has become more aggressive and targeted stealth in victim's system remains hidden. Throughout past year, phishers are targeting e-Banking webpage like Citibank, PayPal, US Bank, eBay, etc. to steal account and credit assets of the victim (Aburrous *et al*., 2010a, b).

Many works has done to prevent victims from phishing attacks. By user verification by pattern matrix technique involves verifying the webpage weather the webpage is legitimate or phony and it is done by dynamic text hashing technique. The webpage would respond when user login into it (Gaurav *et al*., 2012). Digitally signed emails genuinely verify the senders' information. This digital signature email system that prevents from the phishing emails by the standards named OpenPGP and S/MIME. Spam filters mechanism checks the emails and classifies the incoming mails as either normal mails or spam mails. Internet service providers and other very large organizations normally use gateway spam filtering, it adjudges arriving mails at gateway. Web browser extensions rate the webpage by their features and block the phony webpage. Internet Explorer uses EarthLink Toolbar to block the phony webpage. Mozilla Firefox has inbuilt feature to checking vipaged webpage with known phishing pages and this feature in browser updated regularly to provide up to date protection. Online brand monitoring system such as NameProtect, Netcraft and Cyveillence and offers monitoring spam emails, domain name, webpage content and other features content of webpage (Damodaram and Valarmathi, 2012). These are some of the systems that detect the phishing webpage with normal features.

Contributions of the paper: In this study, classification of phishing webpage by incorporating key tectonic features in phishing webpage and employ learning algorithm to the extracted dataset for the classification process. The learning algorithm trains the dataset to learn labels of instances, i.e., legitimate or phishing webpage. This work insight into the effectiveness of using NEuro-Fuzzy CLASSification Back Propagation data mining learning algorithm for the purpose of accurate phishing webpage classification. To address a gamut of operational problems, soft computing techniques like NEFCLASS Back Propagation data mining algorithm is used. Classification is a type of supervised learning; supervised learning assumes that there is previous knowledge about the class membership of the observations, i.e., totally different features from legitimate webpage. The purpose of using learning algorithm in e-Banking webpage detection is to directly extort structure from a dataset. Although, unsupervised learning approaches provides for a dependable results, supervised learning approach provides for a much improved accuracy to gain knowledge from a dataset.

Literature review: A number of methods are being addressed by various researchers in the past; in this study some of the methods related to phishing webpage detection are provided in this study.

Dhanalakshmi *et al*. (2011) have proposed human proficient phishing webpage detection by comparing with trained features with claimed identity of a webpage. An MD5 is a password hashing algorithm helps to increasing the strength of web password authentication for secure transaction. Salt values are added with hashed password, so it is difficult to track original password. For further access, the valid user gets the session key via mobile device and it provides high user security.

Fu *et al*. (2006) have proposed an effective Earth Mover's Distance (EMD) method for phishing webpage detection by using visual similarity of webpage. The image signature represents by converting the corresponding webpage page into low resolution images and color features are used. Earth Mover's Algorithm Method calculates the distances of the trained images with testing images by using a fixed EMD threshold vector for classification of webpage. This online classification solution leads to high precision level and good performance of time. The webpage have tested over large scale with 10,281 suspicious webpage pages.

Damodaram and Valarmathi (2012) have proposed an approach for predicting and detecting phishing webpage which reduced complexity of detection. They concentrated URL and domain identity and security and encryption characteristic based rules are generated to detecting the phishing webpage. Based on the characteristic phishing webpage rate was calculated for classification process. Here, the optimal solution for the detection of phishing webpage using MBAT algorithm that uses meta-heuristic functions, the performance is compared with ACO and PSO algorithms.

Aburrous *et al*. (2010a, b) have proposed an investigation of Phishing Detection Model using combined Fuzzy Logic Data Mining algorithms which categorize the features of webpage and classification phishing webpage form normal webpage. The fuzzy logic was applied to framed 6 criteria of phishing webpage for defining and classification of webpage. The criteria gave more importance to URL and domain entity layers for the phishing webpage classification. The e-Banking phishing webpage classification rate is calculated for the classification.

Dong *et al*. (2010) have proposed method analysis the behavior of online users for detection of phishing webpage by analyzed the submitted of users' data to such webpage and vipaged webpage by users. The manipulations of those users' data are possible by the hackers. The accuracy rates of those data based detection are more predictive detection rate and it was flexible against changing trickery methods.

Komiyama *et al*. (2010) have proposed phishing detection based on content of webpage and this content-based phishing webpage detection has reported that detect the webpage written pages with English. The other content based detection techniques are done in small scale and measurement of detection error rate but won't analyze the cause for it. Their proposed CBD Method effectiveness is tested with English rather than non-English pages like Chinese, Japanese, etc. The CBD refers the depth evaluation of English content-based webpage classification over 843 actual webpage and limitation was discussed.

Martin *et al*. (2011) have proposed a competitive solution for detection and classification of phishing webpage from authorized webpage. This dynamic problem deals with many dominant feature criteria of phishing webpage and prediction problem was discussed about phishing webpage using neural network. The multi-layer neural network system helped to increase the performance as well as reduces the error rates. The neural network method, the framework results the good prediction and classification of phishing webpage. Zhang *et al*. (2011) have proposed a framework for content-based phishing webpage detection using Bayesian approach. This Bayesian framework takes the visual and textual contents of webpage to measure the distances, i.e., similarity between the phished webpage and protected webpage. For good classification, the proposed framework fuses text classifier and an image classifier. The Bayesian Model uses the image and text features for this classification model to calculate matching threshold and it was used to classify whether the webpage is phishing webpage or not. The naive bayes rule is used in text classification to estimate the probability rate of the phishing webpage and the trusted one. The earth mover's distance was used in image classification to estimate the visual similarity distance between the webpage. The Bayesian Model framework was intended for estimating the threshold. In the fusion algorithm, the decision making classification results from text and visual content are estimated by bayes theory. Their proposed fusion approach is tested over the long scale dataset which gave the results depends on only the visual and text characteristic of webpage (Fig. 1).

Singh *et al*. (2011) have proposed dynamic watermarking technique for anti-phishing approach. This approach completely depends on user verification, by querying the information from the server like watermark image and its position and secret key when user registration process time. This approach helpful for the user itself verify the webpage whether the webpage phishing or legitimate by dynamic nature of login credentials. And for every login phase, the user verifies the login webpage by analyzing the watermarks and if the webpage responds properly the webpage is trusted one otherwise not.

Kim and Cha (2011) have proposed Webpage Risk Assessment System (WRAS) for phishing webpage detection. The security risks index are checked and computed by WRAS. WRAS ensures the webpage trustworthiness by providing warning to the user's side. So, the inexperienced users against satire phishing webpage attacks and the trusted webpage occurs attempting exploit-based webpage phishing. Zhuang *et al*. (2012) have proposed a principled cluster ensemble framework based on agreement wise partitioning. The framework combines individual clustering solutions which are based on agreement wise partitions for spam categorization as well as phishing webpage classification by clustering. The clustering is based on knowledge of domain with webpage level constraints can be naturally integrated into the framework. This approach is tested from the large scale dataset from Kingsoft Internet Security Laboratory contains spam collection and phishing webpage.

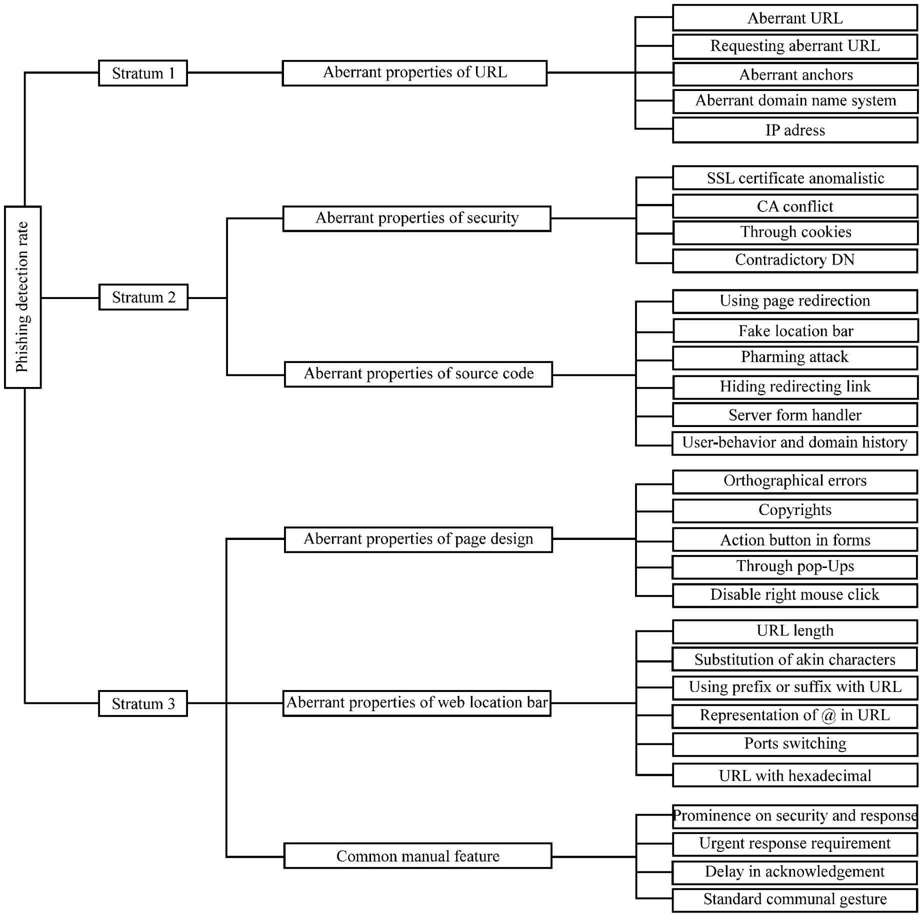
Phishing detection rate
Stratum 1
Stratum 2
Stratum 3
Aberrant properties of URL
Aberrant URL
Requesting aberrant URL
Aberrant anchors
Aberrant domain name system
IP adress
Aberrant properties of security
SSL certificate anomalistic
CA conflict
Through cookies
Contradictory DN
Aberrant properties of source code
Using page redirection
Fake location bar
Pharming attack
Hiding redirecting link
Server form handler
User-behavior and domain history
Aberrant properties of page design
Orthographical errors
Copyrights
Action button in forms
Through pop-Ups
Disable right mouse click
Aberrant properties of web location bar
URL length
Substitution of akin characters
Using prefix or suffix with URL
Representation of @ in URL
Ports switching
URL with hexadecimal
Common manual feature
Prominence on security and response
Urgent response requirement
Delay in acknowledgement
Standard communal gesture

Fig. 1: Stratums of e-Banking webpage

Problem defnitions: An e-Banking phishing webpage camouflaged as a legitimate e-Banking webpage by screening the distinctiveness of the legitimate webpage in its fraud criteria stratums. Observing the bank's titles or logos exhibited in their browsers, online users could mistakenly deem the phisher's lure. In this study, the phishers lure is classified into three phish stratums with different priority weight according to harmful characteristic of phish stratum. In these three stratums, totally thirty features are extracted for e-Banking phishing webpage rate.

Feature extraction and analysis: The phishing feature extraction is the curious part of this research. The webpage features are analyzed and taken from the phished webpage and further divided into three stratum according to the phished density makes harm. The anatomy of the URL is marked in Fig. 2 and URL is the core lure for the phishers. The analyses in three stratum features of phishing webpage are as follows.

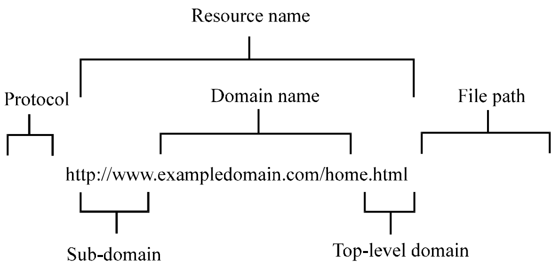

Fig. 2: Uniform resource locator anatomy

Stratum 1: The webpage's first stratum contains the properties of Uniform Resource Locator (URL) and Domain Name System (DNS) records. This stratum of page reflects the nature of webpage with high percentage.

Aberrant properties of URL (P1)

F1 (Aberrant URL): The hostname is qualified name of domain and hostname identity of the URL or IP address should match its identity but the phishers URL/IP doesn't. The hostname is the combination of local host name and domain name. This identity feature can be extracted and checked from WhoIS database (WhoIS).

F2 (Requesting aberrant URL): The object references used in the legitimate webpage are mostly mapped between the same domains but the phishing webpage's object references are externally called from other domain sources. For example, the webpage www.example.com sources are loaded from www.anotherdomain.com/ object_reference and it supports the phishing webpage seems like legitimate one.

F3 (Aerrant anchors): The legitimate webpage/e-mail will have anchor tags which offer navigation to corresponding webpage. The anchor tag has the structure of <a href=""></a>. The phishers uses the anchor tags by leaving the tag empty or referenced with another domain. For example, the phished webpage/email anchor tag could be lured as<ahref="www.anotherdomain.com/index. html" >www.example.com</a>.

F4 (Aberrant domain name system): The Legitimate Webpage Domain Name System records of each domain are being maintained in WhoIS database which operated by Regional Internet Registries (RIR). The WhoIS hold the registered users domain information that includes domain name, sub-domains list, etc. The WhoIS doesn't hold phished webpage domain information which won't register or holds aberrant domain information.

F5 (IP address): IP address can be used instead of web address URL. The IP address has the structure of 0-255.0-255.0-255.0-255, so the phishers can use the IP address as lure to steal user private information. The IP address can be in hexadecimal format or octal format or decimal format.

Stratum 2: The second stratum of the webpage that contains the security control properties of page and the properties of page source code is evaluated.

Aberrant properties of security (P2)

F6 (SSL certificate anomalistic): The Secure Socket Layer (SSL) certificate which offers security layer for a webpage. The webpage that holds the SSL certificate can transact securely using cryptographic protocols in SSL. The web location bar uses the https protocol for navigating transaction pages in secure manner. The https protocol that changes location bar color green with a padlock sign and it denotes the transaction of webpage is secured. The phishers can't able to use SSL in their webpage due to the unsecured transaction between the unknown domains.

F7 (CA conflict): Certificate Authority (CA) is a signed digital certificate and CA hold individuality of the genuine organization domain in the certificate. When SSL handshake between user and server happens, the user's resource requests the CA of the domain to verify the individuality of the certificate. The phished webpage's CA conflicts, if the signed digital certificate matched with other legitimate CA.

F8 (Through cookies): The user holds the cookies that dumped by internet server and this generate a bond between user machine browser and internet server. These cookies are used for tracking the webpage information as well as authentication depend on history and preferences of webpage. But the phishing webpage doesn't point legitimate domain and not consistent to its own domain. So, the identity of the webpage information gets conflicts.

F9 (Contradictory DN): The Distinguished Name (DN) is identifying unique information in Secure Socket Layer (SSL) certificate. The DN contains organization name and location and its unique proficient domain name that user for servers DNS lookups. The Certificate Signing Request (CSR) verifies the Distinguished Name (DN) on the web server, the DN contain unique content that differentiate from other domains. But the phished webpage's DN gets conflicts.

Aberrant properties of source code (P3): These feature properties are extracted from the source code of the webpage. The source code of the webpage is as follows:

F10 (Using page redirection): The phishing URL can have legitimate domain name but it using redirection, it redirects to the phished page and it won't consider the URL before redirection. For example, the webpage can be redirected by the URL as "www.example.com/redir?http:// www.anotherwebpage.com/" and double redirection also possibly used by phishers.

F11 (Fake location bar): The phishers tricks the user by put on view on browser duplicate location bar that contains the phisher assigned legitimate webpage's address. Phishers uses JavaScript to show a duplicate location bar that close and replaces the real browsers location bar or a secondary tiny browser opens on the exact part of location bar in the browser. This fake location bar phishing trick is used web world are rare.

F12 (Pharming attack): The webpage in WWW can be accessed down to the verity they has an distinctive identifier named internet protocol address which help to locate and access the webpage. The phishing webpage's more sophisticated variant is pharming attack and the word pharming is the combination of phishing and farming. The pharming attack poisons the DNS server by handling the IP address of legitimate domain on server that changed it in phishers accordance. The user enters the legitimate web address but IP address of the domain reflects the fraudulent webpage.

F13(Hiding redirecting link): The phishers tricks the user by hiding the redirecting URL by using onMouseOver event handler. This onMouseOver event handler shows the aberrant URL in the browsers status bar. For example the phishers can use onMouseOver event handler as <a onmouseover="window.status=' https://www.example. com'; return true"onmouseout =" window.status='https://www.example.com''"href=" http://www.anotherwebpage.com">https://www.example. com</a>. The status bar shows example.com instead of anotherwebpage.com.

F14 (Server form handler): The phishers gets the users confidential information on the submission of web forms. The phishers tricks the fraudulent webpage with aberrant server form handler by leaving blank or null strings in webpage action. For example the null web form action strings can be used <form action="about: blank">, <form action="#">, <a href="#skip">, <a href="#content">, <form action="javascript:void(0)">, <form action = "javascript:true">, etc.

F15 (User-behavior and domain history): User behavior on the webpage that deals the user traffic that extracted from Alexa server and the domain history can be taken the webpage creation date from WhoIS database. The phishing webpage lifetimes are more possibly within 3 days periods. So, the user behavior on webpage and domain history helps to manipulate the activities of webpage.

Stratum 3: The third stratum of the webpage which contains the properties of page content and design, web location bar and some manual features phishing mails.

Aberrant properties of page design (P4):

F16 (Orthographical errors): The legitimate webpage is designed in orthographical manner. But the phishing webpage are not orthographically designed and so the phished webpage content may have orthographical errors. The phished can be identified by checking the content of webpage in orthographical form.

F17 (Copyrights): The phishers concentrates to trick the user by webpage page design resembles the legitimate one and so the phishers copies entire properties of the legitimate webpage and its content. But the domain of the webpage is varying from the properties of the webpage. All properties of the webpage are loaded from the legitimate webpage and it is considered as doubtable webpage.

F18 (Action button in forms): The phishers gets the confidential information of the users in action of web forms. The domain name used in the action of webpage is varying from the webpage domain name. Normally, the confidential information submitting actions are processed in 'Post' Method because of security reasons but the phishers may use 'Get' Method to get the user information.

F19 (Through pop-ups): The popup window is generally used as for the purpose of advertisement, subscription conversions, organization or domain information, etc. but it is very rare that popup window that collects the confidential credential information from users.

F20 (Disable right mouse click): The right mouse click can be disabled or shows a warning message by the phishers to restrict the user to view its source. Most probably, JavaScript is used by phishers to disable the right mouse click option.

Aberrant properties of web location bar (P5)

F21 (URL length): The URL length can be considered with three factors are number of dots of host part or character length of host part or total character length of the URL. Possibly, the phishing URL can have more than four dots at host part or more than thirty characters at host part or the URL length exceeds seventy five characters.

F22 (Substitution of akin characters): The phisher replaces the akin characters of domain name and which tricks the user to believe the webpage as a legitimate webpage. For example, the phisher's replaced webpage www.exanple.com from the legitimate webpage www.example.com.

F23 (Using prefix or suffix with URL): By adding prefix names or suffix names with the legitimate domain name to the URL, so the user feels that they are dealing with legitimate webpage URL is good. The webpage uses dash or dot to represent the prefix or suffix in webpage URLs. For example, the URL "www.example.com" can be lured as "www.prefixes-example.com or www.example-suffixes. com".

F24 (Representation of @ in URL): The representation of @ symbol in the URL that do not take into account the link before @ symbol used and redirects the link after @ symbol. For example, the representation of @ symbol in URL is “www.examplepage.com%00@anotherpage.com” and here %00 is used to hide the text in URL after the code is used. The phishers take this symbol’s advantage to hiding the details of the link from the users, here is a phishing @ symbol example is “http://www.example-e-banking-page.com.aw-ebankISAPI.dll%00@ 210.93.131.250/my/index.htm”.

F25 (Ports switching): The port number of webpage URL uses port 80 as default port, if not specified in URL. So, the phishers uses port numbers in URL to redirect the users. For example, the standard port switching URL is “www.example.com:6700”. The phishers use this trick to innocent users make them to enter sensitive records, here is a phishing example for switching ports is “www.example-e-banking-page.com:ac-KTtF4BD6y4TZ1cv6 GT5D@64.29.173.91:8034”.

F26 (URL with hexadecimal): Hexadecimal codes shows the characters that in %CODE format based on the ASCII table and the user can’t predict the webpage name which it directing. For example, the URL “www.example.com” can write as “www.%65%78%61%6D%70%6C%65.com”. The phishers uses these codes to direct the webpage to suspected pages without the knowledge of user, the hexadecimal character coded phishing URL may look like “http://www.example-e-Banking-page.com%00@ %32 %32%30%2E%36%38%2E%32%31%34%2E%32%31%33”.

Common manual feature (P6)

F27 (Prominence on security and response): The phishers giving emphasizing the security concern to the users to believe them. For example, the phishers lure as follows: “Keep in mind: ****** will not ask you for sensitive private records (such as your password, credit card and bank account numbers, etc.) in an electronic mail.”

F28 (Urgent response requirement): The phishing webpage have very short lifetime. The phishers have to gather the users sensitive data before that suspected webpage lifetime ends and so, they would convince the customers by acknowledging the users to respond quickly. For example, the phishers convincing fraud emails as follows: “If you don’t react within 24 h after getting this mail records your account will be neutralized and detached from the server and your account suspension will be made due to several inconsistencies in your registration records as explained in Section 11 of the ****** User Agreement.”

F29 (Delay in acknowledgement): The phishers visually tricks the users by buying time to access their account. For example, the acknowledgements of phishers email statements as follows: “This process will take 3 days, time when you will not be able to access your ****** account. After this period you will get commands to enter and securize your ****** account.”

F30 (Standard communal gesture): Phishers uses standard greeting in the bogus emails like Dear Customer or the Valued Customer or Dear Member or Dear <e-mail address&gt as a salutation greetings.

MATERIALS AND METHODS

This study is organized as brief research about proposed NEFCLASS Back Propagation Algorithmic Model as shown in Fig. 2 consisting of training stage and intelligibility of phishing webpage classification. The NEFCLASS Back Propagation algorithm is a Neuro-Fuzzy System, i.e., both the combination of Fuzzy system and Neural networks. The Fuzzy System and neural network is the two independent models and this combination brings both models advantages for problem solving. The proposed NEFCLASS Back Propagation Algorithmic Model holds this advantage for better phishing webpage classification.

Fuzzy systems: The fuzzy system is the approach for computational problem that depends on human behavior and connoisseur as like human’s usual communication and offers the terms of computational representation. Earlier the solution fuzzy systems helps only in control problems but it used in many problem solving system areas. In this research, the Fuzzy System is used for the classification of phished webpage. The idea of the Fuzzy Classification System is deriving the human connoisseur linguistic rules actions and this process is called cognitive analysis. For example:

- R1: if A is small and B is medium then the class is C1
- R2: if A is medium and B is small then the class is C2

Here, small and medium are the two connoisseurs discussing about ‘C’ types. The results of fuzzy classification rules have identify the absolute class by assigned crisp value. The fuzzy classification system is simple that deals many ways to find solution for a problem when compared with other machine learning algorithms. Fuzzy classification groups the difficult multivariate data into certain number of classes according to their category which reduces the difficult multivariate data’s dimensionality. This classification provides intelligent decision and constructing and maintaining the results loyalty. The cons of Fuzzy Systems are as follows: It Fuzzy System doesn't a learning system, there is no formal methods in Fuzzy System for tuning and adaption is complicated in redefined environment.

Neural networks: Neural network system uses the expected or recognized principles of human intelligence and it has number of independent neurons i.e. can be viewed as nodes or simple processors. The neurons or nodes in the neural networks keep interaction with one another through connected link weights. The values of input and the state of the neural networks produce the layered output and the output is assumed as input for the new state. The neural network consists of an input layer, one or more than one hidden layers and an output layer. The input and output of the neural networks are deals directly with applications or users and the hidden layers are not directly communicated with the application or users. Each node is considered as processing unit of the neural network which is connected as directed graph with weighted edges. Each layer in the neural networks is processed with heuristic function in order to obtain output. The cons of neural network are as follows: In neural network the extraction of rules are not possible, no prior knowledge is used and adaption is may be complicated in redefined environment when relearning is required.

In this study, researchers use NEFCLASS back propagation neural network for detecting e-Banking phishing websites. Using the input values rules will be formed and the network will be trained to give output. The most luring characteristic in neural networks is the possibility of learning. During the learning phase, flows through the network may change its structure based on external or internal information. The network learns with the identified results are existing to it. Initially network allocates with arbitrary weights. To bring the concluding output nearer to existing result, weighting factors are adjusted by back propagation algorithm. Each and every unit in the network will receives signals from its input links and calculates a new activation level that it sends together with each of its output links. The calculation of the activation level is based on the values of each input signal received from an adjacent node and the weights on each input link. This calculation is split into input function that computes the weighted sum of the unit's input values and the activation function that transforms the weighted sum into the concluding value that serves as the unit's activation value. Networks is then updated and try to make it reliable. This is done by making small modifications in the weights to reduce the dissimilarity between the observed and predicted values. Here, the weight update rule is mainly straightforward. If the predicted output for the particular output unit is M and the target output should be N then the error is acquired by taking dissimilarity between predicted output and target output. The back-propagation algorithm is a sensible approach to isolating the contribution of each weight. These weights connecting to each and every input to an output, a weight contributes to more than one output. The weight update rule at the output layer is based on the obtained error value. Error back propagation is done for updating the links among the input and the hidden units for obtaining the output of these units. Here, this hidden unit is in charge for some portion of error in each of the output unit to which it joins. This process is repeated until it reaches the estimated output.

Structure of proposed NEFCLASS Back Propagation algorithm: NEFCLASS Back Propagation is Neuro-Fuzzy System which is the combination of Fuzzy System and neural networks in a homogenous architecture. So, the NEFCLASS Back Propagation Model holds both pros of the linguistic human connoisseur rules of Fuzzy System and learning ability of the neural networks with insertable prior knowledge in network. The NEFCLASS Back Propagation algorithm finds learning method for phishing webpage classification with Fuzzy System parameters. The proposed phishing classification model considers structure learning, i.e., feature parameter learning and fuzzy rule set creation. The neural network learning inspires the algorithm of feature parameter learning (fuzzy rules or fuzzy decision tree). The NEFCLASS Back Propagation Phishing Learning Algorithm represents architecture of neural network with applied Fuzzy System as shown in Fig. 2. The fuzzy learning system used in the neural networks for better performance or enhance the learning capabilities faster. The NEFCLASS Back Propagation phishing learning algorithm is represented by, feed-forward neural network architecture is not prerequisite to train the system but to envisage the classification flow of the phishing feature in network (Fig. 3).

Where, l_w denotes Input Webpage (can be URL or Email), w_{s3} denotes the weight, Sn denotes Stratum of phishing, Pn denotes Stratum's property, R_w Rate of the webpage and O_w (O_{PW} - Phished or O_{LW}-Legitimate) Output of phishing webpage.

The NEFCLASS Back Propagation Model derives a fuzzy phishing rule set from all features of three stratums and formalizes the neural learning algorithm. It always

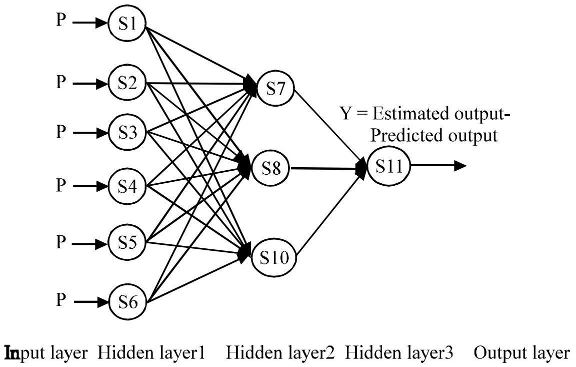

Fig. 3: Model for phishing webpage detection

interprets the fuzzy rule set in the form of If then rules and executes function approximation with indiscriminate logical fuzzy rules. The proposed NEFCLASS Back Propagation architecture provides three-layer fuzzy perception represents fuzzy-neural network and the weights of each layer is modeled as per the effectiveness of the features. The stratum one holds highest weight because of its harm URL and domain features, the stratum two which is weighted secondly with security certificates and page source scripts and the stratum three holds the lowest weight with webpage design and location bar are modeled as fuzzy sets. Figure 2 shows the proposed NEFCLASS Back Propagation learning classification algorithm with six fuzzy rules set in three hidden layers. The three hidden layer represents three phishing stratums holds three different weights according to the harmfulness. The neurons in the NEFCLASS Back Propagation Phishing Network Model act as activation function that calculates with weights to the corresponding fuzzy rule set based threshold unit. This is how the way that fuzzy rule set can be interpreted as special neural network for classifying the phished webpage as NEFCLASS Back Propagation Model and this network can be modify the feature parameters or the flow of network i.e., structure corresponding to the feature set taken for classification by inserting or deleting the nodes/neurons and weights in the network.

System design: The first layer contains only URL and domain identity criteria with a weight equal to 0.3 for its importance; the second layer contains Security and Encryption criteria and Source Code and Java script criteria with a weight equal to 0.2 each; the third layer contains Page Style and Contents criteria, Web Address Bar criteria and Social Human Factor criteria with a weight equal to 0.1 each. The six criteria have been classified and prioritized through mining the e-Banking phishing website archive database using the classification and association algorithms mentioned earlier:

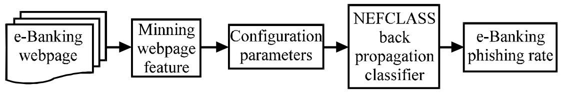

Fig. 4: Architecture for proposed NEFCLASS Back Propagation Model

```
Input: e-Banking webpage (can be URL or email);
Output: e-Banking Phishing Percentage (Label: Phished or Legitimate);
   Step 1: Mining the input URL from PhishTank [Online];
   Step 2: Extract features from different stratum level;
   Step 3: Derive the fuzzy rule system for each stratum;
   Step 4: Assign weight for each stratum in classification network;
   Step 5: For each $F_i$ in $S_{Pj}$, apply Fuzzy rule
                 IF $F_i$ in $S_{Pj}$ is phished
                         THEN assign threshold value;
                 ELSE assign neutral value zero;
         Return value;
   Step 6: Calculate the rate of webpage with neural function;
   Step 7: Display the labeled percentage of webpage;
```

In this phishing classification model, the NEFCLASS Back Propagation classification percentage is calculated based on the six properties in different stratum density. Aberrant properties of URL, Security, Source Code, Page Design, Web Location Bar and Common Manual Feature are the six different properties taken for phishing classification as shown in the Fig. 1. Based on the six properties totally thirty feature phishing parameters are analyzed in this research. The source code and web location bar property has six features, respectively, the properties of URL and page design has five features, respectively and four features for properties of security and manual features, totally thirty features parameters taken into the account for NEFCLASS Back Propagation Learning Model. The NEFCLASS Back Propagation Model has three stratums, the first stratum's weight assigned as 0.3 according to the URL properties that reflects harmful phishing effects. The stratum two holds the weight of 0.1 for each property and at last 0.1 for each property in stratum three. The designed NEFCLASS Back Propagation phishing classification pseudo code is described in box.

Architecture: The architecture of proposed NEFCLASS Back Propagation e-Banking webpage detection describes the flow of detection described in Fig. 4. The e-Banking phishing features are mined from the dataset and configure the fuzzy rules based on the phishing parameters analysis. The NEFCLASS Back Propagation network neurons assigns the threshold values of each rules sets and activation function for each neuron is calculated for classification. The percentage of phishing webpage is calculated from the NEFCLASS Back Propagation rate.

Mining phishing features: This e-Banking phishing detection features are thirty in counts that taken for classification. The average e-Banking phishing webpage's life time should be two and half days and the webpage data must have some errors while checking. So, it is very difficult to extract all archives of phishing webpage features from short lived webpage and therefore the age of e-Banking phishing webpage is the major problem faced for datasets record. Furthermore, the webpage used for learning is dynamically changed by often updating the dataset. The parameters are configured based on the features that explained in study.

Rule generation in NEFCLASS Back Propagation Model: The rules are generated for each e-Banking phishing webpage features that are grouped in three stratum levels. The rule generation of each stratum is discussed.

Stratum one rules: The stratum one has five features set is categorized in a URL property such as aberrant URL, requesting aberrant URL, aberrant anchors, aberrant domain name system and IP address. These five features has five individual fuzzy rules that checks the each features is phished or not brings binary output either true or false. And collects all the stratum one binary results, NEFCLASS Back Propagation creates an aggregated rule that assigns the threshold values. The threshold values of the stratum one is the output of the aberrant properties of URL. The stratum one rule set produce one threshold value output based on ten feature rule entries.

Stratum two rules: The stratum two has two rule base sets are aberrant properties of security and aberrant properties of source code. The security feature rule set such as SSL certificate anomalistic, certification authority conflict, through cookies and contradictory distinguished name and the feature set of source code such as using page redirection, fake address bar, pharming attack, hiding redirecting link, server form handler and user-behavior and domain history. These two fuzzy rule set contains four and six features, respectively. Each and every feature is treated in fuzzy rule to generate binary IF-ELSE output. The final one threshold value output of stratum two rule set property derives based on twenty (eight+twelve) feature rule entries.

Stratum three rules: The stratum three has three rule sets are aberrant properties of page design, web location bar and some manual features. The page design feature set has five feature properties such as orthographical errors, copyrights, action button in forms, through pop-ups and disable right mouse click which deals the phishers rules on webpage design. The properties of web location bar has six rules based on URL length, substitution of akin characters using prefix or suffix with URL, representation of @ in URL, ports switching and URL with hexadecimal. And finally some stratum three has some phishers common manual feature rule set that rules are four in count such as prominence on security and response, urgent response requirement, delay in acknowledgement and standard communal gesture. The stratum three rules contains three rule set properties that produce three threshold value output based on ten, twelve and eight (thirty) features rule entries, respectively.

The phishing features are modeled in fuzzy rules as mentioned in feature extraction and analysis which derives the property of each features. The feature fuzzy rule is derived as follow:

$$F_i = \begin{cases} \text{IF } F_i \text{ is true,THEN } F_i \text{ is phishing} \\ \text{ELSE } F_i \text{ is legitimate} \end{cases}$$
$$\forall \text{for each feature}$$

Training and classification: In the e-Banking phishing webpage rule set training in NEFLCASS Back Propagation assigns threshold values for each stratum property. The threshold values initialized as α, β and γ for the stratum property has four and five features, respectively and τ, μ, ε, and ρ for the stratum property which has six features. The stratum three has manual features which the threshold value assigned as η. The threshold values are assigned between 0 and 1. The stratum has four or five feature rule, the rule set modeled as:

$$S_{i_{pj}} = \begin{cases} \text{IF three rules are true,THENassign } '\alpha' \\ \text{ELSE IFtworules are true,THENassign } '\beta' \\ \text{ELSE IFonerules are true,THENassign } '\gamma' \\ \text{ELSE assign}'\text{zero}' \end{cases}$$
$$\forall \text{ for stratum has 4 or 5 rules}$$

The stratum has six feature property rules, the rule set of stratum property modeled as:

$$S_{i_{pj}} = \begin{cases} \text{IFfour rules are true,THENassign } '\tau' \\ \text{ELSE IFthreerules are true,THENassign } '\mu' \\ \text{ELSE IFtworules are true,THENassign } '\varepsilon' \\ \text{ELSE IFonerules are true,THENassign } '\rho' \\ \text{ELSEassign } '\text{zero}' \end{cases}$$
$$\forall \text{ for stratum has 6 rules and } S3_{P6} = \eta$$

The activation of each neuron in NEFCLASS Back Propagation network is based on the stratum rule sets. The weight w_i assigned for each stratums are 0.3, 0.2 and 0.1, respectively. The activation function of the proposed NEFCLASS Back Propagation network has three different functions A_1, A_2 and A_3, respectively because of different phishing stratum. For stratum one, the activation function A_1 is:

$$A_1 = w_1 \times S1_{P1} \tag{1}$$

For stratum two, the overall summation of activation function A_2 is:

$$A_2 = \sum\nolimits_{j=2}^{3} w_2 \times S2_{Pj} \tag{2}$$

For stratum three, the overall summation of activation function A_3 is:

$$A_3 = \left[\sum\nolimits_{j=4}^{5} w_3 \times S3_{Pj} \right] + (w_3 \times S3_{P6}) \tag{3}$$

The NEFCLASS Back Propagation Model for e-Banking webpage phishing classification is performed by the features that extracted from the webpage. The threshold values are applied to the feature set of each stratum by using the feature property that modeled in study. The classification network provides the activation function on each neuron and performs the function on each stratum to make them as one output. The output layer calculates the overall rate (R_W) of the webpage. The rate R_W is:

$$R_W = \sum\nolimits_{i=1}^{3} A_i \tag{4}$$

The NEFCLASS Back Propagation classifies the webpage by calculating the percentage of the webpage using R_W. The e-Banking webpage percentage is:

$$\text{e-Banking webpage percentage} = R_W \times 100(\%) \tag{5}$$

Thus, the proposed NEuro-Fuzzy CLASSification Back Propagation Model classifies the webpage by calculated percentage to find whether the webpage is phishing or legitimate.

Performance evaluation: The performance of proposed NEFCLASS Back Propagation phishing classification model is evaluated by thoroughly analyzing the phishing features and parameter.

Dataset: "PhishTank" is an open source dataset available from the webpage phishtank.com (PhishTank) for researchers and developers to amalgamate anti-phishing information into their applications. PhishTank is a shared domicile for records and aberrant properties about phishing lures on the World Wide Web (WWW). The PhishTank is considered as the key phishing report systematization from 2007 collections to till this date which contains 800+e-Banking phishing webpage. The PhishTank collects the phish database of lures like URLs, IP address, domain entities, server, registrar, reported details like time of the report, screenshots of suspected webpage (if available), etc. are openly on hand. This study's aim to congregate the lures of attackers to indentify the suspected webpage and so, the thirty features used in this implementation are extracted from the PhishTank webpage collections. The dataset contains the above mentioned thirty features are extracted after a complete investigation about the phishing attackers lure techniques that changed over time. The users are not aware to check the security constraints of webpage, so the security of the sensitive information requisition makes much tricky to identify the either legitimate webpage or suspected webpage. This investigation results the thirty features are grouped under three phishing stratums with different weight according to the effectiveness of the phishing lure to calculate the rate the webpage. Also, visual phishing tricks are used as lures by phishers that may fool even most refined users and so it is considered in third stratum after investigation. The particular webpage's domain features are extracted from "WhoIS" () and "Alexa" (). The WhoIS and Alexa database hold whole domain information about the webpage.

RESULTS AND DISCUSSION

In this study, the experimental studies are conducted using data collection from PhishTank to evaluate the classification model that proposed in this study. In this experiment setup, on the basis of extracted core term frequency features of phishing webpage are taken into an account to for effective classification from the legitimate webpage. In this study, the phishing webpage classification performance of the NEFCLASS Back Propagation Algorithmic Model is evaluated using evaluation metrics. All the experimental studies are analyzed under the infrastructure of Windows XP operating system with Intel Pentium(R) Dual-Core processor 2.30 GHz CPU and 2 GB of RAM.

Evaluation of proposed NEFCLASS Back Propagation for Phishing Website Classification Model: Using phishing webpage URL collection and its webpage collected from

Table 1: Evaluation of phishing website classification sample classification percentage model

URL	Security	Source code	Page design	Web location bar	Common manual feature	Percentage
5	1	4	4	3	1	100
4	1	3	0	0	0	70
3	1	3	2	0	0	77
3	0	0	4	4	0	55
1	1	0	3	3	0	53
0	1	4	4	4	0	60
0	1	2	1	2	0	54

PhishTank are evaluated. The term thirty features are extracted from the phishing webpage is considered classification process. Here, the term frequency features are taken under three stratums according to the phishing mischief. These features are trained under NEFCLASS Back Propagation algorithm that fuzziness plays a major role and neural structure helps to provide the accurate results. Three different weights W_i are allocated to each stratum $s_{i,j}$ according to phishing features strategy. This experimental setup assigns 50% to classify the phishing webpage, i.e., above 50% is considered as phishing webpage and below 50% is legitimate one. Because all the term frequency features not present in all phishing webpage.

In this experiment, the weights for different stratums are w1 = 0.3, w2 = 0.2 and w3 = 0.2 are assigned in the NEFCLASS Back Propagation structure to classify the webpage. The fuzzy rules for each six stratum properties have different threshold values as discussed above. The threshold values of fuzzy rules should be within 0-1. The threshold values initialized as α, β and γ for the stratum property which has four and five features, respectively and τ, μ, ε and ρ for the stratum property which has six features. The stratum three has manual features which the threshold value assigned as η. The threshold values taken in this experiment are $\alpha = 1$, $\beta = 0.75$ and $\gamma = 0.50$, $\tau = 1$, $\mu = 0.80$, $\varepsilon = 0.70$ and $\rho = 0.60$ and $\eta = 0.5$ (for manual feature constant). The sample classification feature data are tabularized in Table 1 which contains a sample calculation of classification percentage by taking number of fuzzy feature set values.

Table 1 shows the classification percentage of proposed NEFCLASS Back Propagation Algorithmic Model in e-Banking phishing webpage clarifies with feature importance. It clearly shows the importance of each stratum feature to detect the phishing page. A point should be noted that all the phishing features in all stratum property doesn't present in all phishing webpage. Most probably, a phishing webpage doesn't have security properties, i.e., HTTPS protocol. The manual feature is set to a constant threshold value. Based on these feature sets the evaluation is carried out.

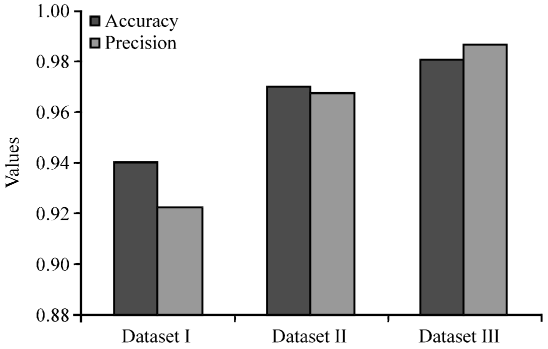

Fig. 5: Accuracy vs. precision

Evaluation metrics: This study aim is to classify phishing webpage from the cluster of e-Banking webpage. The evaluation metrics is carried out for the known URLs that calculates true positive, true negative, false positive and false negative. Here, the phished webpage are categorized under true because the idea is to detect the phished webpage and False is for legitimate webpage. The true positive is number of webpage that classify as phishing, true negative is number webpage judged legitimate webpage as legitimate webpage, false positive is number of legitimate webpage instances wrongly classified as phishing and false negative is number of phishing webpage shows as legitimate webpage.

Accuracy vs. precision: The term accuracy and precision have same dictionary but different conceptual or testing meaning. Accuracy is the measure of standard true values that gives accurate classification rate. Precision is the degree of measurement that the classification webpage are very close to one another (Fig. 5):

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{7}$$

Sensitivity vs. specificity: Sensitivity is the rate of true positive that indicates the ability of classification rate for

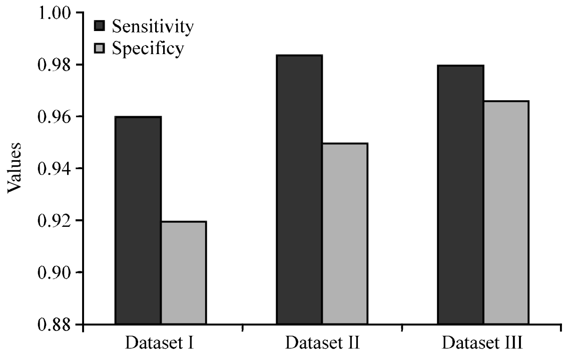

Fig. 6: Sensitivity vs. specificity

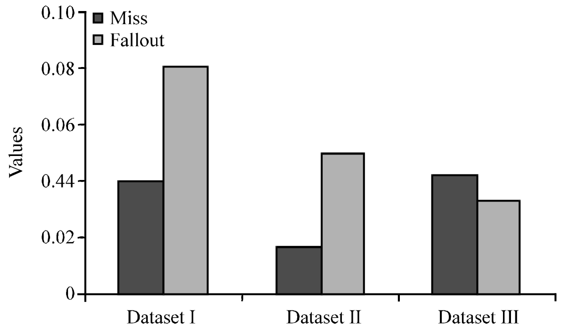

Fig. 7: Miss vs. fallout

phishing webpage. Sensitivity is otherwise called as recall. Specificity is the rate of true negative that indicates the ability of classification rate of legitimate webpage (Fig. 6):

$$\text{Sensitivity} = \frac{TP}{TP+FP} \tag{8}$$

$$\text{Sepecificity} = \frac{TN}{TN+FP} \tag{9}$$

Miss vs. fallout: Miss is the rate of false negative values that indicates the rate of phished webpage is wrongly classified as legitimate webpage. Fallout is the rate of false positive values that indicates the rate of legitimate webpage is wrongly classified as phished webpage (Fig. 7):

$$\text{Miss} = \frac{FN}{TP+FP} \tag{10}$$

$$\text{Fallout} = \frac{FP}{TN+FP} \tag{11}$$

RPP vs. RNP: The Rate of Positive Prediction (RPP) indicates predicted positives value rate as well as the Rate of Negative Prediction (RNP) indicates predicted negative value rate (Fig. 8):

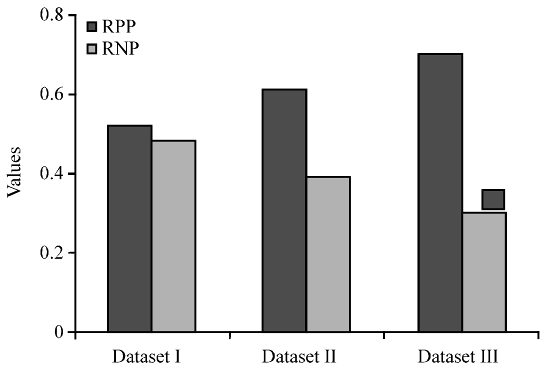

Fig. 8: RPP vs. RNP

Table 2: Performance metrics values for each datasets

Parameters	Dataset I	Dataset II	Dataset III
Accuracy	0.9400	0.9700	0.980
Precision	0.9231	0.9672	0.980
Sensitivity/Recall	0.9600	0.9833	0.980
Specificity	0.9200	0.9500	0.966
Miss	0.0400	0.0166	0.042
Fallout	0.0800	0.0500	0.033
RPP	0.5200	0.6100	0.700
RNP	0.4800	0.3900	0.300

$$\text{RPP} = \frac{TP+FP}{TP+FN+FP+TN} \tag{12}$$

$$\text{RNP} = \frac{TN+FN}{TP+FN+FP+TN} \tag{13}$$

The performance metrics such as accuracy, precision, sensitivity/recall, specificity, miss, recall, rate of positive prediction and rate of negative prediction shows that proposed NEFCLASS phishing webpage classification is highly accurate with very low error rate. And these values are tabularized in Table 2.

CONCLUSION

Phishing webpage ensembles the appearance and properties of legitimate webpage that phishers uses as lure to get the confidential information from the customers. This syudy categorize thirty term frequency phishing features into three different stratum are analyzed. The proposed NEFCLASS Back Propagation Model provides the efficiency of both effective decision making fuzzification and accurate classification tendency of neural network structure.

The empirical studies on the real phishing webpage collection datasets from the PhishTank demonstrates an effective classification of phishing webpage from cluster of e-Banking webpage. The performance of the

NEFCLASS Back Propagation Model is evaluation metrics. This evaluation metrics shows good classification with very low error rate. Thus, NEFCLASS Back Propagation Classification Model is an accurate method that classifies e-Banking phishing webpage.

REFERENCES

Aburrous, M., M.A. Hossain, K. Dahal and F. Thabtah, 2010a. Experimental case studies for investigating e-Banking phishing techniques and attack strategies. Cogn. Comput., 2: 242-253.

Aburrous, M., H.K. Dahal and F. Thabtah, 2010b. Intelligent phishing detection system for e-Banking using fuzzy data mining. Exp. Syst. Appl., 37: 7913-7921.

Al-Khatib, A.M., 2006. E-Banking: Survey. Int. J. Adv. Res. Compu. Sci. Soft. Eng., Vol. 2.

Damodaram, R. and M.L. Valarmathi, 2012. Phishing webpage detection and optimization using modified bat algorithm. Int. J. Eng. Res. Appl., 2: 870-876.

Dhanalakshmi, R., C. Prabhu and C. Chellapan, 2011. Detection of phishing webpage and secure transactions. Int. J. Commun. Network Secur., Vol. 1.

Dong, X., J.A. Clark and J.L. Jacob, 2010. Defending the weakest link: Phishing websites detection by analysing user behaviours. Telecommun. Syst., 45: 215-226.

Fu, A.Y., W.Y. Liu and X.T. Deng, 2006. Detecting phishing webpage with visual similarity assessment based on Earth Mover's Distance (EMD). IEEE Trans. Depend. Secure Comput., 3: 301-311.

Gaurav, M. Mishra and A. Jain, 2012. A Preventive anti-phishing technique using pattern matrix. Int. J. Eng. Res., and Appl., 2: 1825-1828.

Kim, Y.G. and S. Cha, 2011. Webpage risk assessment system for anti-phishing. Future Info. Technol. Commun. Comput. Info. Sci., 185: 131-138.

Komiyama, K., T. Seko, Y. Ichinose, K. Kato, K. Kawano and H. Yoshiura, 2010. In-depth evaluation of content-based phishing detection to clarify Its strengths and limitations. U-and E-Serv. Sci. Technol. Commun. Comput. Info. Sci., 124: 95-106.

Martin, A., N.B. Anutthamaa, M. Sathyavathy, M.M.S. Francois and P. Venkatesan, 2011. A framework for predicting phishing webpage using neural networks. Int. J. Comput. Sci., 8: 330-336.

Singh, A.P., V. Kumar, S.S. Sengar and M. Wairiya, 2011. Detection and prevention of phishing attack using dynamic watermarking. Info. Tech. Mobile Commun., 147: 132-137.

Wang, J., T. Herath, R. Chen, A. Vishwanath and R. Rao, 2012. Phishing susceptibility: An investigation into the processing of a targeted spear phishing email. IEEE Trans. Profess. Commun., 55: 345-362.

Zhang, H., G. Liu, T.W.S. Chow and W. Liu, 2011. Textual and visual content-based anti-phishing: A bayesian approach. IEEE Trans. Neural Networks, 22: 1532-1546.

Zhuang, W., Y. Ye, Y. Chen and T. Li, 2012. Ensemble clustering for internet security applications. IEEE Trans. Syst. Man Cyber. Part C: Appl. Rev., 42: 1784-1796.