2. RESEARCH FRAME WORK

2.1 Introduction

Today internet user community is remarkably large. In many business and service sectors such as finance transaction, banking, forensic investigation, modern communication system, entertainment, selling and purchasing of goods and social networks are require web page for instant references. This practice by the present public is inevitable. To deceive the identities of most frequently using popular pages are described as phishing pages. The phishing pages mimic almost equal as that of legitimate pages. This is a menace for the internet users and the business peoples of incur heavy financial losses.

Computer security providers often introducing advanced methods and techniques for the prediction of fast growing internet crime. Besides these methods are not proved worthy. Phishers are also creating new ways for the designing and development of these fake pages. This is because suitable tools are available for making duplicate pages. Hence this becomes a challenge for the web users and business and service providers. Web security researches are underway in this line to overcome this problem. In this present research work an attempt is made to introduce a technique for the better prediction of phishing pages and also a new novel authentication system for the prevention of viewing the phishing pages.
2.2 Generation of the Synchronized Feature Vector (SFV)

Two data sets of URLs of phishing and legitimate page are collected from internet web sources for this present study. The experimental dataset contains the list of URLs and their kind as phishing or legitimate. The URLs are allowed for the process of feature extraction. There are sixteen significant characteristics of web pages are considered. Based on the extracted features all the web pages are classified. Two thousand seven hundred URLs of web pages are collected and investigated.

There are 1498 phishing pages which have shown the features of fake web pages. On the other hand there are 502 legitimate pages. They have exhibited the real characteristics. The results of the above classification are used for the analysis and for the designing of multilayer authentication. Out of this 2700 collected pages 700 URLs are used as training data set (200 legitimate and 500 phishing pages) of experimental pages. A neural network based classifier is introduced for the classification of legitimate and phishing among this training data set. Each character of web page is assigned a value. The values assigned have represented the type of pages. (O- Suspicious, 1- Neutral,-1- Legitimate). By which the nature of 700 URLs are obtained. Once the training process is completed, the classifier is ready to test any web page with the feature vector. The classifier is well trained for the identification of labeled phishing page as phishing and legitimate as legitimate. This shows that the
classifier may be used for the prediction of an unknown web page kind whether it is original or not.

2.3 Classification of Phishing Pages

The present established synchronized feature extraction is also applied for the test data (2000 pages). Thereby all the entire data set is classified as phishing and non-phished pages. This is kept for future analysis. The identified test web pages are also studied for their features. Many algorithms are put forwarded for the prediction and prevention of illegitimate pages. Besides continuous studies on the development of various step wise algorithms, artificial neural network has been shown by many earlier workers for its efficiency (Ningxia and Yongqing, 2012; Luong et.al., 2015; Swetha and Radha, 2016). However it needs certain modification in this present context. Therefore in this present work the value of the 16 page features are summed up. The summation value, if it is exceeds 2 it indicates the nature of page as phishing one. Similarly if it is less than 2 it may be considered as genuine page. This small modification in neural network classifier gives precise labeling of legitimate as well as phishing pages. This is evidenced from the present analysis by following the stepwise proposed new algorithm. To substantiate the above phenomenon, a comparison is given between original data and the predicative data by adopting confusion matrix. The result has shown that there are 1495 phishing pages as true positive results. 6 phishing pages are looked like legitimates (True Negative). Likewise 3 legitimate pages predicated as phishing (FP) and 497 legitimate pages as legitimate page (TN)
out of 2000 page data set 1498 predicated as phishing a 502 pages as legitimate. These findings are subjected for precision, recall and accuracy value on the prediction efficiency. The results have a disclosed that there is 0.99798 precision value, 0.99666 recall value and 0.996 accuracy. In view of the above foregoing results it would be allowed to assume that the classifier functional efficiency is not promising. There is a negligible error in prediction of the web pages. The appearance of 3 false positive and 5 false negative should be eliminated to achieve 100% prediction of phishing pages. In this situation a multi layer authentication system is proposed in this present work. The following paragraph gives the details of the same and the assumption of the multi layer authentication system.

2.4 Multi Layer Authentications

Finally a multi layer authentication system is suggested for the prevention of illegitimate web pages. Generally user name, password, reading security answers and finger printing are in practice. Apart from this the introduction of one time password and session key are also recently added. In this present study along with above said security component an animated picture selection, mathematical expression evaluation are also added with above conventional security system for the prevention of phishing pages.
The overall plan of the present research proposal in flow diagram is given below.

2.5 Overview of the Research Work – Flow Diagram.

Fig.1. Outlay of the research work