**LITERATURE SURVEY**

In emerging technology, industry, which deeply influence today’s security problems, has given a headache to many employers and home users.

Occurrences that exploit human vulnerabilities have been on the upsurge in recent years. In these new times there are many security systems being enabled to ensure security is given the outmost priority and prevention to be taken from being hacked by those who are involved in cyber-offenses and essential prevention is taken as high importance in organization to ensure network security is not being compromised. Cyber security employee are currently searching for trustworthy and steady detection techniques for phishing websites detection. Due to wide usage of internet to perform various activities such as online bill payment, banking transaction, online shopping, etc. Customer face numerous security threats like cybercrime.

Many cybercrime is being casually executed for example spam, fraud, identity theft cyber terrorisms and phishing. Among this phishing is known as the most common cybercrime today. Phishing has become one amongst the top three most current methods of law breaking in line with recent reports, and both frequency of events and user weakness has increased in recent years, more combination of all these methods result in greater danger of economic damage. Phishing is a social engineering attack that targets and exploiting the weakness found in the system at the user’s end. This paper proposes the Agile Unified Process (AUP) to detect duplicate websites that can potentially collect sensitive information about the user. The system checks the blacklisted sites in dataset and learns the patterns followed by the phishing websites and applies it to further given inputs. The system sends a pop-up and an e-mail notification to the user, if the user clicks on a phishing link and redirects to the site if it is a safe website. This system does not support real time detection of phishing sites; user has to supply the website link to the system developed with Microsoft Visual Studio 2010 Ultimate and MySQL stocks up data and to implement database in this system. Phishing costs Internet user’s lots of money. It refers to misusing weakness on the user side, which is vulnerable to such attacks.

The basic ideology of the proposed solution is use to all the three-hybrid solution blacklist and whitelist, heuristics and visual similarity. The proposed system carries out a set of procedures before giving out the results. First, it tracks all “http” traffic of client system by creating a browser extension. Then compare domain of each URL with the white list of trusted domains and the blacklist of illegitimate domains. Further various characters in the URL is considered like number of ‘@’, number of ‘-‘and many more. Next approach is to extract and compare CSS of doubtful URL and compare it with the CSS of each of the legitimate domains in queue. This method will look into visual based features of the phished websites and machinelearning classifiers such as decision tree, logistic regression, random forest are applied to the collected data, and a score is generated. The match score and similarity score is evaluated. If the score is greater than threshold then the URL marked as phishing and blocked. This approach provides a three level security block. Phishing is a dangerous effort to steal private data from users like address, Aadhar number, PAN card details, credit or debit card details, bank account details, personal details etc.

The various types of phishing attacks like spoofing, instant spam spoofing, Hosts file poisoning, malware-based phishing, Manin-the middle, session hijacking, DNS based phishing, deceptive phishing, key loggers/loggers, Web Trojans, Data theft, Content-injection phishing, Search engine phishing, Email /Spam, Web based delivery, Link Manipulation, System reconfiguration, Phone phishing, etc. are discussed in the paper. The recent approaches to prevent the attacks like heuristics approach, blacklist approach, fuzzy rule-based approach, machine learning approach etc. are also discussed and finally filtering all detection techniques.

based on accuracy and performance proposed a framework to detect and prevent phishing attacks. A combination of supervised and unsupervised machine learning techniques is used to detect malicious attacks.

**2nd**

In this section, few of the research works that deploy the above mentioned algorithms are reviewed and their results are summarized.

In the paper [12], the authors Rishikesh Mahajan and Irfan Siddavatam chose three algorithms for classification Decision Tree, Random Forest and Support Vector Machine. Their dataset contained 17,058 benign URLs and 19,653 phishing URLs collected from Alexa website and PhishTank respectively, with 16 features each. The dataset was divided into training and testing set in the ratios 50:50, 70:30 and 90:10 respectively. The accuracy score, false negative rate and false positive rate were considered as performance evaluation metrics. They achieved 97.14% accuracy for Random Forest algorithm with the lowest false negative rate. The paper concluded that accuracy increases when more data is used for training.

The study conducted by Jitendra Kumar et al. in [13] trained different classifiers like Logistic Regression, Naive Bayes Classifier, Random Forest, Decision Tree and K- Nearest Neighbor based on the features extracted from the lexical structure of the URL. They created the dataset of URLs in such a way that it solved the issues of data imbalance, biased training, variance and overfitting. The dataset contained an equal number of labeled phishing and legitimate URLs, and was further split in the ratio 7:3 for training and testing. All the classifiers had almost the same AUC (area under ROC curve), but the Naive Bayes Classifier turned out to be more suitable as it had the highest AUC value. Naive Bayes achieved the highest accuracy of 98% with a precision=1, recall=0.95 and F1-score=0.97.

Mehmet Korkmaz et al. proposed in [14] a machine- learning based phishing detection system by using 8 different algorithms on three different datasets. The algorithms used were Logistic Regression (LR), K-Nearest Neighbor(KNN), Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), XGBoost, Random Forest (RF) and Artificial Neural Network (ANN). It was observed that the models using LR, SVM and NB have low accuracy rate. In terms of training time, NB, DT, LR and ANN algorithms gave better results. They concluded that RF algorithm or ANN algorithm may be used because of less training time along with a high accuracy rate.

Mohammad Nazmul Alam et al. [15] proposed a system to detect phishing attacks using Random Forest and Decision Tree. The Kaggle dataset with 32 features was used along with

feature selection algorithms like principal component analysis (PCA). Feature selection reduces redundancy of data that is irrelevant or unnecessary in the dataset. The proposed model used REF, Relief-F, IG and GR algorithm for feature selection before applying PCA. Random Forest achieved an accuracy of 97%. It had less variance, and it could handle the over-fitting problem.

Abdulhamit Subasi et al. in [16] presented an intelligent phishing detection system using UCI dataset. Different machine learning tools namely, Artificial Neural Networks (ANN), K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), C4.5 Decision Tree, Random Forest (RF), and Rotation Forest (RoF) were used as classifiers for detection of phishing websites. The performance of proposed RF classifier was higher than the others in terms of accuracy, F-measure and AUC. RF was faster, robust and more accurate than the other classifiers.

The rest of the paper is organized as follows: In section IV, the characteristics of the datasets that are used for training machine learning models are discussed. Section V explains the feature extraction process. The parameters used for performance evaluation of algorithms are discussed in Section

1. The observations obtained from the survey are pointed out in Section VII. Section VIII concludes the paper.

1.DATASETS

Usually, the phishing website data is collected from Phish Tank or OpenPhish. PhishTank.com is a website where phishing URLs are detected and can be accessed via API call. Their data is used by companies like McAfee, Kaspersky, Mozilla and APWG. Since it does not store the content of webpages, it is a good source for URL based analysis[14]. The legitimate sites are generally collected from Alexas top- ranking websites database or from common-crawl. There are publicly available datasets like the UCI machine learning repository dataset used in [16] which contains 11,055 records, each record having 31 features and the Kaggle phishing dataset used in [15].

2.FEATURE EXTRACTION

URLs have certain characteristics and patterns that can be considered as its features. The Fig. 3 shows the relevant parts of a typical URL.

In case of URL based analysis for designing machine learning models, we need to extract these features in order to form a dataset that can be used for training and testing. There are four categories of features that are most commonly considered for feature extraction as in [18]. They are as follows:

Address Bar based features

Abnormal based features

HTML and JavaScript based features

Domain based features