**LITERATURE SURVEY**

**AN EFFECTIVE DETECTION APPROACH FOR PHISHING WEBSITES USING URL AND HTML FEATURES** (JANUARY 2022)

This section provides an overview of the proposed phishing detection techniques in the literature. Phishing methods are divided into two categories; expanding the user awareness to distinguish the characteristics of phishing and benign webpages14, and using some extra software. Software-based techniques are further categorized into list-based detection, and machine learning-based detection. However, the problem of phishing is so sophisticated that there is no definitive solution to efficiently bypass all threats; thus, multiple techniques are often dedicated to restraining particular phishing offenses.

List-based detection:

List-based phishing detection methods use either whitelist or blacklist-based techniques. A blacklist contains a list of suspicious domains, URLs, and IP addresses, which are used to validate if a URL is fraudulent. Simultaneously, the whitelist is a list of legitimate domains, URLs, and IP addresses used to validate a suspected URL. Wang et al.15, Jain and Gupta5 and Han et al.16 use a white list-based method for the detection of suspected URL

**WEB PHISHING DETECTION USING MACHINE LEARNING** (MARCH 2022)

The current circumstance is that the population's maturity has been wisecracked, causing them to unknowingly give their private information to hackers. Several banned websites have already been established to seem like that actual points of contact through obtaining stoners’ private information. Passcode, savings account, and shipping information are just a few examples. Late in 2016, the amount of hacking activities was at an all-time high since the company started monitoring this in 2004. The overall identified phishing attacks in 2016 were 1,609. This represents a 65 percent increase over 2015. Within the final quarter of 2004, there would be scamming attempts each month. Machine Learning was used to find the phishing website. The use of machine literacy to surround the supplied features is the basis of Grounded Malware Monitoring Systems. Features are generated by assembling items in a specific order, such as URLs, sphere names, website features, and website content.

Because of its nonlinear system, it has a high level of fashion ability in terms of web security, particularly for the detection of anomalies on internet spots. The features retrieved utilizing machine literacy approaches are compared to extracting features through URLs, primary law, or third-party services. A process of machine trust ability on a particularity meant for the reflection of the besieged deceit of stoners through electronic communication is a relevant approach for detecting these attacks.

This method can be used to find phishing websites or textbook dispatches sent over email to confuse the victims. This method was presented by S. Marchal et al. to distinguish Malicious URLs based on the assessment of legitimate point garcon record data. By the off operation or the detection of a malicious site. Open source demonstrates several remarkable characteristics, including high proximity, total autonomy, excellent linguistic flexibility, quickness in choosing, inflexibility towards active phishing, and inflexibility towards development in phishing methods. Mustafa Aydin et al. presented the bracket method to fraudulent site detection that involves rooted websites ‘URL properties and evaluating subset- grounded Point selection approaches. For the detection of phishing websites, it uses point birth and selecting styles.

**Phishing Website Detection Based on Deep Convolutional Neural Network and Random Forest Ensemble Learning** (JANUARY 2021)

Although attacks use different techniques to create phishing websites to deceive users, most have similarly designed phishing website features. Therefore, researchers have conducted extensive anti-phishing research using phishing website features. Current methods for phishing detection include black and whitelists, heuristics, visual similarity and machine learning, among which heuristics and machine learning are more widely used. The following is an introduction to the aforementioned phishing detection techniques.

Black and whitelist:

To prevent phishing attack threats, many anti-phishing methods have been proposed. Blacklisting methods are the most straightforward ways to prevent phishing attacks and are widely used in the industry. Google Safe Browsing uses a blacklist-based phishing detection method to check if the URL of the matching website exists in the blacklist. If it does, it is considered a phishing website; otherwise, it is a legitimate website. Jain and Gupta proposed an automatic update whitelist technique to prevent phishing attacks in 2016. This method uses the hyperlink function to check the legitimacy of webpages; it extracts the hyperlinks from the source code of the web pages when accessed, and applies them to a phishing detection algorithm. This method can effectively detect various types of phishing attacks. Lung-Hao and Kuei-Ching et al. proposed a framework to automatically update the blacklist of phishing websites, Phish Track, in 2014. This framework explores existing blacklists to discover suspicious URLs. Phish Track includes redirect and form tracking components to update phishing blacklists, and it proactively discovers phishing URLs as early as possible. This proactive phishing update approach effectively improves blacklist coverage and complements existing anti-phishing techniques to provide safe web surfing. Black and whitelist-based phishing detection has high accuracy and can prevent phishing attacks, to some extent. It has low system overhead for fast client access only if they are included in the blacklist of phishing websites. However, phishing websites can be created at any time and place, and their average life span may be a few hours; the black and white list-based phishing detection approach in itself has low efficiency in prevention from these types of attacks. Other technical means are needed to detect websites not detected by this method.

.

**PHISHING AND ANTI PHISHING TECHNIQUES** (AUGUST 2021)

When Kiren presented a review on different types of phishing attacks and detection techniques. Also they presented some mitigation techniques of phishing. The paper proposed that 100% accuracy to detect phishing can be made possible by using machine learning approach among all other anti-phishing approaches. Rana presented a review and comprehensive examination of the modern and state of the art phishing attack techniques to spread awareness of phishing techniques among the reader and to educate them about different types of attacks. The author proposed this paper to encourage the use of anti-phishing methods as well. Justine proposed a phishing attack taxonomy based on E-mail. which is covering all the drawbacks of already existing phishing taxonomies. It is concluded that the proposed taxonomy has broader classification and depending on classes which are two times greater in numbers as compared to the classes of the existing taxonomies.

**PHISHING DETECTION USING MACHINE LEARNING BASED URL ANALYSIS** (DECEMBER 2021)

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 Phish Tank 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, and variance and over fitting. 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.

**PHISHING ATTACK TECHNIQUES** (APRIL 2016)

Protecting user against phishing using Anti-phishing:

Anti-phish is used to avoid users from using fraudulent web sites which in turn may lead to phishing attack. Here, Anti phish traces the sensitive information to be filled by the user and alerts the user whenever he/she is attempting to share his/her information to an untrusted web site. The much effective elucidation for this is cultivating the users to approach only for trusted websites. However, this approach is unrealistic. Anyhow, the user may get tricked. Hence, it becomes mandatory for the associates to present such explanations to overcome the problem of phishing. Widely accepted alternative are based on the creepy websites for the identification of “clones” and maintenance of records of phishing websites which are in hit list.

Learning to Detect Phishing Emails:

An alternative for detecting these attacks is a relevant process of reliability of machine on a trait intended for the reflection of the besieged deception of user by means of electronic communication. This approach can be used in the detection of phishing websites, or the text messages sent through emails that are used for trapping the victims. Approximately, 800 phishing mails and 7,000 non-phishing mails are traced till date and are detected accurately over 95% of them along with the categorization on the basis of 0.09% of the genuine emails. We can just wrap up with the methods for identifying the deception, along with the progressing nature of attacks.