**CHAPTER 1**

**INTRODUCTION**

With the widespread usage of the internet for online banking and trade, phishing attacks and forms of identity theft-based frauds are becoming extremely popular among hacker communities. In 2004 alone, more than fifty million phishing emails were sent. Their result was ten billion dollars of damage to banks and financial institutions. Most of the recent phishing attacks are carried out as a three-step process. In the first step, the phishers send emails to their victims from social engineering attacks, webpages, and forums. Large volumes of phishing emails with legal banking domains are sent out using anonymous servers or compromised machines.[1] These emails contain hyperlinks with an appearance like a legitimate website. The fake webpage contains input forms requesting personal critical information such as credit card, social security numbers, mother's maiden name, etc. Although existing spam filtering techniques can be employed to combat phishing emails, these measures are not entirely scalable. Several readily available tools can bypass both the statistical and rule-based spam filters. As these mechanisms are not uniquely tuned for the detection of phishing emails despite their existence, the threats caused by phishing emails are prevalent. Furthermore, unlike spamming, which impacts bandwidth, phishing attacks directly affect their victims by inflicting a thefty loss due to monetary damage. Moreover, attackers can use technical vulnerabilities to construct socially engineered messages (i.e., use of legitimate, but spoofed, domain names can be far more persuasive than using different domain names), which makes phishing attacks a severe problem. Effective mitigation would require addressing issues at the technical and human layers. Since phishing attacks aim at exploiting weaknesses found in humans (i.e., system end-users), it is difficult to mitigate them. For example, as evaluated, end-users failed to detect 29% of phishing attacks, even when trained with the best performing user awareness program. On the other hand, software phishing detection techniques are evaluated against phishing attacks, which makes their five performances practically unknown with targeted forms of phishing attacks. These limitations in phishing mitigation techniques have almost resulted in security breaches against several organizations, including leading information security providers. Now days, there are so many people are being aware of using internet to perform various activities like online shopping, online bill payment, online mobile recharge, banking transaction. Due to wide use of this customer face various security threats like cybercrime. There are many cybercrimes that are widely performed for example spam, fraud, cyber terrorisms, and phishing. Among this phishing is new cybercrime and very popular nowadays. Phishing is fraud attempt, which performed to obtain sensitive information of user. Phisher design website which looks same as any legitimate site and spoof user for obtaining private information of user such as username, password, banking details for miscellaneous reasons. According to APWG 2Q report, the total number of phish detected in 2Q 2018 was 233,040, compared to 263,538 in 1Q 2018. These totals exceed the 180,577 observed in 4Q 2017 and the 190,942 seen in 3Q 2017. There were increases in SAAS/webmail targeted sector with 21% of overall phishing attack. Payment sector is continuing as most attractive target for phishing. According to APWG 1Q report, the total number of phish detected in 1Q 2018 was 263,538. This was up 46 percent from the 180,577 observed in 4Q 2017. It was also significantly more than the 190,942 seen in 3Q 2017. The number of unique phishing reports submitted to APWG during 1Q 2018 was 262,704, compared to 233,613 in 4Q 2017 and 296,208 in 3Q 2017.

1. **Problem Statement**

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them, but it is very important to enhance phishing detection techniques. There are too many approaches to detect a phishing website we will use Machine Learning to do. We use few machine learning algorithms like Decision Tree Classifier, Random Forest and XGBoost.

1. **Industry / Society benefited**
2. The Project can be implemented by many E-Commerce or other websites in order to have good customer relationship.
3. Users can make online payment securely.
4. Eliminate the cyber threat risk level.
5. Increase user alertness to phishing risks.
6. In still a cyber security culture and create cyber security heroes.
7. Change behaviours to eliminate the automatic trust response.
8. **Proposed Solution**

Machine learning technique detects phishing sites based on markup visualization. Machine learning models trained on the visual representation of website code can help improve the accuracy and speed of detecting phishing websites. Use anti-phishing protection and anti-spam software to protect yourself when malicious messages slip through to your computer. Anti-malware is included to prevent other types of threats. Similar to anti-spam software, anti-malware software is programmed by security researchers to spot even the stealthiest malware.

1. **Feasibility**

The Phishing Detection System is the best technique to distinguish between the website is legitimate or a phishing website. Machine learning technique detects phishing sites based on markup visualization. Machine learning models trained on the visual representation of website code can help improve the accuracy and speed of detecting phishing websites.

**1.5 Literature Survey**

Rao et al. [2] proposed a novel classification approach that use heuristic based feature extraction approach. In this, they have classified extracted features into three categories such as URL Obfuscation features, Third-Party-based features, Hyperlink-based features. Moreover, proposed technique gives 99.55% accuracy. Drawback of this is that as this model uses third-party features, classification of website dependent on speed of third-party services. And also this model is purely depends on the quality and quantity of the training set and Broken links feature extraction has a limitation of more execution time for the websites with more number of links.

Chunlin et al. [3] proposed approach that primarily focus on character frequency features. In this paper, they have combined statistical analysis of URL with machine learning technique to get result that is more accurate for classification of malicious URLs. Also they have compared six machine-learning algorithms to verify the effectiveness of proposed algorithm which gives 99.7% precision with false positive rate less than 0.4%.

Sudhanshu et al. [4] used association data mining approach. They have proposed rule based classification technique for phishing website detection. They have concluded that association classification algorithm is better than any other algorithms because of their simple rule transformation. They achieved 92.67% accuracy by extracting 16 features but this is not up to mark so proposed algorithm can be enhanced for efficient detection rate.

M. Amaad et al. [5] presented a hybrid model for classification of phishing website. In this paper, proposed model carried out in two phase. In phase 1, they individually perform classification techniques, and select the best three models based on high accuracy and other performance criteria. While in phase 2, they further combined each individual model with best three model and makes hybrid model that gives better accuracy than individual model. They achieved 97.75% accuracy on testing dataset. There is limitation of this model that it requires more time to build hybrid model.

Hossein et al. [6] developed an open-source framework known as “Fresh-Phish”. For phishing websites, machine-learning data can be created using this framework. In this, they have used reduced features set and using python for building query. They build a large labelled dataset and analyse several machine-learning classifiers against this dataset. Analysis of this gives very good accuracy using machine learning classifiers. These analyses how long time it takes to train the model.

Gupta et al. [7] proposed a novel anti phishing approach that extracts features from client-side only. Proposed approach is fast and reliable as it is not dependent on third party but it extracts features only from URL and source code. In this paper, they have achieved 99.09% of overall detection accuracy for phishing website. This paper have concluded that this approach has limitation as it can detect webpage written in HTML. Non-HTML webpage cannot detect by this approach.

Bhagyashree et al. [8] proposed a feature based approach to classify URL’s as phishing and non-phishing. Various features this approach uses are lexical features, WHOIS features, Page Rank and Alexa rank and Phish Tank-based features for disguising phishing and non-phishing website.

Mustafa et al. [9] developed safer framework for detecting phishing website. They have extracted URL features of website and using subset based selection technique to obtain better accuracy. In this paper, author evaluated CFS subset based and content based subset selection methods and Machine learning algorithms are used for classification purpose.

Priyanka et al. [10] proposed novel approach by combining two or more algorithms. In this paper, author has implemented two algorithm Adaline and Backpropotion along with SVM for getting good detection rate and classification purpose.

Pradeepthi et al. [11] In this paper, author studied different classification algorithm and concluded that tree-based classifiers are best and gives better accuracy for phishing URL detection. The author also uses various features such as lexical features, URL based feature, network based feature and domain based feature.

Luong et al. [12] proposed new technique to detect phishing website. In proposed method, Author used six heuristics that are primary domain, sub domain, path domain, page rank, and alexa rank, alexa reputation whose weight and values are evaluated. This approach gives 97% accuracy but still improvement can be done by enhancing more heuristics.

Ahmad et al. [13] proposed three new features to improve accuracy rate for phishing website detection. In this paper, Author used both type of features as commonly known and new features for classification of phishing and non-phishing site. At the end author has concluded this work can be enhanced by using this novel feature with decision tree machine learning classifiers.

Mohammad et al. [14] proposed model that automatically extracts important features for phishing website detection without requiring any human intervention. Author has concluded in this paper that the process of extracting feature by their tool is much faster and reliable than any manual extraction.