Project Design Phase-I
Proposed Solution

Date	08 October 2022
Team ID	PNT2022TMID07498
Project Name	Project – Web Phishing Detection
Maximum Marks	2 Marks

Proposed Solution:

The proposed solution for the web phishing detection has the following parameters and their respective descriptions:

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	'Phishing sites' are some type of the internet security issues that mainly targets the human vulnerabilities compared to software vulnerabilities. Phishing sites are malicious websites that imitate as legitimate websites or web pages and aim to steal user's personal credentials like user id, password, and financial information. Spotting these phishing websites is typically a challenging task because phishing is mainly a semantics-based attack, that mainly focus on human vulnerabilities. Our main aim of this project is classification of a phishing website with the aid of various machine learning techniques to achieve maximum accuracy and concise model.
2.	Idea / Solution description	• The technique comprises of host based, page based and lexical feature extraction of collected websites. The primary step is the collection of phishing and benign websites. In the host-based approach, admiration based and lexical based attributes extractions are performed to form a database of attribute value. This database consists of knowledge mined that uses different machine learning techniques. On evaluating the algorithms, a selective classifier is opted and is implemented in Python.

		• The proposed system collected URLs of benign websites from www.alexa.com www.dmoz.org and personal web browser history. The phishing URLs were collected from www.phishtak.com . The data set consists of 17000 phishing URLs and 20000 benign URLs.
3.	Novelty / Uniqueness	The proposed system used the dataset provided by UCI Machine Learning repository[4] collated by Mohammad et al[3]. The dataset has 11055 datapoints with 6157 legitimate URLs and 4898 phishing URLs. Each datapoint had 30 features subdivided into following three categories: • URL and derived features • Page's souce code based features: Includes URLs embedded in the webpage and HTML and Javascript based features. • Domain based features
4.	Social Impact / Customer Satisfaction	The project helped most of the public(users)to find whether the website is a phishing website or not.It helped them to categorize the dangerous sites.The algorithms used in this project are Machine Learning algorithms.The URL is copied and pasted in the input and it will detect and give the users an accurate result.
5.	Business Model (Revenue Model)	In literature, several approaches have been proposed for detection and filtering phishing attack. However, researchers are still searching for such a solution that can provide

		better results to secure users from phishing attack. Phishing websites have certain characteristics and patterns and to identify those features can help us to detect phishing. To identify such features is a classification task and can be solved using machine learning techniques
6.	Scalability of the Solution	This project provides an efficient solution for phishing detection that extracts the features from website's URL and HTML source code. Specifically, we proposed a hybrid feature set including URL character sequence features without expert's knowledge, various hyperlink information, plaintext and noisy HTML data-based features within the HTML source code. . Extensive experiments show that the proposed anti-phishing approach has attained competitive performance on real dataset in terms of different evaluation statistics. Our anti-phishing approach has been designed to meet the following requirements. • High detection efficiency • Real-time detection • Target independent • Third-party independent

Contributed by,

Navaneetha Krishnaveni.A	721719104056
Subavarshini.S	721719104081
Varshini.P	721719104089
Mythili.R	721719104053