Project Design Phase-I
Proposed Solution Template

Date	17 November 2022
Team ID	PNT2022TMID46905
Project Name	Project - Web Phishing Detection
Maximum Marks	2 Marks

Proposed Solution Template:

Project team shall fill the following information in proposed solution template.

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	As opposed to software vulnerabilities, "phishing sites" are a particular kind of internet security problems that primarily target human vulnerabilities. Phishing sites are harmful websites that pretend to be trustworthy websites or web pages in order to steal users' personal information, including their user name, password, and credit card number. Since phishing is mostly a semantics-based attack that focuses on human vulnerabilities, identifying these phishing websites can be difficult. The main goal of this project is to classify phishing websites using a variety of machine learning approaches in order to produce a model with the highest level of accuracy and simplicity.

2.	Idea / Solution description	● The method includes the extraction of lexical features from collected webpages as well as host- and page-based feature extraction. The first stage is gathering phishing and legitimate websites. In the host-based technique, attribute extractions based on admiration and lexical bases are carried out to create a database of attribute value. This database contains knowledge that has been extracted using various machine learning methods. A selective classifier is chosen after comparing the methods, and it is put into practise in Python.

		● The suggested approach gathered URLs of safe websites from sites like www.alexa.com, www.dmoz.org, and browsing history. We gathered the phishing URLs from www.phishtak.com. 20000 benign URLs and 17000 phishing URLs make up the data collection.
3.	Novelty / Uniqueness	The dataset provided by UCI Machine Learning repository4 and compiled by Mohammad et al3 was used by the suggested system. The dataset contains 6157 legal URLs and 4898 phishing URLs across 11055 datapoints. Each data point had 30 features that were sorted into the three categories below: ● Features extracted from the URL ● Features based on the page's source code, such as URLs that are incorporated into the webpage and HTML and Javascript-based features. ● Features based on domains.

4.	Social Impact / Customer Satisfaction	The majority of the public (users) were assisted by the project in determining if a website was a phishing website or not. It assisted them in classifying the hazardous locations. Machine learning methods were employed in this research. The URL is entered, and it will recognise it and provide users with precise results.
5.	Business Model (Revenue Model)	In the literature, a number of methods for phishing attack detection and filtering have been suggested. Researchers are still looking for a solution that can protect consumers from phishing attacks and produce better outcomes. It might be easier to spot phishing websites if we can recognise the specific traits and patterns they exhibit. The classification problem of identifying such traits can be resolved using machine learning approaches.

6.	Scalability of the Solution	This project offers an effective method for phishing detection that pulls features from the URL and HTML source code of websites. In particular, we suggested a hybrid feature set that included features for the HTML source code's plaintext and noisy HTML data, different hyperlink information, and URL character sequence characteristics without the knowledge of experts. The suggested anti-phishing technique has demonstrated competitive performance on actual datasets in terms of several assessment statistics, according to extensive trials. The following criteria have been established for our anti-phishing strategy. ● Target independent ● Real-time detection ● High detection efficiency ● Third-party independent