# AN EFFECTIVE APPROACH FOR WEB PHISHING DETECTION USING THE MACHINE-LEARNING ALGORITHM WITH DATA MINING PROCESS TOWARDS DATA SCIENCE


## ABSTRACT

Phishing is a common attack against Internet users that causes them to revealtheir information using fake websites. The goal of the fake website is to steal personal information such as usernames, passwords and online banking transactions. Scammers use websites that are visually and semantically similarto the real ones.

As technology continues to advance, phishing techniques begin to advance rapidly, andthis should be prevented by using anti-phishing mechanisms such as spoofed URL detection. Machine Learning is a powerful tool used to combat spoofing attacks. This report covers machine learning technology to detect fake URLs by extracting and analyzing different characteristics of legitimate and fake URLs. Random Forest, LogisticRegression and algorithms are used to detect fake websites.


## Introduction

Nowadays, the Internet plays an important role in communication, where people create an online environment to manage business functions, online activities of banks, social networks… However, the Internet also contains hidden things. a lot of risk because when users operate in an online environment they can be vulnerable to attackers. And their identity is often a fake URL. And spoofed URLsare often placed on popular websites or sent to user email.

## Literature Review

*Construction of Phishing Site*. In the first step attacker identifies the target as a well-known organization. Afterward, attacker collects the detailed information about the organization by visiting their website. The attacker then uses this information to construct the fake website

URL Sending. In this step, attacker composes a bogus e-mail and sends it to the thousands of users. Attacker attached the URL of the fake website in the bogus e-mail. In the case of spear phishing attack, an attacker sends the e-mail to selected users. An attacker can also spread the link of phishing website with the help of blogs, forum, and so forth [43].

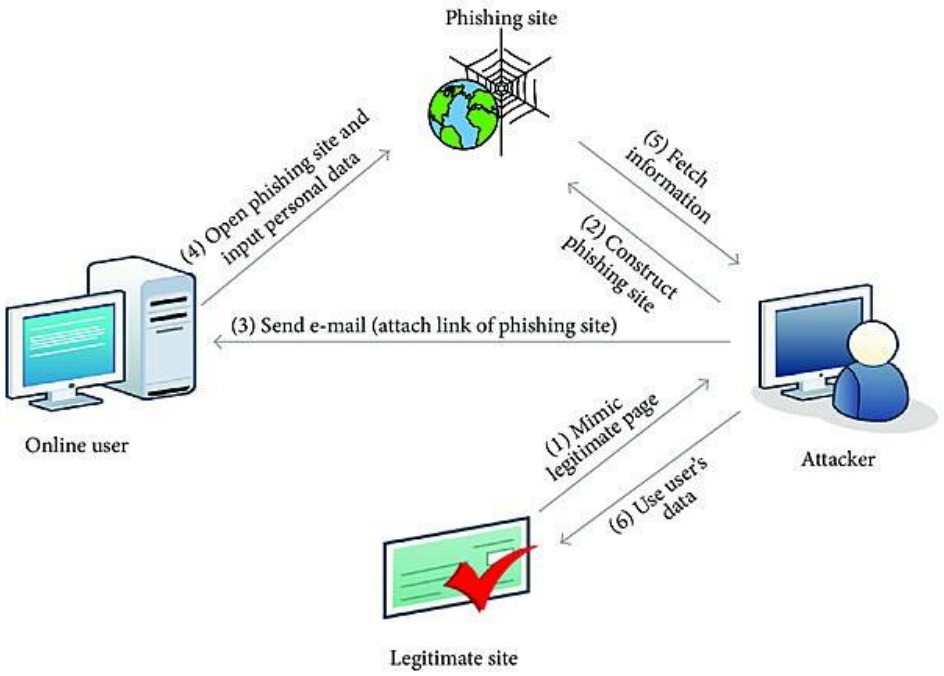
Stealing of the Credentials. When user clicks on attached URL, consequently, fake site is opened in the web browser. The fake website contains a fake login form which is used to take the credential of an innocent user. Furthermore, attacker can access the information filled by the user.

Identity Theft. Attacker uses this credential of malicious purposes. For example, attacker purchases something by using credit card details of the user.

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.




**FEATURES OF PROPOSED SYSTEM:**

1. **FUNCTIONAL CAPABILITIES:** The ultimate aim of this project is to detect phishing attacks in real-time. This model checks the website with machine learning server for any maliciousness in the accessed site.
2. **PERFORMANCE LEVEL:** At the client side, it takes 1-2 seconds to detect whether a site is phishing or not. **DATA STRUCTURES:** The data in this project are maintained in the CSV form. It provides easy access to the user.
3. **SAFETY:** No data loss occurs in this system.
4. **RELIABILITY:** We assure that the project is completely authenticated in order to enhance security and corruptions of database as well as the software.

**QUALITY:** The project is developed with the help of Anaconda Navigator softwarewhich meets the requirement of the user, the project is checked whether the phases individually have a served its purple.