Phish Catcher Client-Side Defense Against Web Spoofing Attacks Using Machine Learning

BITS ZC4999T: Capstone Project

by

Shalini Jindam

202117B3557

Capstone Project work carried out at

HCL Tech Ltd., Nagpur

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE
PILANI (RAJASTHAN)

October 2025

Birla Institute of Technology & Science, Pilani

Work-Integrated Learning Programmes Division

BITS ZC499T Capstone Project Outline

ID No.	**: 202117B3557**
NAME OF THE STUDENT	**: Shalini Jindam**
EMAIL ADDRESS	**: 202117B3557@wilp.bits-pilani.ac.in**
STUDENT'S EMPLOYING	**: HCL Tech Ltd., Nagpur**
ORGANIZATION & LOCATION	
MENTOR'S NAME	**: Mohiddin Shaik**
MENTOR'S EMPLOYING	**: HCL Tech Ltd., Hyderabad**
ORGANIZATION & LOCATION	
MENTOR'S EMAIL ADDRESS	**: mohiddin_s@hcltech.com**
CAPSTONE PROJECT TITLE	**: PhishCatcher: ML-Based Web Spoofing Detection**

_____

Contents

1. Broad Academic Area of Work

Cybersecurity / Web Spoofing Detection / Machine Learning

2. Background

Web spoofing attacks, commonly known as phishing, continue to threaten digital security by imitating legitimate web pages to steal sensitive user information such as passwords, PINs, and personal identifiers. Despite the availability of multiple anti-phishing solutions, challenges such as high false-positive rates, latency issues, and incomplete feature utilization limit their effectiveness. Traditional systems often rely on server-side validation or single-feature detection approaches, making them insufficient against modern, sophisticated attacks.

To overcome these limitations, the PhishCatcher project introduces a stateless client-side defense mechanism implemented as a Google Chrome extension. Using machine learning techniques—specifically a Random Forest classifier—the system analyzes multiple web features including URL patterns, content attributes, and visual elements to classify web pages as legitimate or spoofed. Experimental evaluations conducted over 1500 URLs demonstrated an impressive 98.5% accuracy and a response time of just 62.5 ms, proving the system's reliability and efficiency

3. Objectives

The main goals of this project are:

1. Develop a client-side browser extension to detect spoofed or malicious login pages.
2. Implement a machine learning model (Random Forest) capable of classifying URLs as legitimate or suspicious.
3. Ensure high accuracy, precision, and low latency in phishing detection.
4. Select and extract relevant web features including URL attributes, web content, visual structure, and blacklist checks.
5. Avoid reliance on server-side mechanisms to enhance speed and preserve user privacy.
6. Evaluate system performance through experiments on real-world phishing and legitimate URLs.

4. Scope of Work

- Study phishing attack patterns and identify relevant web attributes.
- Collect and label datasets of legitimate and phishing URLs.
- Extract URL-based, content-based, visual-based, and blacklist-based features.

- Train and optimize a Random Forest classifier.
- Develop a Google Chrome extension integrating the ML model.
- Conduct experimental evaluation for accuracy, precision, recall, and latency.
- Prepare documentation and usage guidelines.

SOFTWARE REQUIREMENTS:

Operating system	:	Windows 7 Ultimate.
Coding Language	:	Python.
Front-End	:	Python.
Back-End	:	Django-ORM
Designing	:	Html, CSS, JavaScript.
Data Base	:	MySQL (WAMP Server).

5. Plan of Work

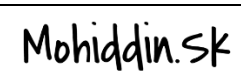
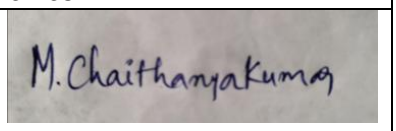
Week 1–2: Literature review, dataset preparation, feature identification.
Week 3–4: Model development using Random Forest classifier.
Week 5–6: Development of Chrome extension infrastructure.
Week 7–8: Integration of the ML model with the browser extension.
Week 9–10: Performance evaluation and optimization.
Week 11–12: Testing with real URLs and refining detection accuracy.
Week 13–14: Documentation and report preparation.

6. Literature Reference

Studies referenced include SpoofCatch (visual similarity-based detection), TF-IDF content analysis models, Gestalt-based web page analysis, RIPPER classification for email phishing detection, machine learning models such as Linear Models, Decision Trees, Neural Networks, and hybrid RDF-based classification frameworks.

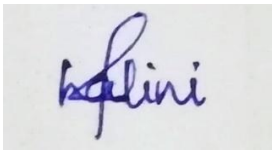
7. Particulars of Mentor and Examiner

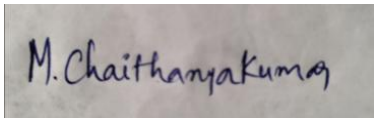
	Mentor	**Additional Examiner**
Name	Mohiddin Shaik	Malli Chaithanya Kumar
Qualification	B.Tech (EEE)	B.Tech (ECE)
Designation	SENIOR TECHNICAL LEAD	SOFTWARE ENGINEER
Employing Organization and Location	Hyderabad	Chennai

Phone No. (with STD Code)	8500161218	9052856746
Email Address	mohiddin_s@hcltech.com	mallichaithan.kumar@hcltech.com
Signature		
Date	20/11/2025	20/11/2025

8. Remarks of Mentor

EC No.	Component	Excellent	Good	Fair	Poor
1.	Capstone Project Outline				





Signature of Student **Signature of Mentor** **Signature of Additional Examiner**

Name: Shalini Jindam **Name: Mohiddin Shaik** **Name: Malli Chaithanya Kumar**