ML-based Phishing Website Detection

Name	ID
Redwan Ali Rafi	1921161642
Talha Taiyoub	1921949042
Nafsi Omar Prokrety	2022326642
Madiha Jarrin	2022223642

Introduction

In today's digital landscape phishing remains one of the most pervasive threats to individuals and organisations. It divulges sensitive information such as passwords, credit card numbers, and personal details. This project aims to develop a robust phishing detection system utilising machine learning algorithms and data analysis techniques by leveraging various data sources including URL features, source code.

Problem Definition

Phishing attacks represent one of the most pervasive cybersecurity threats, often resulting in significant financial and personal data breaches as malicious actors continually refine their techniques to deceive users. The rapid proliferation of phishing websites and the increasing sophistication of their tactics render traditional detection methods such as blacklisting heuristic-based approaches and rule-based systems largely ineffective. Consequently, There is a critical need for intelligent, adaptable and accurate systems that can detect phishing websites accurately. Therefore, we're designing a phishing detection system that utilises machine learning models focusing on two main features sets: URL-based & content-based. We demonstrate how combining both feature types enhances the model's ability to detect phishing sites in real-time.

Objectives

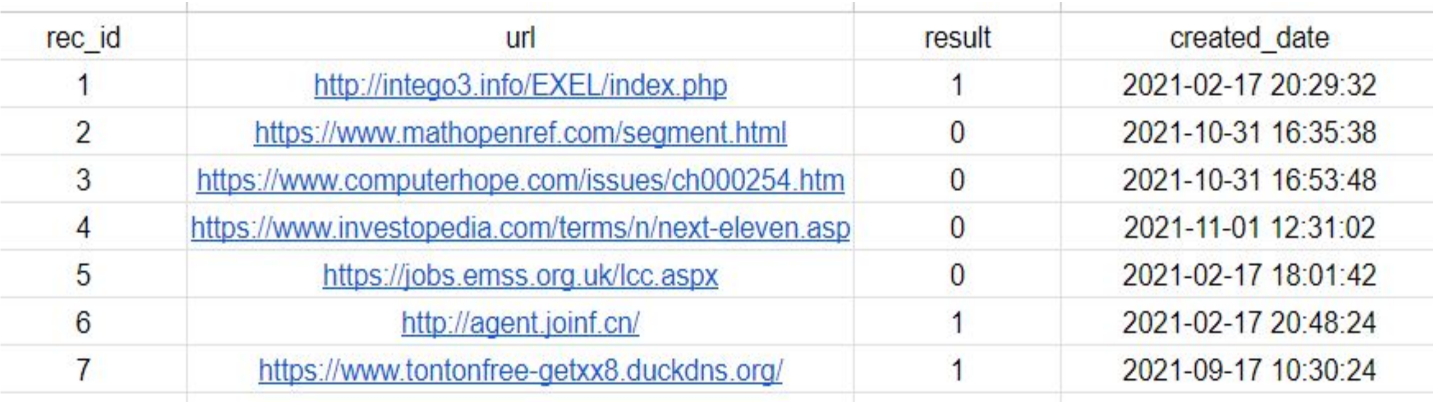
Develop a machine learning based phishing detection system that :

- Detects phishing and legitimate websites.
- Achieves high detection accuracy by extracting key features.
- Will protect users from providing sensitive information to malicious actors.

Related Work

Research on phishing website detection has explored many approaches. The list-based approach was quite useful, but it couldn't protect users from zero day attacks. To address this gap, various machine learning approaches have been introduced focusing on feature extraction from URLs, Transport Layer Security(TLS), and HTML content.

Dataset



rec_id	url	result	created_date
1	http://intego3.info/EXEL/index.php	1	2021-02-17 20:29:32
2	https://www.mathopenref.com/segment.html	0	2021-10-31 16:35:38
3	https://www.computerhope.com/issues/ch000254.htm	0	2021-10-31 16:53:48
4	https://www.investopedia.com/terms/n/next-eleven.asp	0	2021-11-01 12:31:02
5	https://jobs.emss.org.uk/lcc.aspx	0	2021-02-17 18:01:42
6	http://agent.joinf.cn/	1	2021-02-17 20:48:24
7	https://www.tontonfree-getxx8.duckdns.org/	1	2021-09-17 10:30:24

Source: Mendeley data
Size: 83275 instances
Classes:

- Legitimate website instances labeled as 0
- Phishing website instances labeled as 1

Technology Stack

ML Models: Random Forest, MobileNet(CNN), Support Vector Machine(SVM)

Frameworks: Tensorflow, Scikit-learn

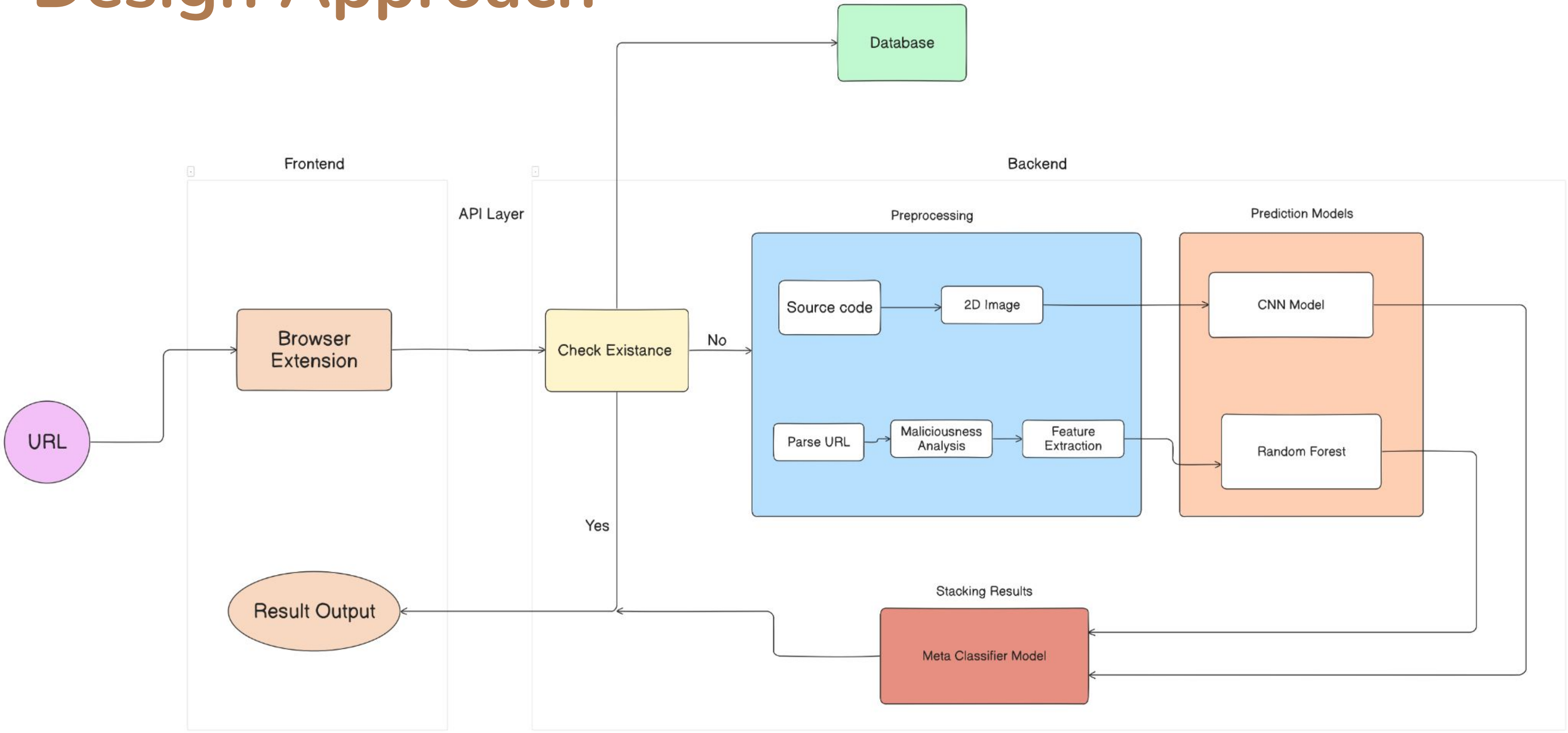
Frontend: React

Backend: Django

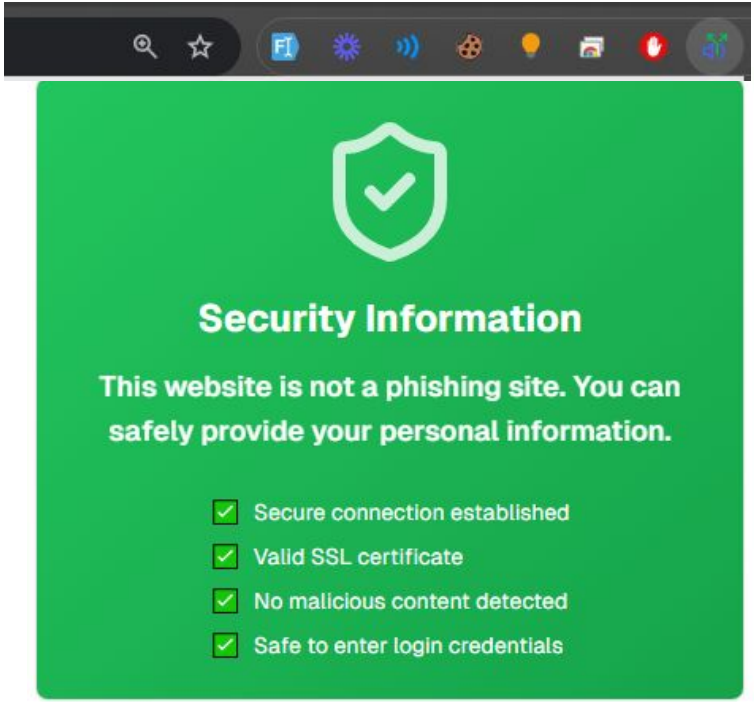
Database: MySQL

Data Processing: Pandas, Numpy etc.

Design Approach



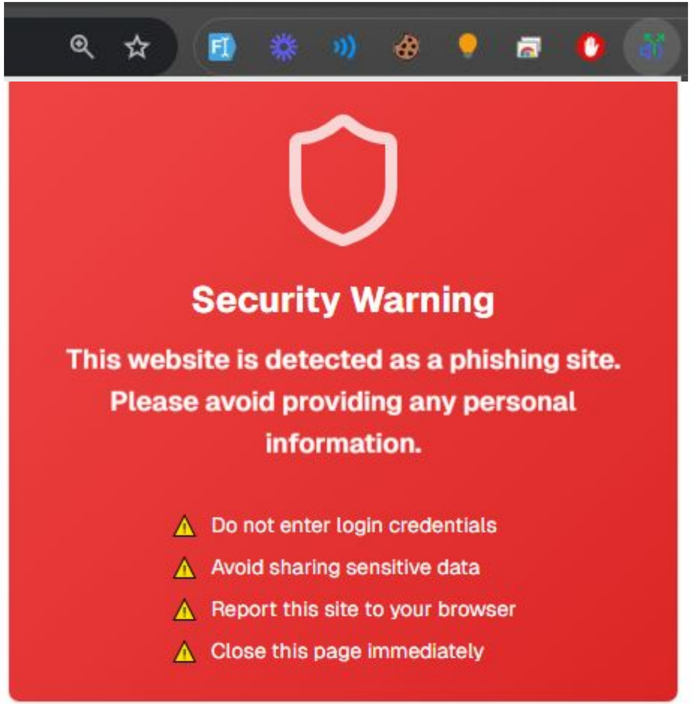
Demo User Interface



Security Information

This website is not a phishing site. You can safely provide your personal information.

- ☑ Secure connection established
- ☑ Valid SSL certificate
- ☑ No malicious content detected
- ☑ Safe to enter login credentials

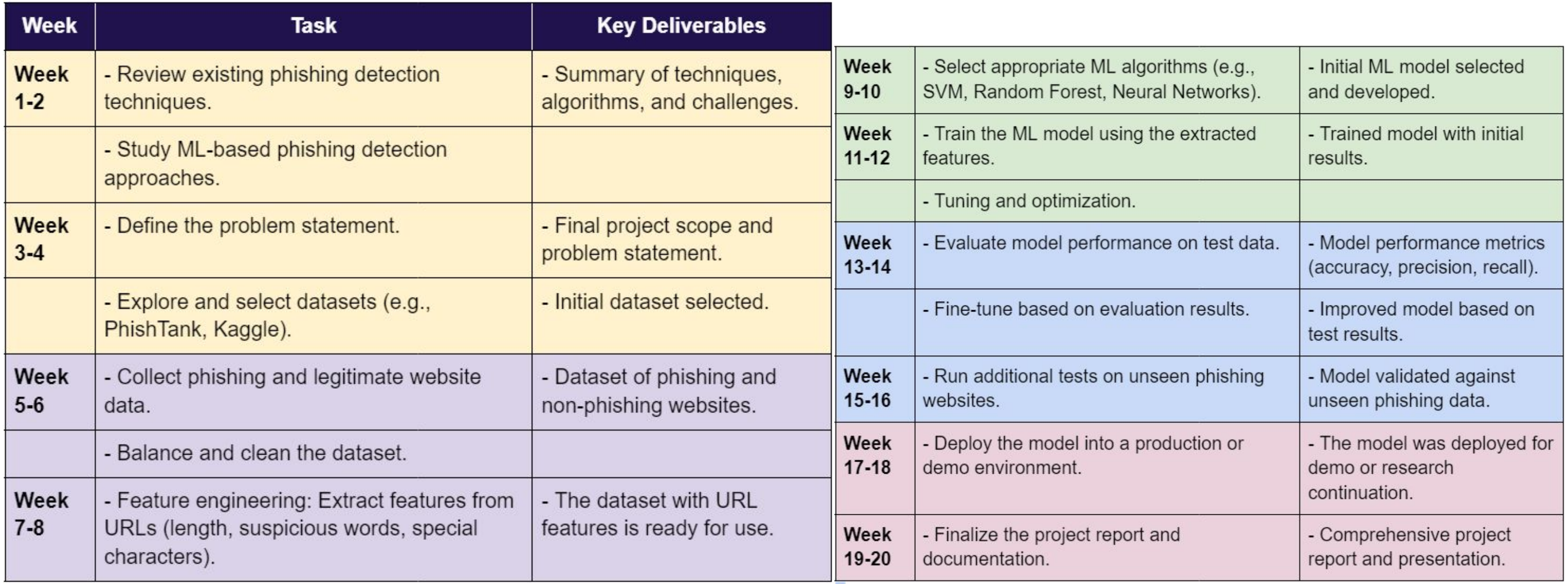


Security Warning

This website is detected as a phishing site. Please avoid providing any personal information.

- ⚠ Do not enter login credentials
- ⚠ Avoid sharing sensitive data
- ⚠ Report this site to your browser
- ⚠ Close this page immediately

Work Plan



Week	Task	Key Deliverables
Week 1-2	- Review existing phishing detection techniques.	- Summary of techniques, algorithms, and challenges.
	- Study ML-based phishing detection approaches.	
Week 3-4	- Define the problem statement.	- Final project scope and problem statement.
	- Explore and select datasets (e.g., PhishTank, Kaggle).	- Initial dataset selected.
Week 5-6	- Collect phishing and legitimate website data.	- Dataset of phishing and non-phishing websites.
	- Balance and clean the dataset.	
Week 7-8	- Feature engineering: Extract features from URLs (length, suspicious words, special characters).	- The dataset with URL features is ready for use.
Week 9-10	- Select appropriate ML algorithms (e.g., SVM, Random Forest, Neural Networks).	- Initial ML model selected and developed.
Week 11-12	- Train the ML model using the extracted features.	- Trained model with initial results.
	- Tuning and optimization.	
Week 13-14	- Evaluate model performance on test data.	- Model performance metrics (accuracy, precision, recall).
	- Fine-tune based on evaluation results.	- Improved model based on test results.
Week 15-16	- Run additional tests on unseen phishing websites.	- Model validated against unseen phishing data.
Week 17-18	- Deploy the model into a production or demo environment.	- The model was deployed for demo or research continuation.
Week 19-20	- Finalize the project report and documentation.	- Comprehensive project report and presentation.

Thank You