Detection of Phishing Websites using Machine Learning

BY

Arnab Kumar Das- 10200220008
Shubhankar Das- 10200220006
Sayan Banerjee- 10200220041
Sarthak Dey- 10200220015

Under The Guidance of : **Prof. Malabika Sengupta**

Problem Statement

Title: Detection of Phishing Websites using Machine Learning
Problem Statement:
The surge in phishing attacks poses a severe threat to online security, demanding a sophisticated solution to accurately detect and prevent malicious URLs.

Objectives:
1.**High Accuracy**: Develop a machine learning model with precise phishing URL classification.
2.**Real-time Detection**: Implement a system for instant analysis, preventing access to harmful websites.
3.**User-Friendly Interface**: Design an intuitive interface for seamless user interaction.
4.**Scalability**: Ensure efficient handling of a large volume of URL requests without compromising performance.

Outcome:
A proactive defense against phishing attacks, enhancing online security for end-users and professionals.

Overview

- Phishing websites pose a significant threat in today's digital landscape.
- This presentation explores the use of machine learning to detect and combat
these fraudulent sites.
- Analyzing factors such as English efficiency, source year, DNS filter, reviews, and more.

Factors to Consider for Analysis

Language Correctness

Measure the English proficiency level of the website content, as phishing sites often contain grammatical errors and awkward phrasing.

Source Year

Evaluate the age of the website, as recently registered domains are more likely to be associated with phishing attempts.

DNS Filter

Analyze the DNS filters used by the website to identify any suspicious or blacklisted domains.

Reviews

Consider the reputation and feedback from users and security experts to determine the legitimacy of the website.

Potential Outcomes of the Project

1 **Enhanced Cybersecurity**

By effectively detecting and blocking phishing websites, users can reduce the risk of falling victim to identity theft, financial fraud, and other cybercrimes.

2 **Time and Cost Savings**

Automated detection of phishing websites saves time and resources by minimizing manual efforts in identifying and reporting fraudulent sites.

3 **User Education**

Through awareness campaigns and educational initiatives, users can become more knowledgeable about the risks associated with phishing and take proactive measures to protect themselves.

Justification for Selecting the Title

The main purpose of building a phishing website detector is to enhance cybersecurity by identifying and preventing phishing attacks. Here are some key reasons for developing a phishing website detector:

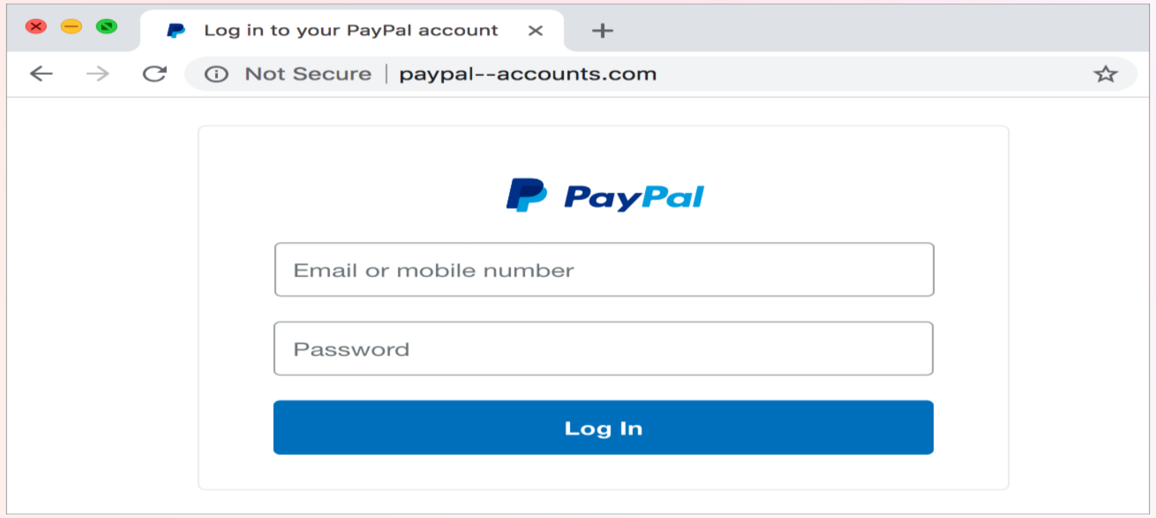
1. **Protect Users**
2. **Prevent Data Breaches**
3. **Maintain Trust**
4. **Financial Protection**
5. **Corporate Security**
6. **Compliance**
7. **Proactive Security Measures**
8. **Educational Purposes**
9. **Global Security**

Literary Survey

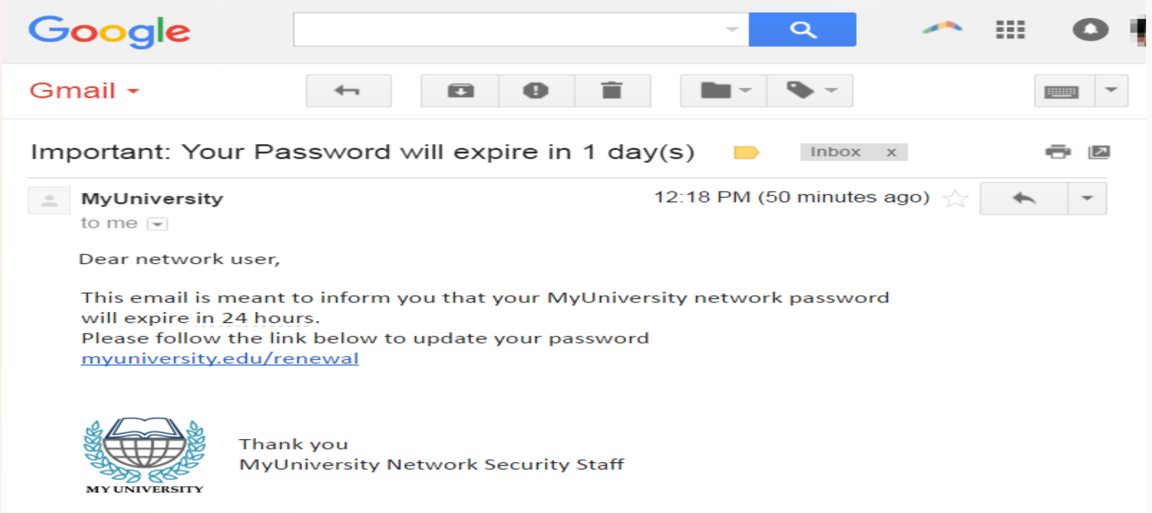
SL No	Author and Year	AIm	Main findings	Limitations
1	**Title:** Detecting phishing websites using machine learning technique **Author:** Ashit Kumar Dutta, 2021	The proposed framework employs RNN- LSTM to identify the properties Pm and Pl in an order to declare an URL as malicious or legitimate	The outcome of this study reveals that the proposed method presents superior results rather than the existing deep learning methods	The future direction of this study is to develop an unsupervised deep learning method to generate insight from a URL
2	**Title:** A systematic literature review on phishing website detection techniques **Authors:** Qabajeh et al., 2018	This review paper compares traditional anti-phishing methods, which includes raising awareness, educating users, conducting periodic training or workshop, and using a legal perspective. The Computerized anti-phishing techniques talk about list-based and machine-learning techniques	Machine Learning and rule induction are suitable to combat phishing due to their high detection rate and, more importantly, the easy-to-understand outcomes.	Sixty-seven studies were analyzed in work, and the research did not discuss Deep Learning techniques.
3	**Title:** Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review **Authors:** Eduardo Benavides,2020	This systematic literature review aimed to evaluate various other scholars' proposals for identifying phishing attacks using Deep Learning algorithms	In conclusion, there is still a significant gap in the area of Deep Learning algorithms for phishing attack detection..	This work includes 19 studies, and only research articles on phishing and Deep Learning are considered in this study.
4	**Title:** Applications of deep learning for phishing detection: a systematic literature review **Author:** Catal et al., 2022	The work answers nine research questions. The main aim is to synthesize, assess, and analyses Deep Learning techniques for phishing detection.	According to this study, 42 studies applied Supervised ML algorithms out of 43 studies. The most used algorithm was DNN, and the best performance was given by DNN and Hybrid DL algorithms.	The work only discusses Deep Learning related studies for phishing detection.

Modes of Phishing

1. Examples of a Phishing Website



2. Phishing link through email



Important: Your Password will expire in 1 day(s)

MyUniversity
to me

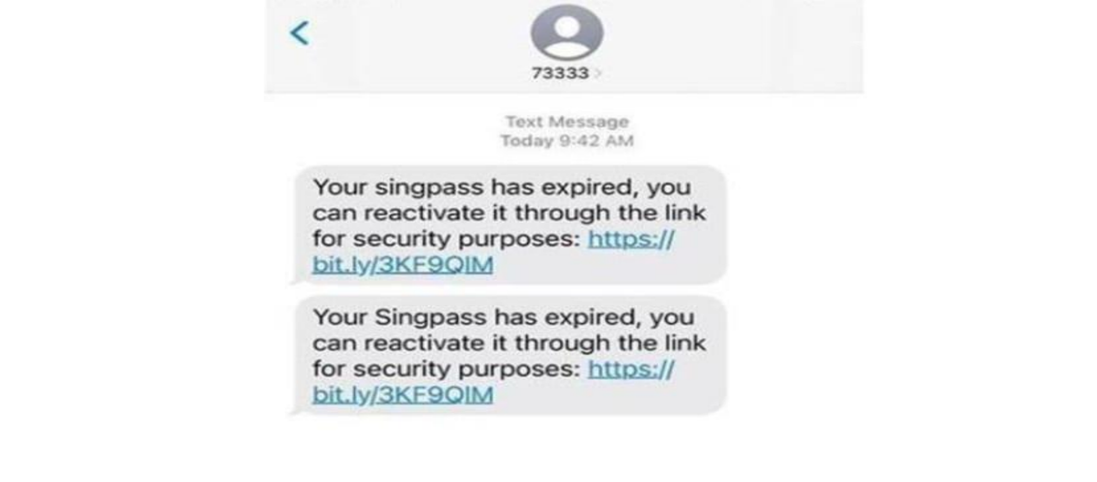
12:18 PM (50 minutes ago)

Dear network user,

This email is meant to inform you that your MyUniversity network password will expire in 24 hours.
Please follow the link below to update your password
myuniversity.edu/renewal

Thank you
MyUniversity Network Security Staff

3. Phishing link through SMS



73333

Text Message
Today 9:42 AM

Your singpass has expired, you can reactivate it through the link for security purposes: https://bit.ly/3KF9QIM

Your Singpass has expired, you can reactivate it through the link for security purposes: https://bit.ly/3KF9QIM

4. Phishing via Call

Implementation

Below are the steps of how we plan on approaching the problem statement:

1.Define Objectives: We are building a website where we will collect links of suspected websites. Then we will check the sites using the machine learning algorithms. If the site is a phishing website, we'll add it in our database and then submit the report to organizations like apwg.

2.Data Collection: We collected our training and test data from the UCI phishing dataset that is publicly available

3.Feature Extraction: Identify relevant features from the URLs that can help distinguish between phishing and legitimate websites. Features might include URL length, presence of HTTPS, domain age, and other relevant characteristics.

4.Data Preprocessing: Clean and preprocess the dataset. This involves handling missing values, encoding categorical variables, and scaling numerical features.

5.Split Data: Divide the dataset into training and testing sets. This allows you to train the model on one subset and evaluate its performance on unseen data.

6. **Model Selection:** We chose to start with the Random Forest classifier to work with. After reading various works on this field, many have approached the problems with this algorithms. We plan on starting off with random forest and testing other algorithms too on the way to determine which works best for our requirement.

7. **Feature Selection:** The difficulty arises when we must determine what are the most relevant features from a set and what combination of features give us near perfect classification accuracies. From the 30 features, we identified a few subsets.

8. **Training:** Train the model using the training dataset. The model will learn to distinguish between phishing and legitimate websites based on the provided features.

9. **Testing:** Evaluate the model's performance on the testing dataset. Use metrics such as accuracy, precision, recall, and F-score to assess its effectiveness.

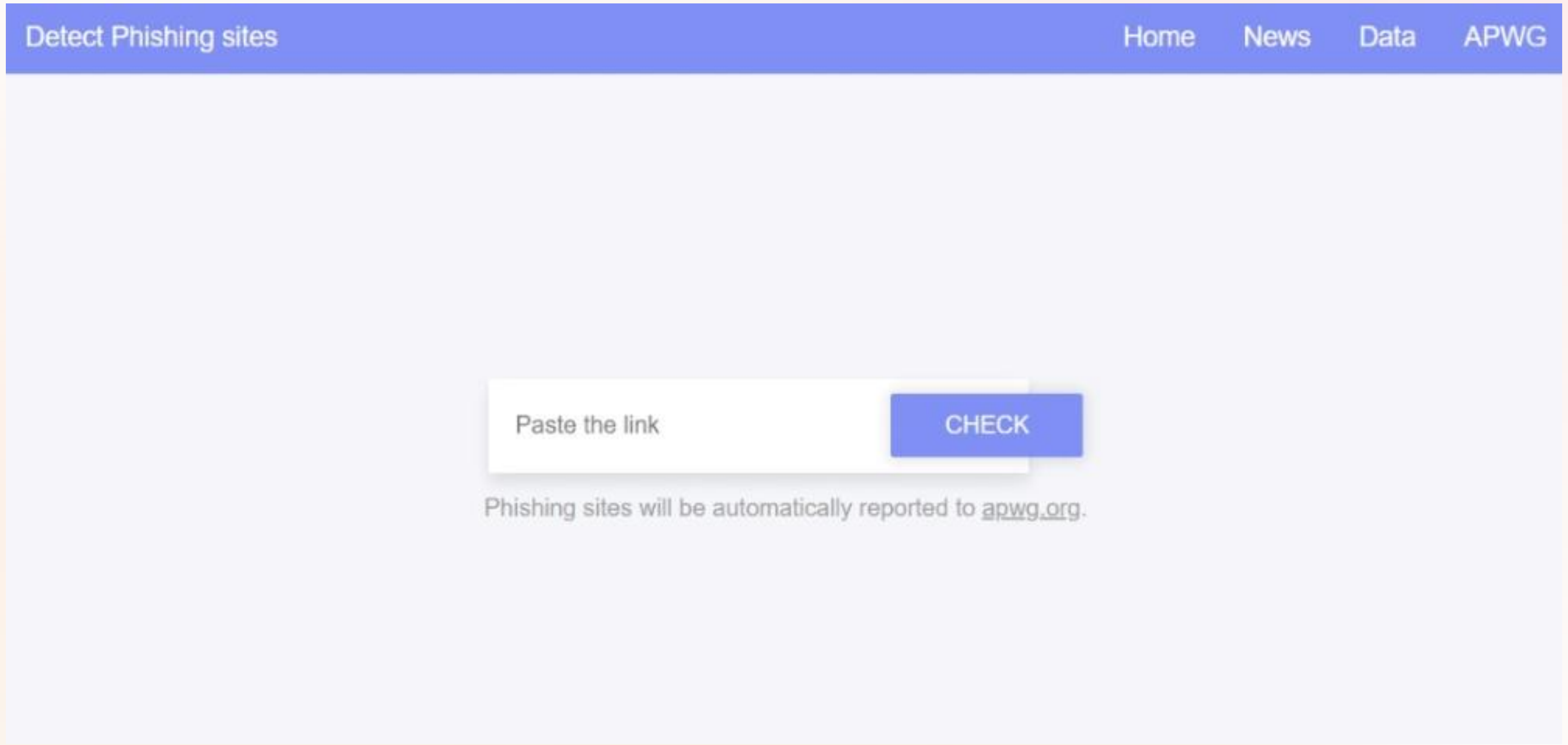
10. **Hyperparameter Tuning:** Optimize the performance of the model by tuning its hyperparameters.

11. **Validation:** Perform additional validation, such as cross-validation, to ensure the model's generalizability and robustness.

12. **Deployment:** Once satisfied with the model's performance, deploy it for real-time detection. This could involve integrating it into a web application, browser extension, or network security system.

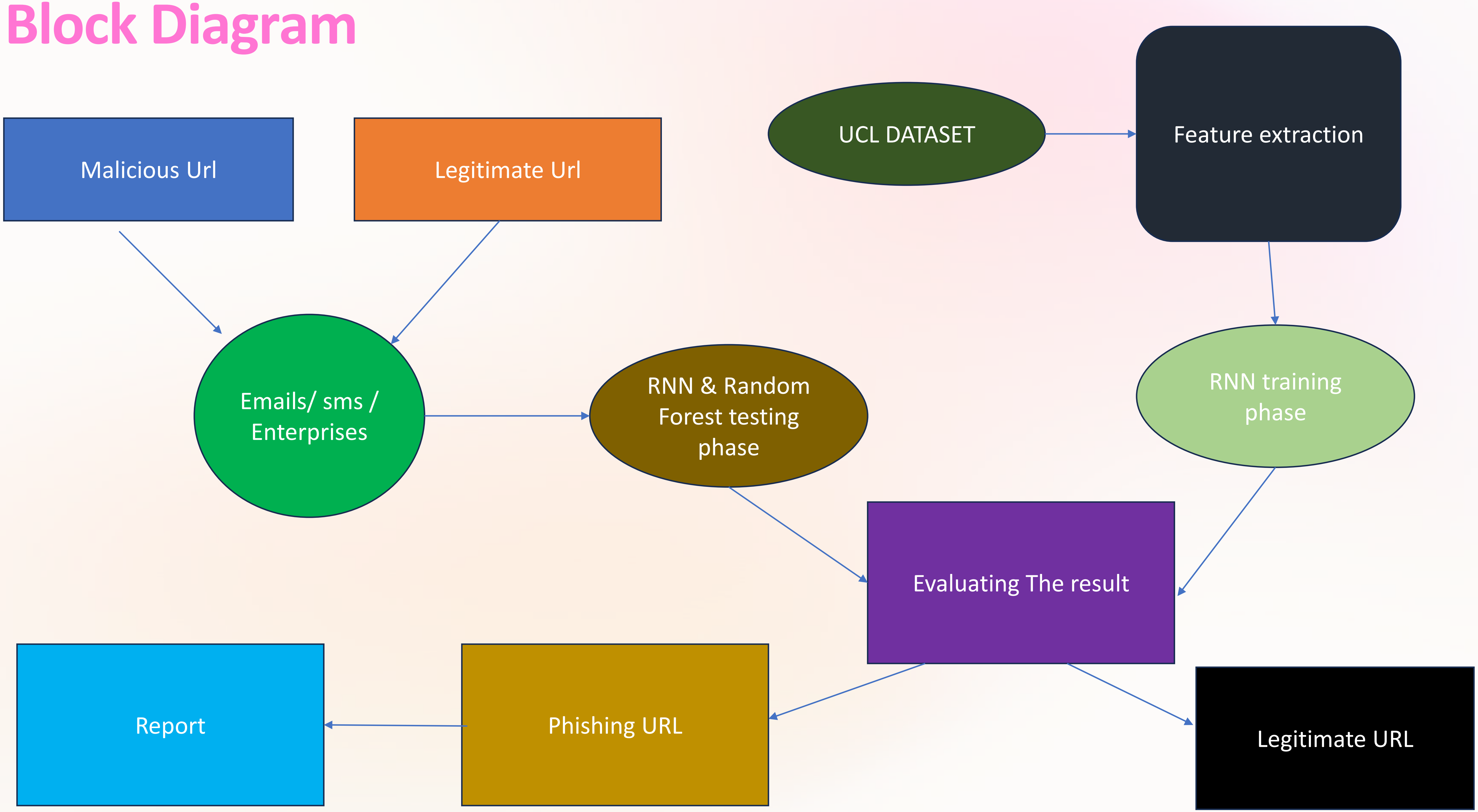
Implementation

Here is a screenshot of the front end of the website we are working on. The simple idea is that, anyone can come and paste an URL in the “Paste the link” box and click “CHECK” to figure out whether a website is authentic or fake.



Phishing sites will be automatically reported to apwg.org.

Block Diagram



Malicious Url

Legitimate Url

UCL DATASET

Feature extraction

Emails/ sms / Enterprises

RNN & Random Forest testing phase

RNN training phase

Evaluating The result

Report

Phishing URL

Legitimate URL

Tech Stacks for Detecting Phishing Websites

- Python
- Jupyter Notebook
- Numpy
- Pandas
- Scikit Learn
- HTML
- CSS
- JavaScript
- JQuery
- Node.js
- Express.js
- PostgreSQl

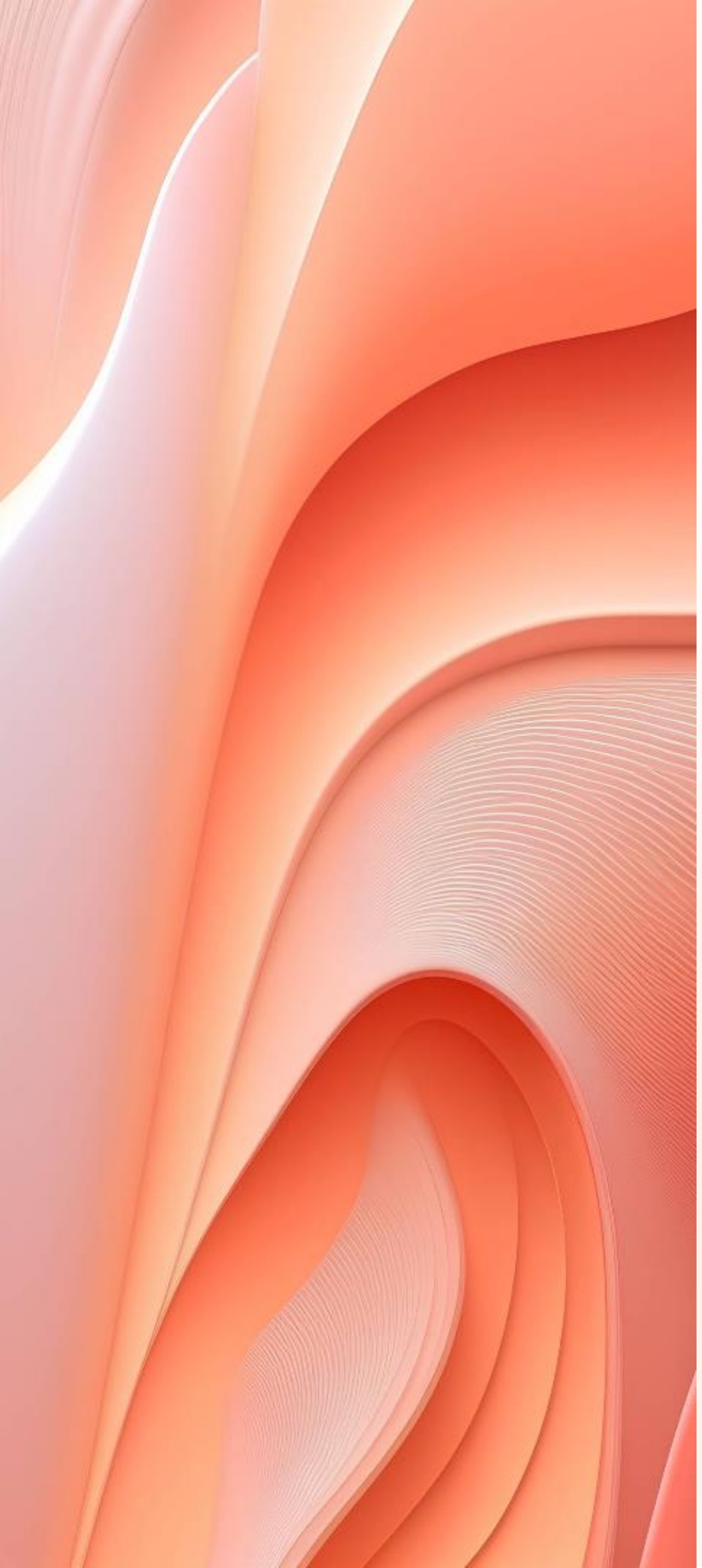
Future Work

Algorithm Implementation: Finalize the development of machine learning algorithms for phishing detection using Python.

API Integration: Seamlessly link the algorithms to the website through APIs, ensuring a user-friendly and responsive interface.

Database Creation: Establish a robust database structure to store information on identified phishing websites, enhancing data management and analysis capabilities.

Submission to APWG: Implement a systematic process to submit reports on identified phishing websites to organizations like APWG, contributing to the broader cybersecurity community.



Thank You!