Graphic Era deemed to be University DEHRADUN

PROJECT AND TEAM INFORMATION

Project Title

An efficient mechanism for machine learning based phishing attack detection.

Student/Team Information

Team Name: Team #	Tech shield CYBER-IV-T003
Team member 1 (Team Lead) Tomar, Ishita 230221280 ishitatomar1992@gmail.com	

Team member 2
Gupta, Anshu
23022478
anshugupta699@gmail.com

Team member 3
Singh, Dashpreet
230213889
dashpreetsingh632@gmail.com

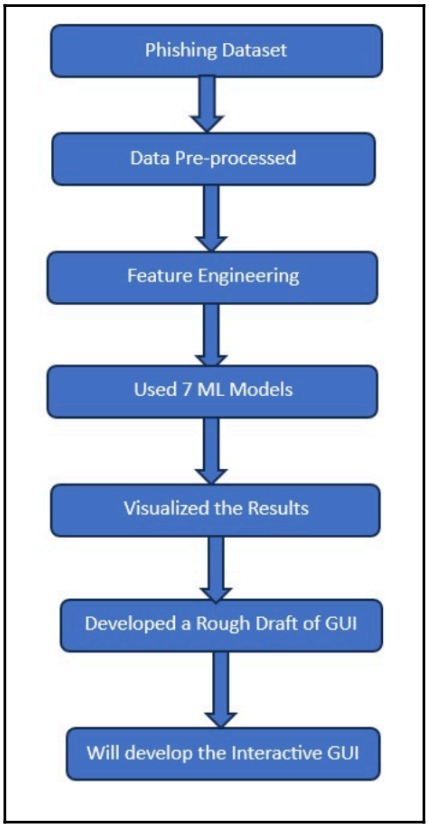
Team member 4
Singh, Ishita
230221084
ishitasingh0804@gmail.com

PROJECT PROGRESS DESCRIPTION (35 pts)

Project Abstract

1. **Develop an ML-based system for phishing detection**: To create a machine learning system capable of identifying phishing websites and emails.
2. **Improve detection accuracy:** To achieve a higher level of accuracy in identifying phishing attacks compared to traditional rule-based methods.
3. **Reduce false positives**: To minimise the number of legitimate websites and emails incorrectly flagged as phishing.
4. **Enable real-time phishing detection**: To implement a system that can identify phishing attempts as they occur, preventing associated cyber fraud.
5. **Enhance cybersecurity**: To contribute to a stronger overall cyber security posture by providing more effective protection against phishing threats.

Updated Project Approach and Architecture



Tasks Completed

Task Completed	Team Member
• Data Collection, preprocessing and feature selection. • Model selection and training • Visualisation of result • Rough draft of GUI	• Ishita Tomar • Dashpreet Singh • Ishita Singh • Anshu Gupta

Challenges/Roadblocks

1. Computational Challenges with CNN and BiLSTM Models:
We faced significant delays while running deep learning models such as CNN and BiLSTM due to their high computational demands. To address this, we reduced the dataset size to an optimal level that allowed for efficient model training without compromising performance substantially.
2. Difficulty in Implementing Ensemble Learning:
Integrating ensemble learning techniques to combine the outputs of previously applied models posed a challenge, as it was a new concept for us. To overcome this, we undertook a thorough study of ensemble methods to identify suitable algorithms and understand their implementation.
3. Issues Importing the 'TensorFlow' Library for GUI Integration:
While working on the GUI integration, we encountered difficulties with importing the TensorFlow library. To resolve this, we explored its structure, including various modules, layers, and utility functions. Additionally, we used TensorFlow functionalities to set the random seed and ensure reproducibility during model training.

Tasks Pending

Task Pending	Team Member (to complete the task)
A finalised GUI has to be created for the rough draft created.	Anshu Gupta

Project Outcome/Deliverables

The expected outcome is a highly accurate, AI-driven phishing detection system capable of identifying and preventing phishing attacks in real time. It will efficiently distinguish between phishing and legitimate websites or emails with high precision and recall, minimising false positives and false negatives for enhanced cybersecurity. Furthermore, the system is expected to provide comprehensive logging and reporting capabilities for security analysis and incident response. Its modular design should facilitate future updates and integration of new detection techniques. Ultimately, this project aims to deliver a robust and scalable solution that significantly reduces the risk and impact of phishing attacks.

Progress Overview

- Completed the data collection and cleaning.
- Completed the feature extraction using SelectKBest method.
- Implemented Logistic Regression, Naive Bayes, Random Forest XGBoost, SVM , CNN and BiLSTM.
- Combined the output of all these algorithms using ensemble learning.
- Visualised the results using confusion metrics and bar plot for result comparison.
- Created a rough draft for the GUI.

The final GUI application is scheduled to be developed after the completion of Phase 2 of the PBL evaluation.

Codebase Information

- **Gitlab repository account:** https://gitlab.com/ishitatomar1992/cyber_pbl.git
- **Branch: Everything is pushed in the main branch.**
- **Commits:**

1. data_cleaning.py by Ishita Tomar
2. feature_selection.py by Ishita Tomar
3. models.py by Dashpreet Singh
4. confusion_metrics.py by Ishita Singh
5. GUI folder which contains all the files required for the GUI development by Anshu Gupta

Testing and Validation Status

Test Type	Status (Pass/Fail)	Notes
We tested it with 2 different datasets	Pass	We achieved high accuracy, F1-score, recall and precision

Deliverables Progress

- The outcome is a highly accurate, AI-driven phishing detection system capable of identifying and preventing phishing attacks in real time. It will efficiently distinguish between phishing and legitimate websites or emails with high precision and recall, minimising false positives and false negatives for enhanced cybersecurity.
- We were able to present the performance metrics with high accuracy, f1-score, recall and precision.
- We also plotted the confusion metrics for all the models used here.
- We also compared the results the of the various algorithms.
- It was able to distinguish between legitimate and phishing URLs.