CS 700
Term Project Proposal
Fall 2023

Md Sultanul Islam Ovi
movi@gmu.edu
G01444040

PhishGuard: A Multi-Layered Ensemble Model for Optimal Phishing Website Detection

1 Introduction

Phishing attacks are a significant cybersecurity threat that demands advanced detection mechanisms. While techniques such as DL, ML, and heuristic algorithms hold value, a multi-layered approach promises enhanced performance and dependability. This project's goal is to design and evaluate an ensemble model for superior phishing website detection. I plan to utilize various datasets, including the Phishing Websites Data Set from the UCI Machine Learning Repository, a Kaggle dataset, and OpenPhish real-time feeds. These datasets provide a comprehensive foundation for the project, and their diverse features, from SSL status to HTML content, will be pivotal for an effective phishing detection system.

Problem Statement

The project seeks to address the query: "How can an ensemble model integrating a limited but diverse set of DL, ML, and heuristic methods enhance the precision and robustness of phishing website detection?"

2 Methodology

2.1 **Data Collection and Preprocessing:** Datasets from Kaggle and UCI will be curated and preprocessed to address missing or inconsistent entries.

2.2 **Feature Engineering:** Relevant features, such as URL length, domain name, and path length, will be extracted for model training.

2.3 **Individual Model Training:** A selective set of three to four DL, ML, and heuristic models will be employed for phishing detection. The emphasis will be on thorough analysis and performance evaluation.

2.4 **Ensemble Strategy:** Techniques like weighted voting and stacking will combine the individual models into a cohesive system.

2.5 **Performance Evaluation:** Multiple metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, will be considered separately to scrutinize the ensemble model's performance.

2.6 **Statistical Analysis:**

- Point Estimates, Confidence Intervals, and Hypothesis Testing will be employed.
- The effectiveness of the ensemble model will be contrasted with individual models using visual tests and the ANOVA technique and contrasts approach.

3 Conclusion

The PhishGuard project aims to revolutionize current methods by introducing a multi-layered ensemble model crafted for phishing detection excellence. By integrating a few well-chosen techniques, the research aspires to redefine standards for phishing website detection. Rigorous statistical evaluation will not only help generate relevant data but will also be pivotal for insightful analysis, leading to impactful conclusions.