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 continue to be a significant cybersecurity threat, necessitating advanced detection mechanisms. While individual methods like Deep Learning (DL), Machine Learning (ML), and heuristic algorithms offer value, a comprehensive, multi-layered approach has the potential for improved performance and reliability. The project aims to design and evaluate an ensemble model for optimal phishing website detection. To build a robust model, I plan to utilize several datasets such as the Phishing Websites Data Set from the UCI Machine Learning Repository, a user-friendly dataset hosted on Kaggle, and real-time feeds from OpenPhish. These datasets collectively offer a comprehensive foundation for the project, and their varied features—ranging from SSL status to HTML content—will help in creating a highly accurate phishing detection system.

Problem Statement

The project intends to answer the following question: "How can an ensemble model integrating DL, ML, and heuristic methods improve the accuracy and robustness of phishing website detection?"

2 Methodology

2.1 **Data Collection and Preprocessing:** The dataset will be sourced from Kaggle, UCI and subsequently preprocessed to rectify any missing or inconsistent data entries.

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

2.3 **Individual Model Training:** Distinct DL, ML, and heuristic models can be employed for phishing detection. ML algorithms such as RF, SVM, and NB may focus on identifying key features and patterns. DL methods like CNN or RNN can offer nuanced feature extraction and sequence recognition. Heuristic approaches may include rule-based classifiers for spotting common phishing elements like suspicious URLs.

2.4 **Ensemble Strategy:** Ensemble techniques like weighted voting and stacking will be applied to combine the individual models into a unified system.

2.5 **Performance Evaluation:** The performance of the ensemble model will be scrutinized using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

2.6 **Statistical Analysis:**

- To delve deeper into model evaluation, Point Estimates, Confidence Intervals, and Hypothesis Testing will be used as key statistical tools.
- For contrasting the effectiveness of the ensemble model with that of the individual models, Comparing Proportions and Paired/Unpaired Observations techniques will be applied.
- To understand how much each individual model contributes to the ensemble, Regression Analysis will be conducted as the final step in our statistical assessment.

3 Conclusion

The PhishGuard project goes beyond existing methods by innovating a multi-layered ensemble model tailored for optimal phishing detection. By synergistically combining different approaches, this research aims to set new standards in phishing website detection accuracy and reliability. Employing rigorous statistical analysis, the project seeks to offer novel insights and a significant contribution to cybersecurity research and applications.