Neil Kumar 10-31-25

# **Project Proposal**: *Reproducing Malware Traffic Classification with nPrintML*

| First Name | Last Name | cnet ID | Project Role |
|---|---|---|---|
| Neil | Kumar | neilkumar | Project Lead |

**Project Description:**
This project aims to reproduce results from the nPrint/pcapML leaderboard for the malware traffic classification task. The goal is to train and evaluate machine learning models such as Logistic Regression, Random Forest, and MLP, on the nPrint Malware Detection dataset to identify malicious network traffic from benign flows. This problem is important because malware detection is a core challenge in modern computer systems and networks, and applying machine learning to packet-level data can be an effective approach and provide a level of automation and accuracy that cannot be achieved from traditional packet analysis. This project is aligned directly with the main goal of the course, by combining machine learning methods with real network traces. I plan to reproduce the baseline results using models such as logistic regression, random forests, and MLP's, and then test whether basic preprocessing or feature scaling can improve performance. The focus of my project will be on clean and clear experimentation and clarity in my results rather than new contributions.

**Data:**
I will use the publicly available netML Malware Detection dataset, accessible through the nPrint benchmark page. This dataset includes labeled packet captures that have already been processed into numerical feature representations, eliminating the need for manual feature extraction/labelling or packet parsing.

**Deliverables:**
1. Clean Jupyter notebook that can be run end-to-end without errors. It will load the dataset, train multiple models, compare the performance to the current leaderboard results, and visualize the key metrics such as accuracy and confusion matrices.
2. ASphinx-formatted project report summarizing the methodology, results, and key takeaways from this project.