Network Feature Extraction from

Traffic Captures to Support Automation of

High-Fidelity Cyber Simulation Environment

Project Team: Christopher Castro, Ryan Collete, William Lagos, Sam Soltanian

*Computer Science*

Project Number CS 025-345

Faculty Advisor(s): Irfan Ahmed

Sponsor: DoD ASPIRE

Mentor: Kim Jae Sung

This project enhances OsirisML, a machine learning application developed by the U.S. Department of Defense that identifies operating systems through passive network monitoring. Our improvements focus on increasing accuracy, reducing computational requirements, and streamlining the implementation.

Training on the original dataset resulted in 72% accuracy, a significant increase from the initial 32%. Additionally, we reached 91% accuracy on a subset of the dataset that excluded DDoS traffic. By implementing CatBoost gradient boosting models rather than XGBoost and applying advanced feature engineering, we reduced the feature count from 784 to 200, representing a 74.4% reduction. Our model requires only 123GB RAM to train the entire dataset compared to the original XGBoost model's 424GB, a 71% reduction in memory requirements.

The enhanced system uses a hierarchical approach that labels the packets as either Linux, Windows, or Mac. These packets are then labeled with specific OS versions by respective sub-models, except for Mac since our training data did not contain any labeled Mac sub-versions. OsirisML identifies operating systems by analyzing TCP/IP headers and leveraging machine learning to detect patterns unique to specific OS versions. Key packet identifiers include Fragment ID/Offset, Time To Live (TTL), Initial Sequence Number (ISN), and other TCP/IP header characteristics. CPU requirements remain consistent with the original implementation.

**Keywords:** Machine learning, Network traffic analysis, Operating system identification, Feature extraction, Passive monitoring

