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.

We achieved significant accuracy improvements using the Friday dataset from CIC-IDS-2017, boosting performance from 30% on the combined dataset to 75-80% through hyperparameter optimization and feature reduction. By implementing CatBoost gradient boosting models rather than XGBoost, applying column compression to reduce features by 240, and feature compression to reduce total features by 500 without accuracy loss, we created a more robust decision-making model. This hierarchical approach utilizes a larger model to sort packets into buckets which are then identified by smaller sub-models.

The enhanced system dramatically reduced training time from 6 hours to between 15 minutes and 1 hour depending on dataset size. 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.

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

