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 by increasing the original model's 32% accuracy on the entire dataset to 72%, with our model reaching 91% accuracy on cleaner datasets where attack traffic doesn't generate misleading data. By implementing CatBoost gradient boosting models rather than XGBoost and applying advanced feature engineering, we reduced the feature count from 84 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 maintains a hierarchical approach that utilizes a larger model to sort packets into buckets which are then identified by smaller sub-models. 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

