Project Proposal: Extending NetML Feature Set for Network Flow Analysis

Team Members: Luis Ojeda (luisojeda) (Collaborator), Kai Lee (ka1) (Collaborator)

Overview/Project summary: The NetML library is useful in processing packet captures but there are other features that could improve model accuracy when calculated and used through the library. Our project aims to extend the NetML library by adding additional flow-based features related to video segment transmission to improve model performance in predicting Quality of Experience (QoE). We intend on adding the following features:

Segment Count per Flow — *number of distinct segments transmitted.*

Segment Rate (segments per second) — *captures delivery pacing and adaptive streaming responsiveness.*

Average Segment Size and Variance — *identifies bitrate switching patterns.*

Segment Throughput — *mean bytes per segment interval, complementing existing packet throughput features.*

In addition to these features we hope to gain a better understanding of the library and perhaps make further enhancements. We've discussed potentially identifying additional useful features or implementing batch processing.

Data: To evaluate whether our contributions to the library improve model performance, we will rely on data that we will capture from a video stream using Wireshark. We were thinking about observing segments from either Youtube or Netflix.

Machine learning and Evaluation: We expect to try linear regression and a random forest classifier when we evaluate the features we will create. We will also evaluate using all of the models in the netML library. We will benchmark the performance of the models without using the added features and compare against the performance when the features are used. To evaluate each performance we will likely use accuracy, ROC curves, and other evaluation methods.

Learning objective: We are aiming to learn more about state-of-the-art ML and networks libraries, contribute to an open-source project, and solidify our understanding of the influence of different features on ML model performance and its network applications.