Reproducing NetML Baseline Results for Network Traffic Application Classification

Group: Connor McGraw, Faiz Hilaly, Hewitt Watkins

Project Summary

This project aims to reproduce the baseline classification results from "NetML: A Challenge for Network Traffic Analytics" (Barut et al., 2020) for the network application identification task. The paper benchmarked Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) models on network flow metadata to classify traffic into application types. Reproducibility in ML research is often hindered by unspecified implementation details, random seeds, and preprocessing differences. By attempting to reproduce these baselines, we aim to (1) validate the published findings, (2) identify missing or ambiguous information that affects replication, and (3) provide a well-documented foundation for future research on network traffic analytics.

The original problem—classifying applications in encrypted, dynamic-port network flows—is critical for network management, security, and QoS, as traditional port-based methods no longer suffice.

Data

We will use the NetML VPN-nonVPN Application Identification Dataset, focusing on the non-vpn2016 subset (derived from the ISCX VPN-nonVPN 2016 dataset). It includes 163,711 flows (131k train, 16k + 16k test), organized at three label granularities:

- 7 top-level classes (e.g., P2P, audio, chat)
- 18 mid-level classes (e.g., Facebook, Skype, Hangouts)
- 31 fine-grained classes (e.g., facebook_audio, skype_chat)

We will follow the paper's 80/10/10 split and verify that the recreated dataset's 31 metadata features match the originals, documenting any discrepancies from the proprietary Intel extractor used in the paper.

We chose the VPN-nonVPN dataset because none of us have prior experience working with VPN traffic, and we're interested in understanding how encryption and tunneling affect network flow behavior. This area feels especially relevant today, as it highlights the trade-off between maintaining user privacy and preserving network visibility and performance.

Machine Learning Models

We will reproduce the three baseline classifiers exactly as described:

- Random Forest: 100 estimators, max_depth = 10 (best baseline overall).
  SVM: RBF kernel, C = 1.0, trained on 10% of the data due to computational limits.
- MLP: one hidden layer (121 units), Adam optimizer, L2 = 0.0001.
- All models will use standardized metadata features (packet/byte counts, flow duration, header/payload statistics, TCP flag histograms).

Evaluation

We will directly compare our F1 and mean Average Precision (mAP) scores to the paper's reported baselines (e.g., RF F1 = 0.6273 @ 7 classes). Performance will be analyzed across all three label granularities, including per-class confusion matrices and imbalance checks. If discrepancies occur, we will perform sensitivity tests on preprocessing, random seeds, and library versions.

Learning Objectives

Through this reproduction study, we expect to:

1. Assess how reproducible the NetML baselines are with the provided information.
2. Identify which implementation details (e.g., data standardization, feature extraction) most affect results.
3. Understand why performance declines as class granularity increases (7 → 18 → 31).
4. Learn best practices for transparent, reproducible ML research in network traffic analytics.

Deliverables / Responsibilities

- Hewitt Watkins: Data acquisition + preprocessing, implements RF baseline, and documents reproducibility setup.
- Connor McGraw: Implements SVM baseline, analyzes runtime/memory trade-offs, and co-develops evaluation pipeline.
- Faiz Hilaly: Implements MLP baseline, leads sensitivity studies across models, and prepares analysis figures + results discussion.
- All Members: Collaborate to finalize the Colab notebook, ensure code readability, and co-write the final report/presentation.