Network Flow Classification and Attack Detection - ADA Assignment 3

Riya Shyam Huddar (MDS202431)

October 19, 2025

1. Introduction

This report presents an analysis of network flow data for multi-class attack detection. We explore dataset characteristics, preprocessing strategies, feature analysis, and model performance across several machine learning approaches including Logistic Regression, Random Forest, XGBoost, and LightGBM.

2. Dataset Analysis

2.1 Dataset Overview

The dataset consists of **network flow records** containing statistical features and labels indicating benign or malicious flows.

Column	Description
IPV4_SRC_ADDR	Source IPv4 address of the flow
L4_SRC_PORT	Source transport-layer port
IPV4_DST_ADDR	Destination IPv4 address of the flow
L4_DST_PORT	Destination transport-layer port
PROTOCOL	Transport protocol (TCP=6, UDP=17, etc.)
L7_PROTO	Application layer protocol identifier
IN_BYTES	Bytes sent from source to destination
OUT_BYTES	Bytes sent from destination to source
IN_PKTS	Packets sent from source to destination
OUT_PKTS	Packets sent from destination to source
TCP_FLAGS	TCP control flags
FLOW_DURATION(ms)	Duration of the flow
Label	0 = Benign, 1 = Malicious
Attack	Type of attack if malicious (Exploits, DoS, etc.)

Dataset size: 1,623,118 entries with numeric and categorical features.

2.2 Duplicate Flow Handling and Aggregation

- **Exact duplicates removed:** 19,740 rows.
- **Duplicate flows (same 5-tuple):** 23,210 rows.
- **Aggregation logic:** sum for bytes/packets, max for duration, median for packet sizes, max/mode for flags/protocols. Label marked as attack if any fragment is malicious. Attack type takes the most frequent among duplicates.

- **Result:** Dataset reduced to 1,580,168 rows representing unique flows.

2.3 Class Distribution

- Benign flows (Label 0): 1,519,637
- Malicious flows (Label 1): 60,531

Observation: Only ∼3.8% of flows are attacks, highlighting severe class imbalance.

2.4 Distribution of Malicious Attack Types

Among malicious flows:

- Exploits: 37.7%, Fuzzers: 27.9%, Reconnaissance: 17.3%
- Other types (DoS, Analysis, Shellcode, Backdoor, Worms) are sparse.

Observation: Few attack types dominate; rare attacks require special attention in modeling.

2.5 Feature Analysis

- **Packets:** DoS, Generic, Worms: incoming < outgoing. Backdoor: inbound-heavy.
- **Flow Duration:** Backdoor longest (∼7.6k ms), Shellcode/Reconnaissance shortest (∼370–570 ms)
- **Correlation:** OUT_BYTES/OUT_PKTS highly correlated (0.97), IN_BYTES/IN_PKTS moderately correlated (0.71), PROTOCOL/TCP_FLAGS negatively correlated (-0.76)
- **Attack Patterns:** Byte ratios and packet ratios differentiate attack behaviors (e.g., outbound-heavy DoS/Worms, inbound-heavy Backdoor/Analysis)

3. Modeling and Evaluation

A binary Random Forest classifier was first trained using only the Label column to distinguish between benign (0) and malicious (1) flows, achieving an accuracy of 0.99.

3.1 Multiclass Classification Results

We then trained different models to classify flows into **benign and multiple attack types**, making it a **multi-class classification** task.

Table 2: Performance Metrics for Multiclass Classification Models

Model	Accuracy	Macro F1-score	Weighted F1-score
Logistic Regression	0.70	0.30	0.81
Random Forest	0.99	0.83	0.99
Balanced Random Forest	0.99	0.82	0.99
XGBoost	0.99	0.77	0.99
LightGBM	0.97	0.68	0.98

Key Observations:

- Logistic Regression serves as a baseline and struggles with minority (attack) detection.

- Tree-based models (Random Forest, XGBoost, LightGBM) significantly outperform the linear baseline.
- Balanced sampling marginally improves minority detection while maintaining high accuracy.

3.2 Multi-Class Attack Classification (Malicious Subset Only)

To further analyze model behavior within malicious traffic, additional classifiers were trained **only on the attack flows** to distinguish between different attack types.

Table 3: Performance Metrics for Multi-Class Attack Classification (Malicious Flows Only)

Model	Accuracy	Macro F1-score	Weighted F1-score
Random Forest (Attack Flows Only)	0.92	0.87	0.92
Random Forest + SMOTE (Attack Flows Only)	0.91	0.83	0.91
Duplicates Present + SMOTE (Attack Flows Only)	0.76	0.61	0.80

Key Observations:

- Random Forest achieved high accuracy in distinguishing attack types, indicating clear feature separability.
- SMOTE balancing slightly reduced accuracy but improved recall for some minority classes.
- Presence of duplicates significantly degraded model performance.

4. Key Insights and Conclusions

- Dataset is highly imbalanced, with benign flows dominating the distribution.
- Feature analysis shows packet counts, flow duration, and byte ratios differ across attack types, which are highly informative for classification.
- Tree-based models provide strong performance for both binary and multi-class detection tasks.
- Removing duplicates and using appropriate balancing techniques improve generalization.
- Robust network attack detection depends on preprocessing (aggregation, deduplication, balancing) and model design choices.