Network Intrusion Detection Using XGBoost on CIC-IDS2017 Dataset

Sangay Thinley

School of Built Environment, Engineering and Computing
Leeds Beckett University

Introduction

- Focus: Develop an efficient NIDS using classical ML.
- Dataset: CIC-IDS2017 – realistic network traffic with benign and multiple attack types.
- Goal: Accurate multi-class intrusion detection with computational efficiency.

Dataset Overview: CIC-IDS2017

- 2.8M network flows, 78 features, 15 attack types.
- Classes include: DDoS, Botnet, PortScan, Heartbleed, Infiltration, Web attacks.
- Raw data required cleaning, encoding, scaling, and handling of missing values.

Data Preprocessing

- Missing values imputed with column mean.
- Categorical features: One-Hot Encoding (protocol) and Label Encoding (labels).
- Feature scaling: StandardScaler (mean=0, std=1).
- Class imbalance handled with SMOTETomek and XGBoost scale_pos_weight.

Model Selection: XGBoost

- Chosen for efficiency, scalability, and performance on tabular data.
- Gradient boosting ensemble of decision trees with L1/L2 regularization.
- Handles missing values and supports parallel processing.

Experimental Setup

- Train/Test split: 70/30 with stratified sampling.
- Metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC, False Positive Rate.
- Hyperparameter tuning: Randomized + Grid Search with 5-fold CV.

Hyperparameter Tuning

- Optimized parameters: deeper trees, lower learning rate, more estimators.
- Result: Improved performance vs baseline (default hyperparameters).
- Balanced high F1-Macro across all attack classes.

Optimized XGBoost Performance

- High True Positive Rate; low False Positive/Negative Rate.
- F1-Macro: 0.98; strong detection of rare attacks (e.g., Heartbleed, Infiltration).
- Feature importance: Flow Packets, Flow Duration, Total Length of Fwd Packets.

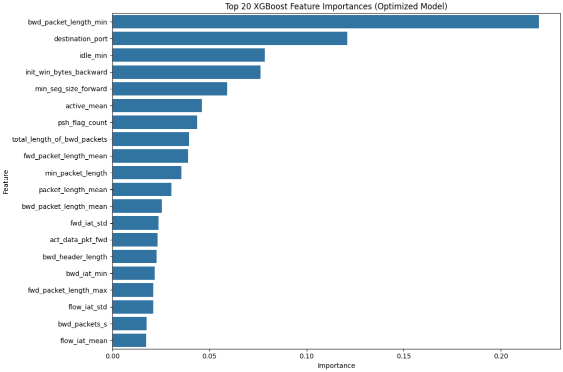

Figure: Top Features by XGBoost Importance

SHAP Feature Impact

- Visualizes contribution of each feature to model predictions.
- High values of Flow Packets/s push predictions towards DDoS class.
- Confirms model uses relevant network flow characteristics.

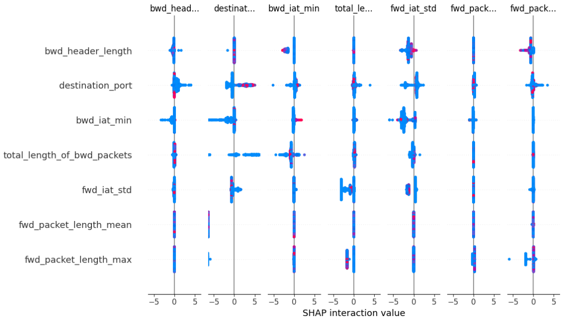

Figure: SHAP Summary Plot

Comparative Analysis

- XGBoost vs RF, LightGBM, CatBoost.
- Outperforms competitors in F1-Macro and Accuracy.
- Demonstrates efficiency-accuracy tradeoff for NIDS deployment.

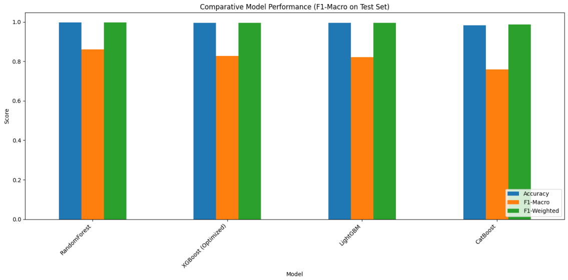

Figure: Comparative Performance Summary

Key Insights

- Classical ML is viable for computationally constrained NIDS.
- Class imbalance handling is critical for rare attack detection.
- Pipeline combines pre-processing, feature selection, tuning, and evaluation effectively.

Limitations and Future Work

- Dataset limitations: generalization to unseen attacks.
- Feature set: only CICFlowMeter features used.
- Future directions:
  - Test on other datasets (CSE-CIC-IDS2018, UNSW-NB15, IoT)
  - Hybrid models combining classical ML + deep learning
  - Real-time deployment and adversarial robustness
  - Explainable AI for operational decision support

Conclusion

- XGBoost effectively detects multi-class network intrusions on CIC-IDS2017.
- SMOTETomek class balancing ensures performance on rare attack types.
- Provides an end-to-end pipeline for NIDS: preprocessing, modeling, tuning, and evaluation.
- Classical ML offers a practical, computationally efficient solution compared to deep learning.

Network Intrusion Detection Using XGBoost on CIC-IDS2017 Dataset

Sangay Thinley

School of Built Environment, Engineering and Computing
Leeds Beckett University