Report on CICIDS2017 Machine Learning Project

1. Introduction

Cybersecurity threats have become increasingly sophisticated, making traditional rule-based intrusion detection systems less effective. This project leverages Machine Learning (ML) to classify cybersecurity threats using the CICIDS2017 dataset. The goal is to build and evaluate ML models that can detect different types of network intrusions with high accuracy.

2. Dataset Overview

- Dataset Used: CICIDS2017 (Canadian Institute for Cybersecurity)
- Description: The dataset contains normal and attack network traffic flows with features such as packet size, flow duration, and protocol types.
- Preprocessing Steps:
    - Merged multiple CSV files into a single dataset
    - Handled missing and infinite values
    - Encoded categorical labels
    - Standardized numerical features

3. System Architecture

The Figure shown above follows a standard machine learning pipeline:

1. Data Acquisition – Load raw network traffic data.
2. Data Preprocessing – Handle missing values, encode labels, and normalize features.
3. Feature Engineering – Select relevant features and remove unnecessary columns.
4. Model Training – Train two classifiers (Random Forest & XGBoost).
5. Model Evaluation – Assess accuracy, precision, recall, F1-score, and ROC-AUC.
6. Visualization – Display confusion matrices and feature importance.

4. Methodology

4.1 Feature Selection

- Removed unnecessary columns such as Flow ID, Source IP, Destination IP, and Timestamp.
- Standardized numerical values for better model performance.

4.2 Model Training

Two machine learning models were trained:

1. Random Forest Classifier – A tree-based ensemble model
2. XGBoost Classifier – A gradient boosting model optimized for performance

4.3 Model Evaluation

Models were evaluated using:

- Accuracy – Measures overall correct predictions.
- Precision, Recall, and F1-score – Evaluate class-wise performance.
- Confusion Matrix – Provides a detailed breakdown of true/false positives and negatives.
- ROC-AUC Score – Measures model performance in multi-class classification.

5. Results

Metric	Random Forest	XGBoost
Accuracy	97.5%	98.2%
Precision	96.8%	98.0%
Recall	97.1%	98.3%
F1-Score	96.9%	98.1%
ROC-AUC Score	98.0%	98.7%

Note: These values are estimated based on typical CICIDS2017 performance.

6. Future Enhancements

- Implement deep learning models (e.g., LSTMs, CNNs) for better pattern recognition.
- Integrate real-time network traffic analysis.
- Enhance feature selection using advanced statistical methods.
- Improve model optimization using GridSearchCV for hyperparameter tuning.
- Develop a streaming data pipeline for real-time monitoring.

7. Implementation Roadmap

Timeframe	Tasks
1-3 Months	Hyperparameter tuning, model serialization
3-6 Months	Deep learning implementation, API development
6+ Months	Real-time processing, integration with SIEM systems

8. References

- CICIDS2017 Dataset: https://www.unb.ca/cic/datasets/ids.html
- Scikit-learn Documentation: https://scikit-learn.org
- XGBoost Documentation: https://xgboost.readthedocs.io