Machine Learning-Based DDoS Detection in SDN Using the SNT Dataset

Ayush Singh (22BDS012)
Hitesh Sharma (22BDS028)
Sanchay Awana (22BCS109)
GitHub: github.com/AyushSingh916/ddos$_{d}etection$
Dataset: SNT Dataset

Abstract

Software-Defined Networking (SDN) makes modern networks flexible and programmable by separating the control and data planes. However, this same design leaves the control plane exposed, and Distributed Denial of Service (DDoS) attacks can easily overwhelm the controller, leading to service disruption.

In this project, we look at how machine learning can help detect such attacks before they cause damage. Using the SNT dataset, we tested several supervised learning models such as Random Forest, XGBoost, Logistic Regression, and Decision Trees. We evaluated them on both binary (attack vs. normal) and multiclass setups. Feature importance analysis helped us figure out which attributes contribute most to accurate detection.

Our results show that the pipeline achieves over 95% accuracy in binary classification, suggesting that machine learning approaches are highly promising for securing SDN environments.

1 Introduction

The rise of SDN has redefined network management by introducing centralization and programmability. While this makes the network easier to control, it also introduces a single point of failure — the controller. DDoS attacks exploit this by overwhelming the control plane with massive traffic flows, slowing the network down or even crashing it.

A possible solution is to use machine learning (ML). By analyzing flow-level statistics, ML models can detect unusual patterns and differentiate between benign and malicious behavior. For this project, we used the SNT dataset, which provides labeled traffic data covering both normal and attack scenarios. Our objective is to design and evaluate a pipeline that uses ML to detect DDoS attacks effectively.

2 Literature Survey

Some of the important works in this area include:

- Mousavi & St-Hilaire (2015): Suggested entropy-based methods for early DDoS detection in SDN. While effective at spotting initial threats, their approach struggled with adapting to varied attacks.
- Braga et al. (2010): Showed that simple flow-level statistics can be enough to detect flooding attacks in networks.
- Bawany et al. (2017): Surveyed existing approaches for SDN attack detection and mitigation, highlighting the lack of consistent datasets for benchmarking.
- Kyriakopoulos et al. (2019): Compared different machine learning techniques and confirmed the strength of supervised models for DDoS detection in SDN.

3 Problem Statement

DDoS attacks pose a major risk to SDN controllers because of their centralized role. The challenge lies in detecting these attacks in real-time despite the evolving nature of attack strategies and the sheer traffic volume.

The SNT dataset provides labeled flow statistics for both normal and malicious traffic. Our project goals are:

1. Train multiple ML models on the SNT dataset.
2. Identify which features are most important for detection.
3. Compare binary vs. multiclass performance.
4. Aim for at least 95% accuracy in binary classification.

4 Methodology

Our pipeline was developed in the following stages:

1. **Data Preprocessing:** Handling missing values, normalizing numeric attributes, and encoding categorical ones.
2. **Feature Selection:** Using Random Forest feature importance to highlight the most useful features.
3. **Model Training:** Training and testing four models — Random Forest, XGBoost, Decision Tree, and Logistic Regression — on both binary and multiclass tasks.
4. **Evaluation:** Using accuracy, F1-score, and ROC-AUC to measure performance.
5. **Final Model:** Selecting the best-performing model and retraining it on the entire dataset for deployment.

5 High-Level Design

```
+------------------+
|   SNT Dataset    |
+------------------+
         |
         v
+------------------+
| Data Preprocess  |
+------------------+
         |
         v
+------------------+
| Feature Selection|
+------------------+
         |
         v
+------------------+
|  Model Training  |
+------------------+
         |
         v
+------------------+
|   Evaluation     |
+------------------+
```

6 Detailed Design

```
[Input]
    -> Flow-level features from SNT dataset

[Preprocessing]
    -> Handle missing values
    -> Normalize numerical features
    -> Encode categorical features

[Feature Engineering]
    -> Random Forest importance ranking

[Model Training]
    -> Train classifiers:
        - Random Forest
        - XGBoost
        - Logistic Regression
        - Decision Tree

[Evaluation]
```

```
    -> Metrics:
        - Accuracy
        - F1 Score
        - ROC-AUC
```

7 Flow Chart

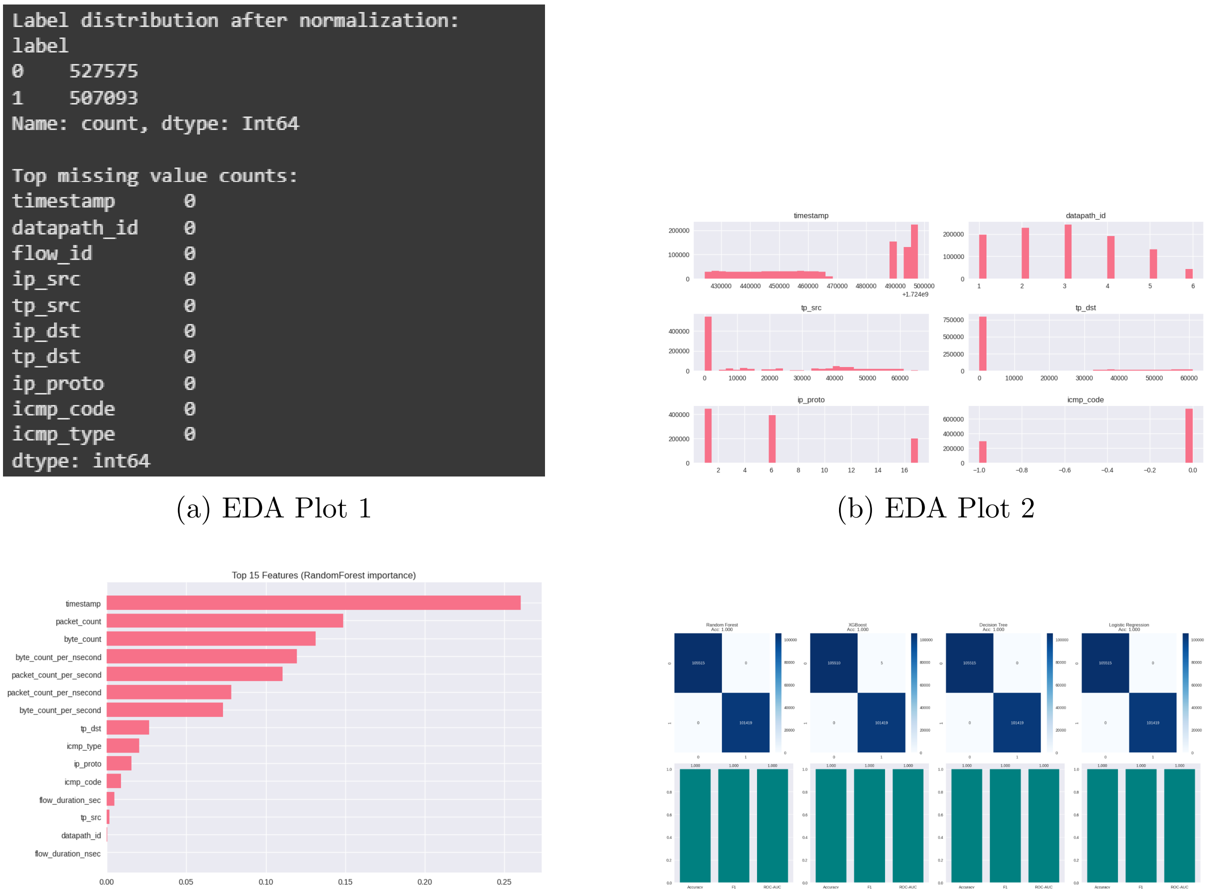
```
Start
   |
   v
Load Dataset
   |
   v
Data Preprocessing
   |
   v
Feature Selection
   |
   v
Train ML Models
   |
   v
Evaluate Performance
   |
   v
Deploy Best Model
   |
   v
End
```

8 Flow Diagram

```
                +----------------------+
                |     SDN Network      |
                +----------------------+
                          |
                          v
                +----------------------+
                |   Traffic Flows      |
                +----------------------+
                          |
                          v
                +----------------------+
                |   ML Detection       |
                | (Trained Models)     |
                +----------------------+
                          |
            +------------+------------+
            |                         |
            v                         v
  +--------------------+     +--------------------+
  | Normal Traffic     |     |   DDoS Attack      |
  +--------------------+     +--------------------+
```

9 Exploratory Data Analysis (EDA) Results

We carried out a simple EDA to get a sense of how the dataset behaves. The following plots summarize the distribution of traffic features and how they differ between normal and attack flows.



```
Label distribution after normalization:
label
0    527575
1    507093
Name: count, dtype: Int64

Top missing value counts:
timestamp      0
datapath_id    0
flow_id        0
ip_src         0
tp_src         0
ip_dst         0
tp_dst         0
ip_proto       0
icmp_code      0
icmp_type      0
dtype: int64
```

(a) EDA Plot 1

(b) EDA Plot 2

(c) EDA Plot 3

(d) EDA Plot 4

Figure 1: Exploratory Data Analysis (EDA) Results

10 Expected Outcomes

By the end of this project, we expect to achieve:

- A working set of ML classifiers that can detect DDoS attacks with high precision.
- A better understanding of which flow features are most useful for real-time detection.

- A clear comparison between binary and multiclass detection.
- Accuracy levels above 95% in binary classification.

11 Conclusion and Future Work

This project shows that machine learning can indeed help in detecting DDoS attacks in SDN environments. Using the SNT dataset, we built a pipeline that achieved strong results, especially in binary classification.

For future work, a few directions seem promising:

- Applying deep learning approaches for improved adaptability to unseen attacks.
- Testing the pipeline in a real SDN testbed to check performance in live conditions.
- Expanding detection beyond DDoS to cover other types of network intrusions.

References

1. H. Mousavi and M. St-Hilaire, ”Early detection of DDoS attacks against SDN controllers,” 2015 International Conference on Computing, Networking and Communications (ICNC), 2015.

2. R. Braga, E. Mota, and A. Passito, ”Lightweight DDoS flooding attack detection using NOX/OpenFlow,” IEEE LCN, 2010.

3. N. Bawany, J. Shamsi, and K. Salah, ”DDoS attack detection and mitigation using SDN: Methods, practices, and issues,” IEEE Communications Surveys & Tutorials, 2017.

4. K. Kyriakopoulos, G. Kambourakis, and S. Kolias, ”Towards machine learning-based DDoS detection for SDN,” 2019 IEEE Symposium on Computers and Communications (ISCC).