Project Report

DDoS Detection Using Machine Learning

Sophal Vaung

American University of Phnom Penh

COSC 221 001-Computer Science B

Dr. Abdallah Altrad

December 2, 2024

DDoS Detection Using Machine Learning

1. Introduction

Distributed Denial of Service (DDoS) attacks are a major concern in cybersecurity, because they flood networks with too much traffic, causing services to stop working. To deal with this concern, I built a project uses machine learning to detect and classify these attacks. To make it easier for users, I created a user-friendly interface using Streamlit UI that allows real-time analysis of network traffic data.

2. Objective

The primary objective of this project is to develop an efficient and scalable machine learning-based system for the detection and classification of DDoS attacks. By analyzing key features from network traffic, the system aims to show the difference between normal and harmful patterns with high accuracy. The project also focuses on building a simple interface that makes it easy to upload data, choose a model, and see the results. This includes preparing the data, selecting the best features, and testing the system to make sure it works well and can be used in real-life cybersecurity situations.

3. Technologies Used

3.1 Programming Language: Python

3.2 Libraries:

- Streamlit for the user interface

- Scikit-learn for machine learning models and metrics

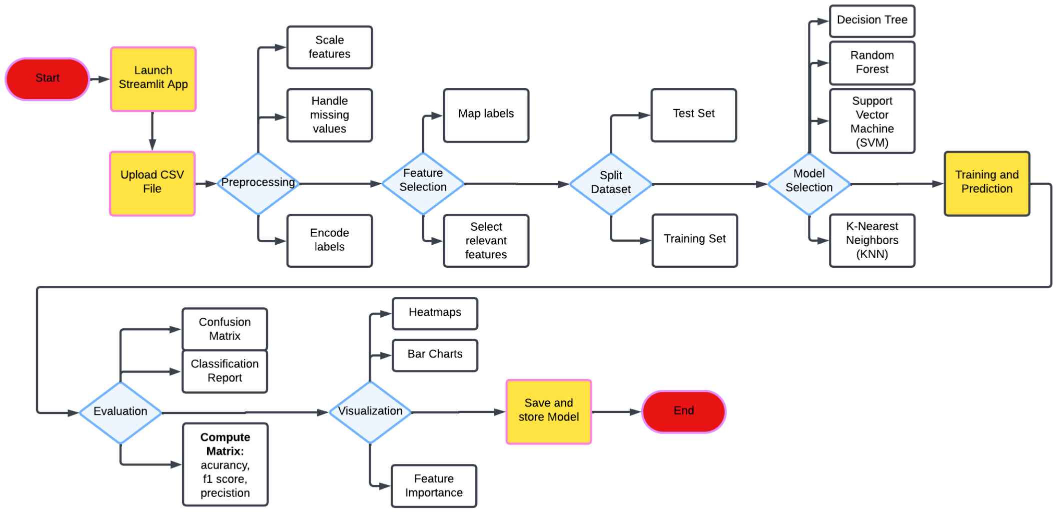
- Pandas and NumPy for data manipulation

- Matplotlib and Seaborn for visualization

- Joblib for model saving/loading

3. System workflow

3.1 Flowchart



4. Dataset and Features

4.1 Dataset

- Input: A CSV file containing network traffic data.
- Key Features: Selected from the dataset based on their relevance to detecting DDoS patterns, such as packet length, flow statistics, and flag counts.

```
# Selecting necessary features
necessary_features = [
    'Bwd Packet Length Mean', 'Avg Bwd Segment Size', 'Bwd Packet Length Max',
    'Bwd Packet Length Std', 'Packet Length Mean', 'Average Packet Size',
    'Packet Length Std', 'Max Packet Length', 'Packet Length Variance',
    'PSH Flag Count', 'Flow IAT Std', 'Flow IAT Mean', 'Fwd IAT Max', 'Flow IAT Max',
    'Fwd IAT Std', 'ACK Flag Count', 'Idle Max', 'Idle Mean', 'Idle Std', 'Idle Min',
    'Subflow Bwd Bytes', 'Total Length of Bwd Packets', 'Fwd IAT Total', 'Active Min',
    'Flow Duration', 'Active Mean', 'Fwd IAT Mean'
]
```

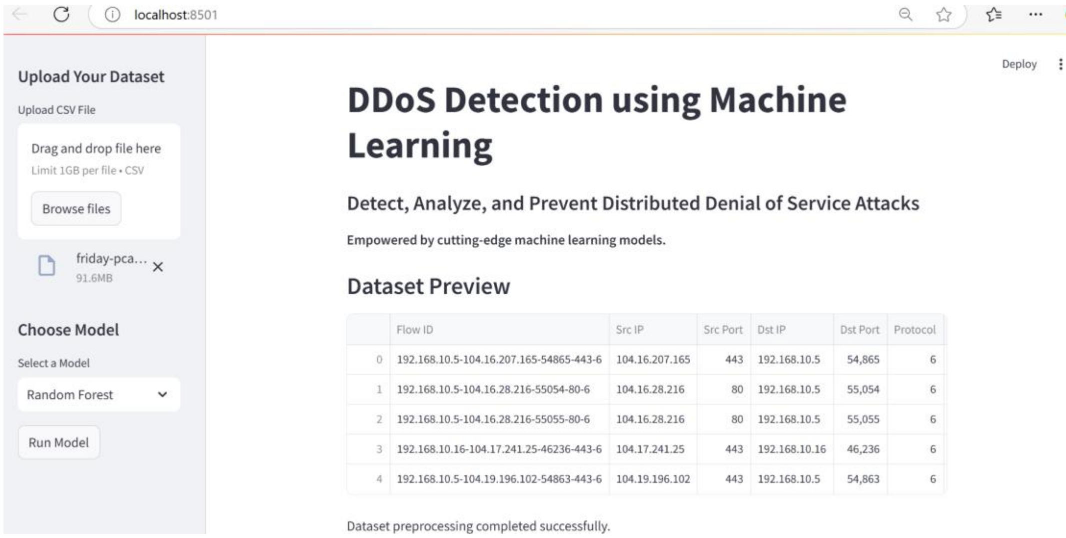
4.2 Preprocessing

- Removal of non-numeric columns.
- Filling missing values using the column mean.
- Label encoding for binary classification:
  - 0 for benign traffic.
  - 1 for DDoS attack traffic.

5. Methodology

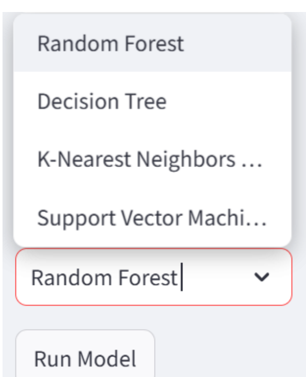
5.1 User Interface

- Built with: Streamlit.
- Functionality: Users can upload datasets, select models, and view results including predictions, performance metrics, and visualizations.



5.2 Machine Learning Models

Four classifiers were implemented and evaluated:
- Random Forest
- Decision Tree
- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)



5.3 Steps Involved

1. Data Splitting: Dataset divided into training (70%) and testing (30%) subsets.
2. Feature Scaling: StandardScaler was used to normalize the features.
3. Model Training: Models were trained using the training dataset.

4. Evaluation Metrics:
    - Accuracy Score
    - Precision
    - F1 Score
    - Confusion Matrix
    - Classification Report

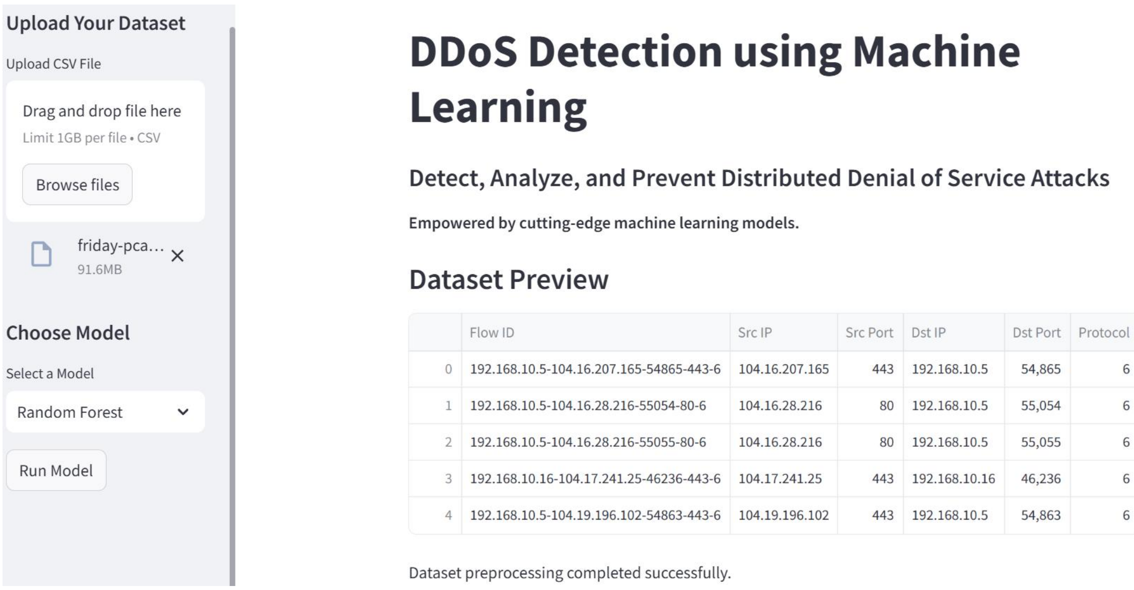
5.4 Label Mapping

Labels were mapped to interpret predictions:
- 0: BENIGN
- 1: DDoS

```
    # Define the label mapping explicitly
    label_mapping = {0: 'BENIGN', 1: 'DDoS'}
```

6. Application Results

6.1 Data Preview



	Flow ID	Src IP	Src Port	Dst IP	Dst Port	Protocol
0	192.168.10.5-104.16.207.165-54865-443-6	104.16.207.165	443	192.168.10.5	54,865	6
1	192.168.10.5-104.16.28.216-55054-80-6	104.16.28.216	80	192.168.10.5	55,054	6
2	192.168.10.5-104.16.28.216-55055-80-6	104.16.28.216	80	192.168.10.5	55,055	6
3	192.168.10.16-104.17.241.25-46236-443-6	104.17.241.25	443	192.168.10.16	46,236	6
4	192.168.10.5-104.19.196.102-54863-443-6	104.19.196.102	443	192.168.10.5	54,863	6

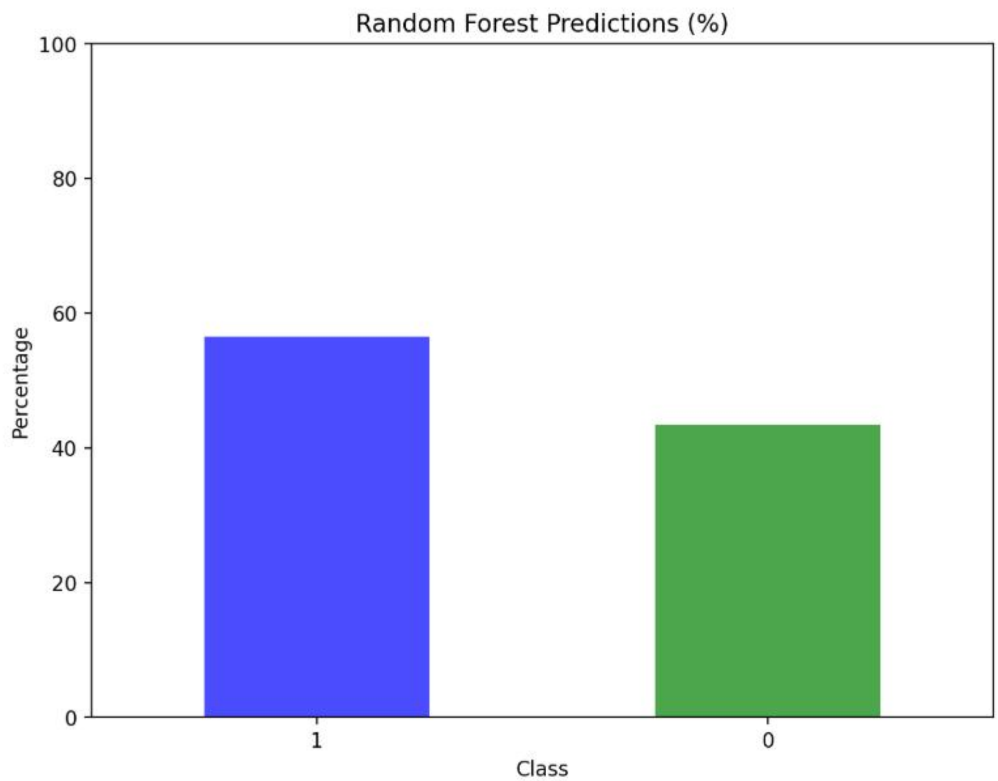
6.2 Prediction with String Label

	Predicted Labels (Numeric)	Predicted Labels (String)
0	0	BENIGN
1	0	BENIGN
2	1	DDoS
3	1	DDoS
4	1	DDoS
5	1	DDoS
6	1	DDoS
7	1	DDoS
8	0	BENIGN
9	0	BENIGN

6.3 Class Distribution

Random Forest Class Distribution (Percentage):

	Class	Percentage (%)
0	DDoS	56.5605
1	BENIGN	43.4395



6.5 Classification Report and Evaluation Metrics

Random Forest Classification Report:

	precision	recall	f1-score	support
BENIGN	0.9992	0.9996	0.9994	29,407
DDoS	0.9997	0.9993	0.9995	38,317
accuracy	0.9994	0.9994	0.9994	0.9994
macro avg	0.9994	0.9995	0.9994	67,724
weighted avg	0.9994	0.9994	0.9994	67,724

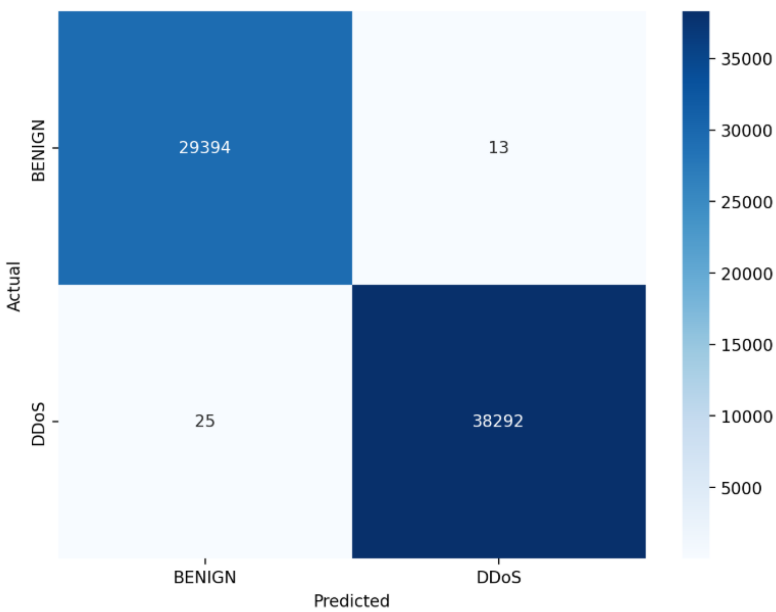
Accuracy: `0.9994388990608942`

F1 Score: `0.9994389123529402`

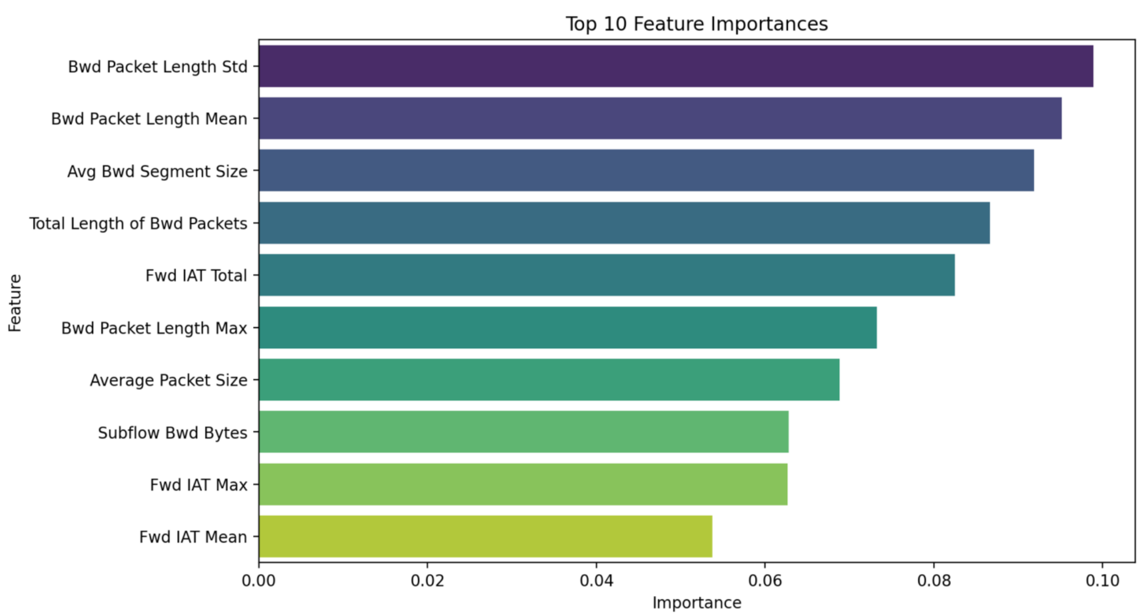
Precision: `0.9994388990608942`

6.7 Confusion Matrix

Random Forest Confusion Matrix:



6.8 Feature Importance



7. Key Features

- User-Friendly Interface: Simplified dataset upload and visualization.
- Visualization: Plots for class distribution, confusion matrices, and feature importance.
- Real-Time Analysis: Supports iterative model selection and evaluation.
- Model Persistence: Trained models were saved using joblib for reuse.

8. Conclusion

This project demonstrates the effective application of machine learning models for detecting DDoS attacks with high accuracy and efficiency. The interactive Streamlit interface enhances usability by enabling users to seamlessly explore datasets, train models, and interpret results, making the system both practical and accessible for real-world cybersecurity challenges.

9. Future Work

- Incorporate additional attack types for multi-class classification.
- Extend feature selection to include temporal characteristics of network traffic.
- Deploy the application on a web server for broader accessibility.

10. References

These references refer to the project idea, concept code, and datasets explored.

Kaggle: Your machine learning and data science community. (n.d.). Retrieved from https://www.kaggle.com/

Najafimehr, M., Zarifzadeh, S., & Mostafavi, S. (2023). DDoS attacks and machine-learning-based detection methods: A survey and taxonomy. ***Engineering Reports, 5****(12). https://doi.org/10.1002/eng2.12697*

Samruddhid. (n.d.). GitHub - DDoS Detection using Machine Learning. Retrieved from https://github.com/samruddhid5/DDoS-Detection-using-Machine-Learning