Task 1 – AI Integration of BlockShield Project

EDA Report: CIC-IDS-2017 Dataset

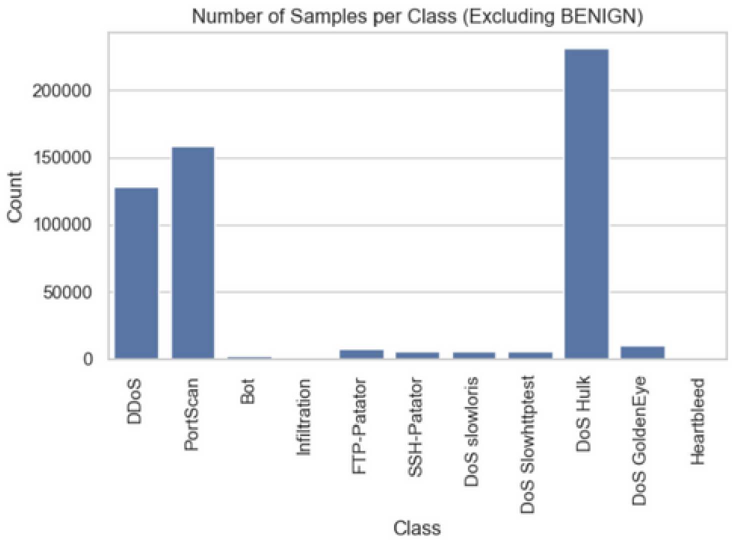
Nashat Alfarajat – BlockShield AI Development Team (intern)

Why This Dataset Was Chosen

The CIC-IDS-2017 dataset, created by the Canadian Institute for Cybersecurity, was selected for its high quality, labeled data, and relevance to real-world intrusion detection. It includes 2,660,377 records and 85 features, making it large and comprehensive enough to train robust machine learning and deep learning models.

The dataset captures a wide range of network attacks, such as:

DDoS, PortScan, Bot, Infiltration, FTP-Patator, SSH-Patator, DoS Slowloris, DoS Slowhttptest, DoS Hulk, DoS GoldenEye, Heartbleed, along with benign traffic.

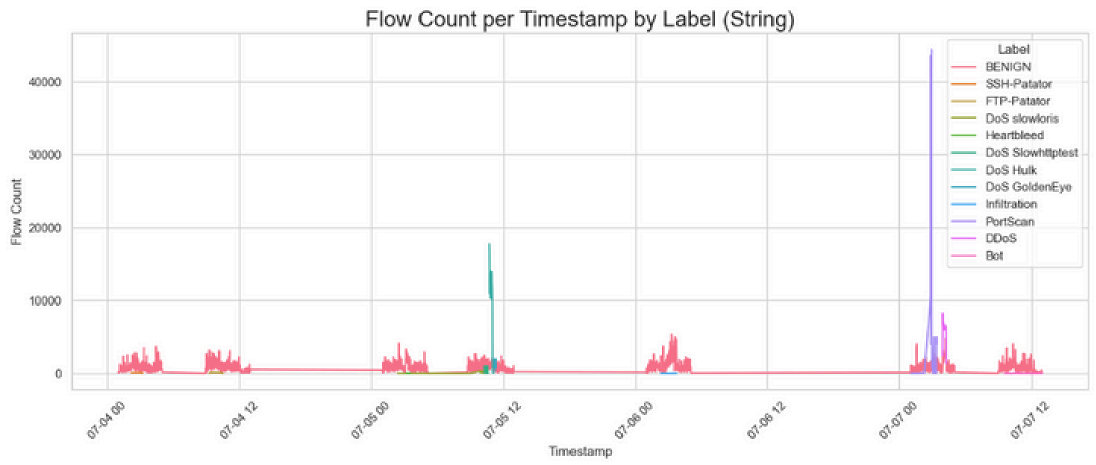


What the Dataset Contains

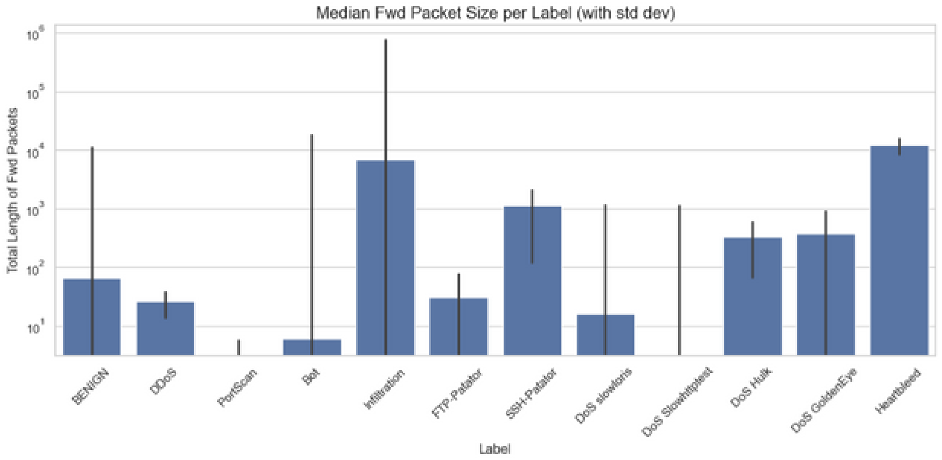
Each row represents a network flow and includes:

- **Timestamp-based Data**: Flows are captured over time, enabling time-series analysis.
- **Source/Destination IPs and Ports**: Making it possible to study attacker and victim behavior.
- **Traffic Metrics**: Packet counts, flow duration, byte rates, and inter-arrival times.
- **Protocol and Flag Data**: Useful for behavioral analysis.
- **Label**: Classifies the flow as either benign or a specific type of attack.

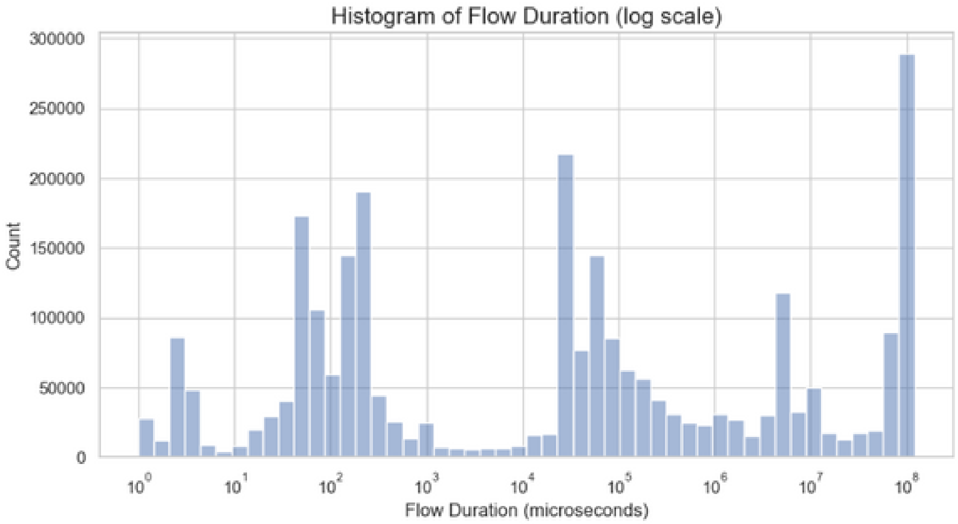
The dataset's size, diversity of attacks, and detailed structure make it ideal for supervised learning tasks, including both binary and multiclass classification.



Median Fwd packet size per Label:



Flow Duration count:



Current Progress

After exploring the data, initial preprocessing has been completed, including:

- Handling Missing Values
- Removing Duplicates
- Outlier Detection and Treatment

Next Steps

Further processing will be done during modeling, including:

- Scaling and Normalization
- Feature Selection and Engineering
- Model Training and Evaluation