SAPIENZA
UNIVERSITÀ DI ROMA

Lab Project - Networking for Big Data

Tran Luong Bang - 1956419
Gaurav Mohan Ramse - 1965564

Statistical Analysis

1) Extract 1 million of packets from the available data,

```
file_name = '../input/nbd-project/data.pcap'
new_file_name = './data_1m.pcap'


cmd('editcap -r ' + file_name +" "+ new_file_name+ ' '+ " 0-1000000")
```

1) Extract general info from trace using capinfos

- Number of packets in capture file

```
! capinfos -c './data_1m.pcap'
```

- The average data rate, in bit/sec

```
! capinfos -i './data_1m.pcap'
```

- The average packet size

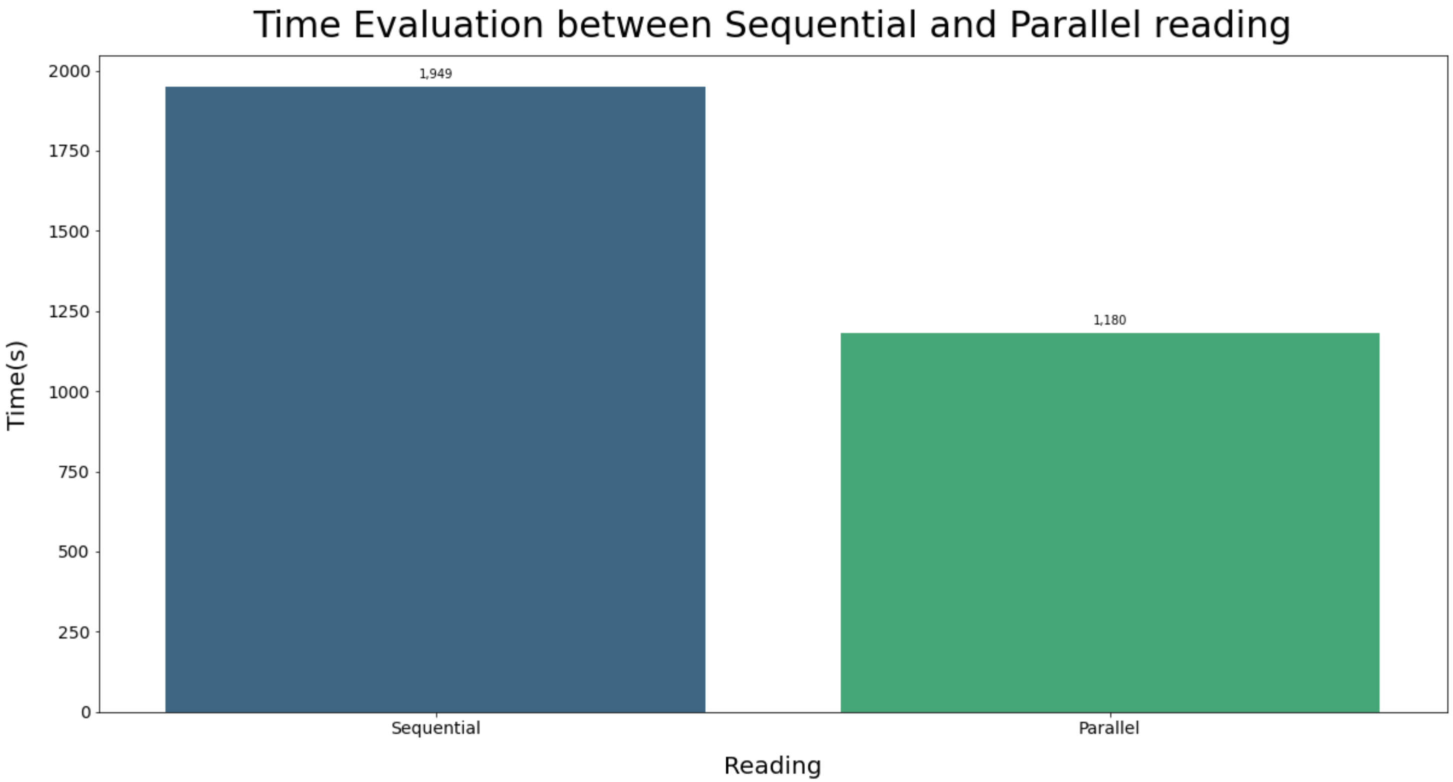
```
! capinfos -z './data_1m.pcap'
```

- Generate all infos

```
! capinfos -A './data_1m.pcap'
```

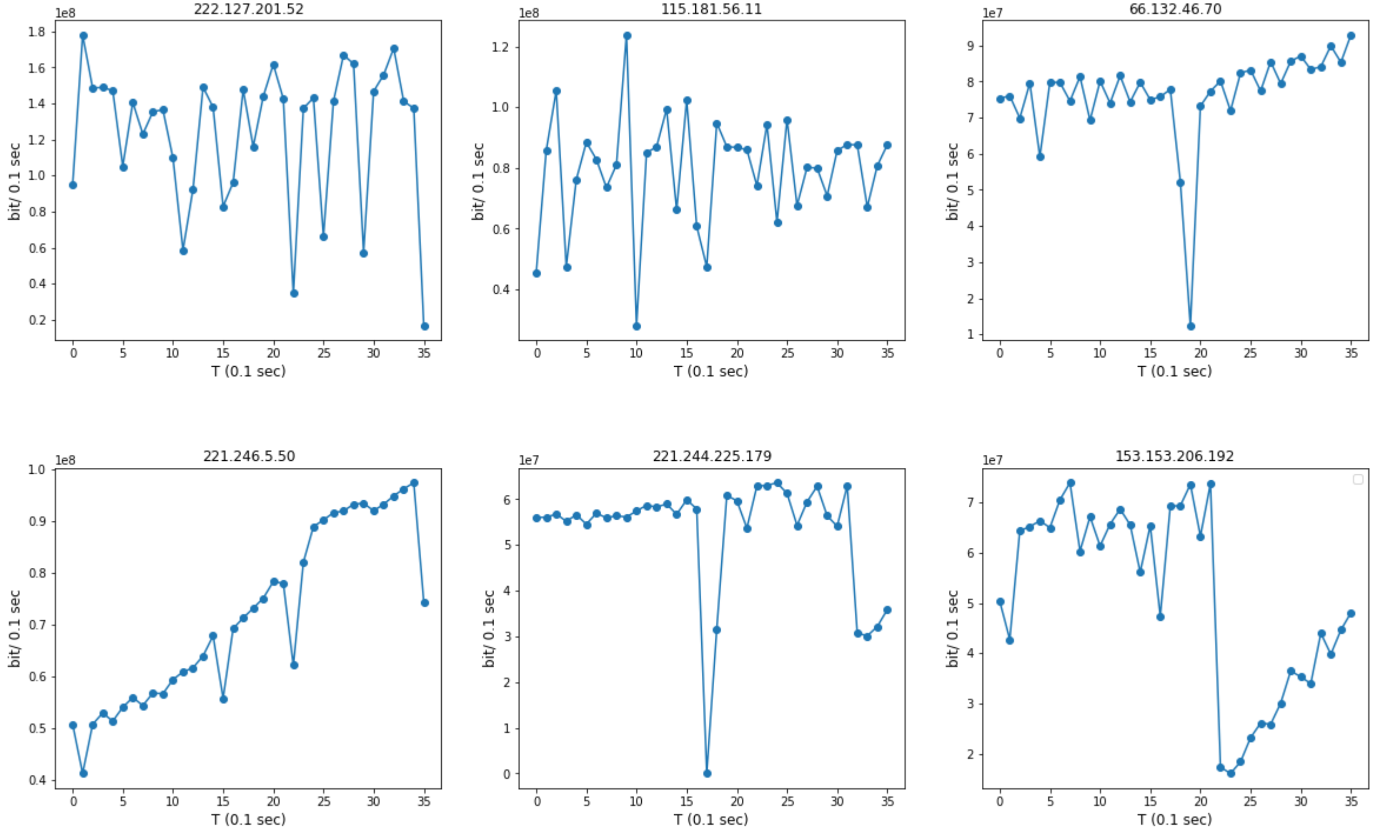
2) Time Evaluation between Sequential and Parallel reading

- Evaluation by executing on Kaggle Notebook with 4 CPUs and 16GB RAM

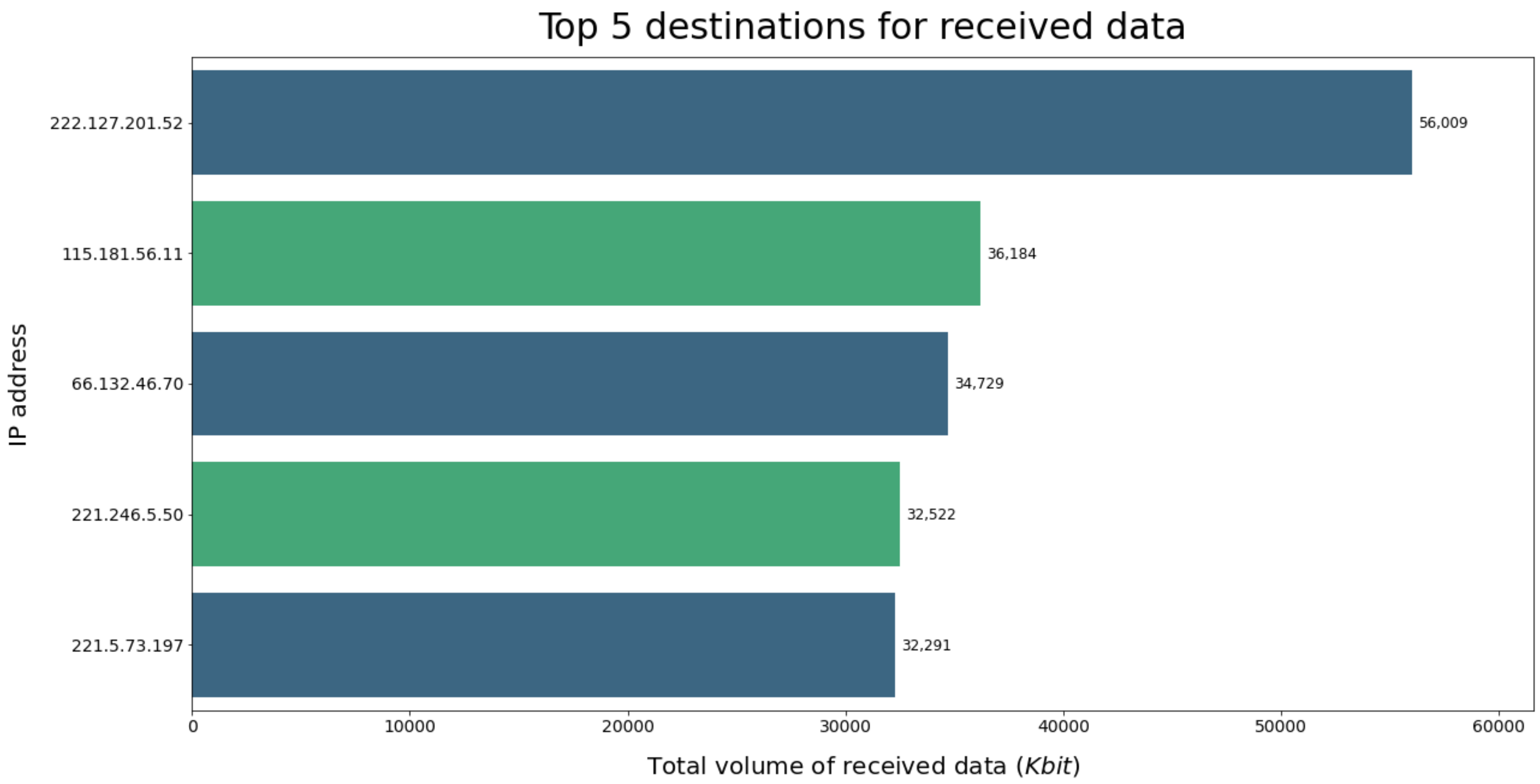


3) Extract the IP which generates the highest amount of sender traffic, evaluate the bit rate (0.1 sec) for the 6 IP addresses mostly used as endpoint

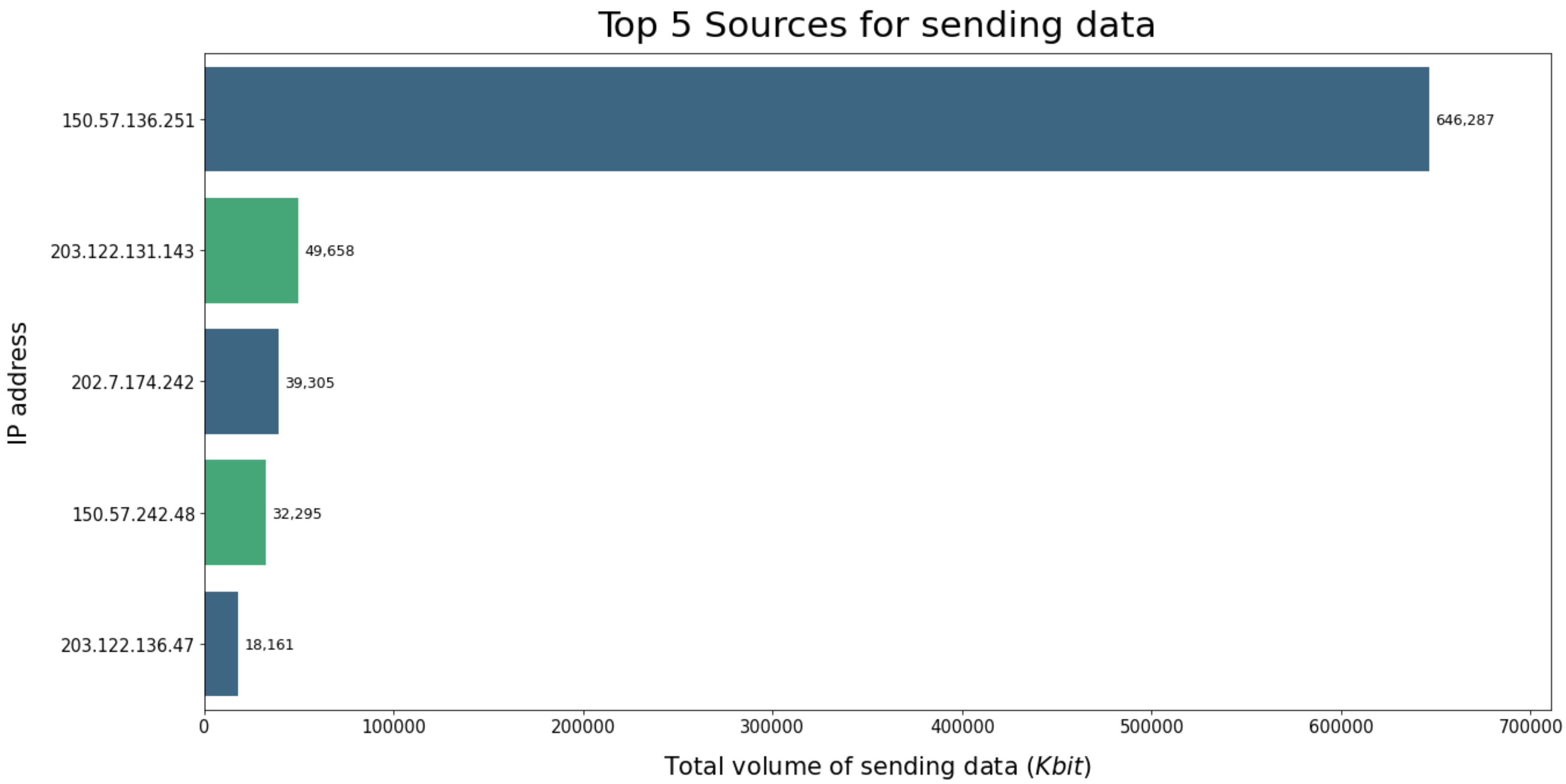
TOP 6 IP Dst for 150.57.136.251



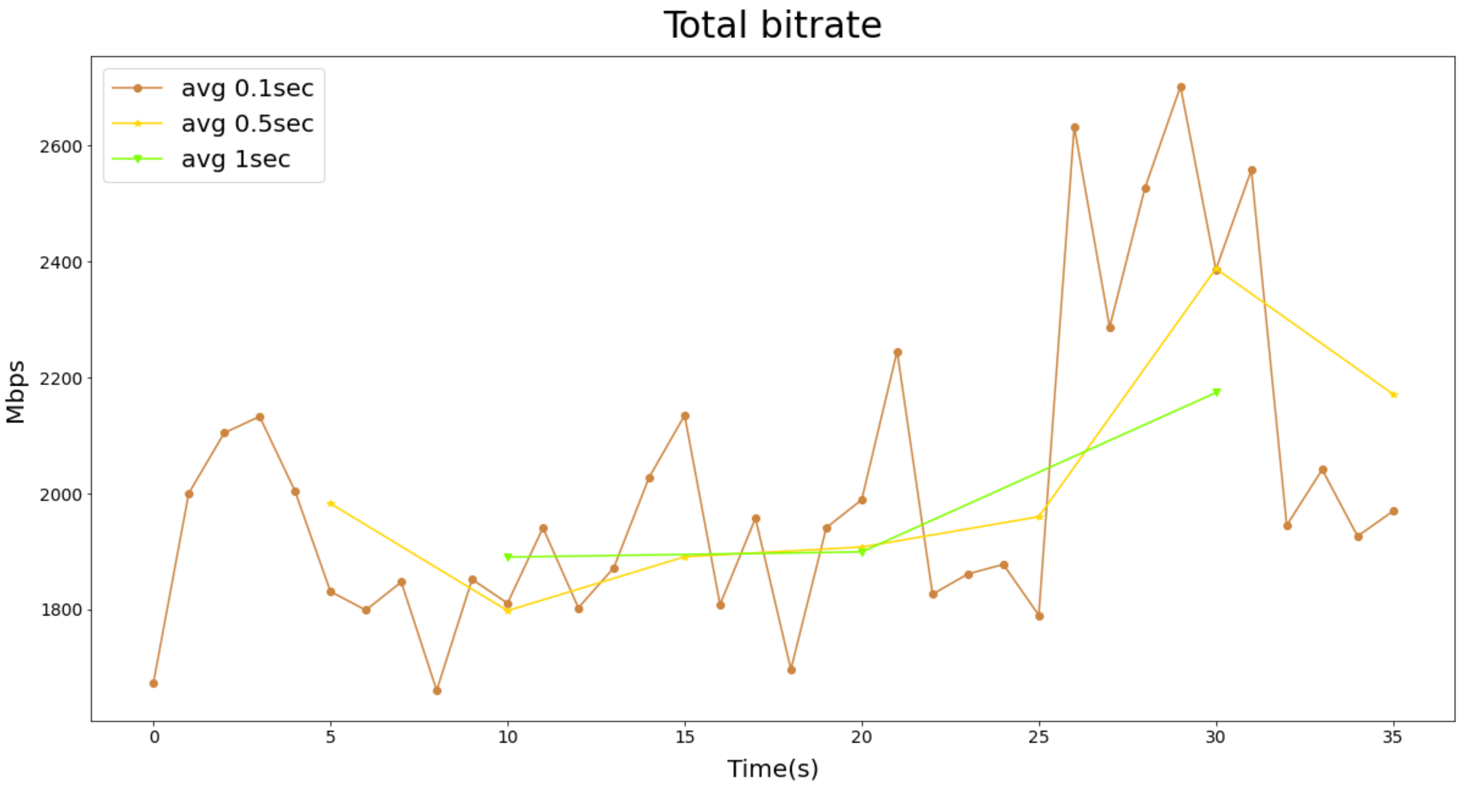
4) Top 5 Destination IP (received bytes)



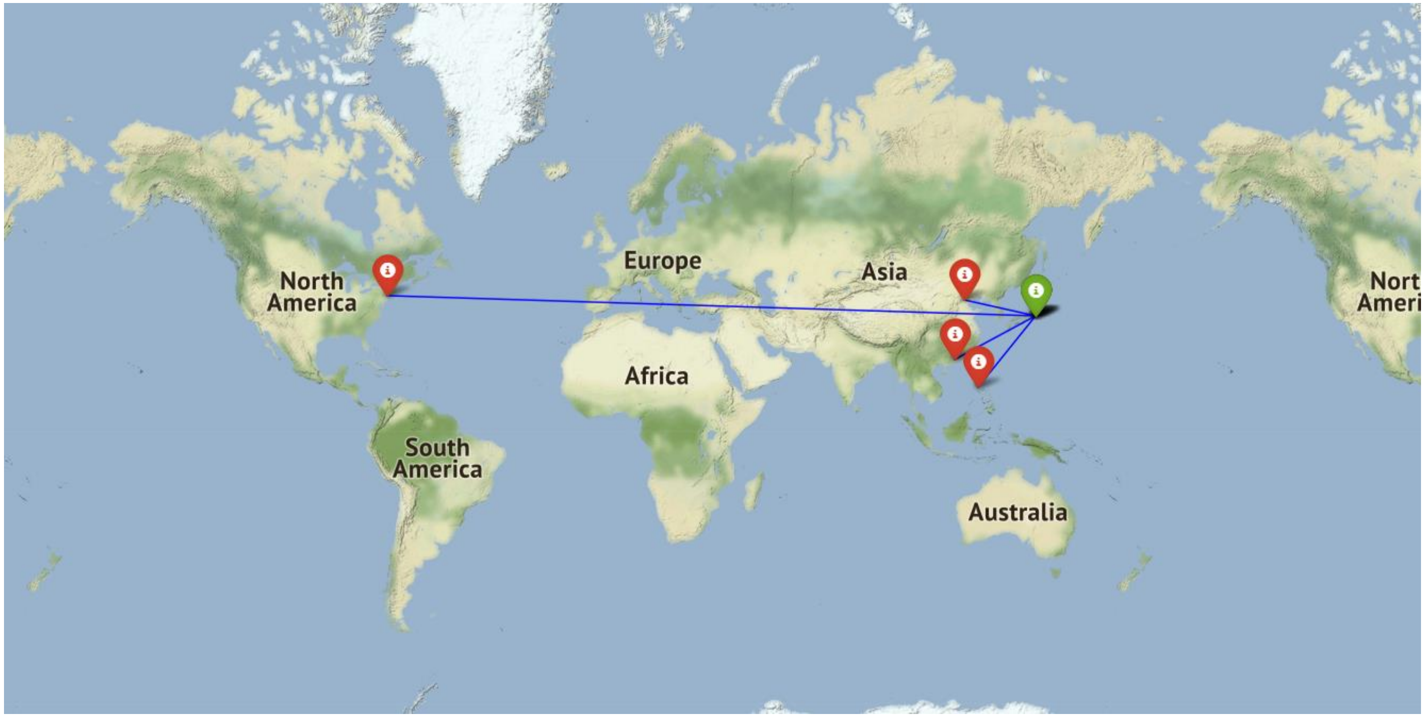
4) Top 5 Source IP (sent bytes)



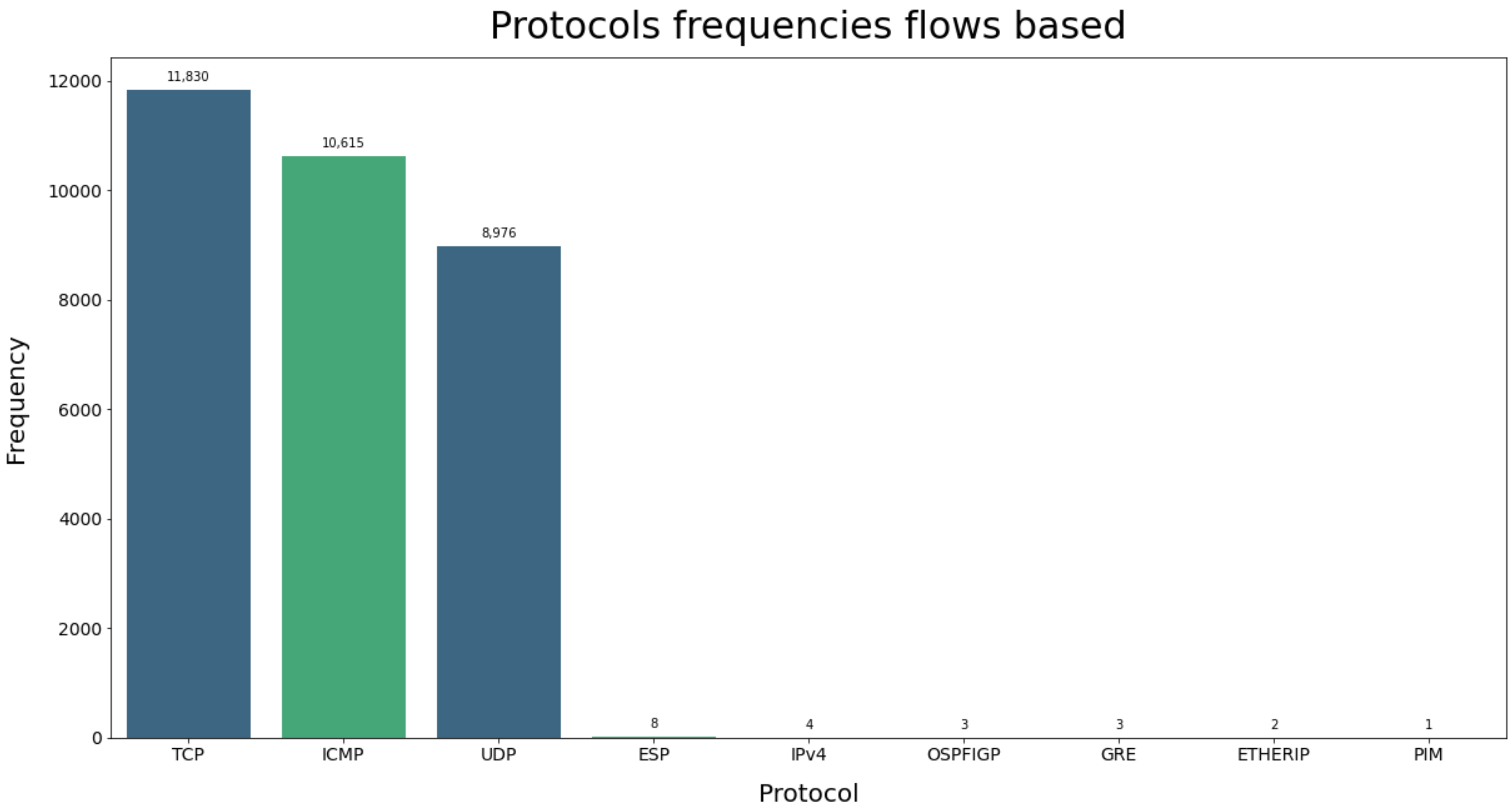
5) Evaluate bitRate considering all the trace with 3 different sampling rate



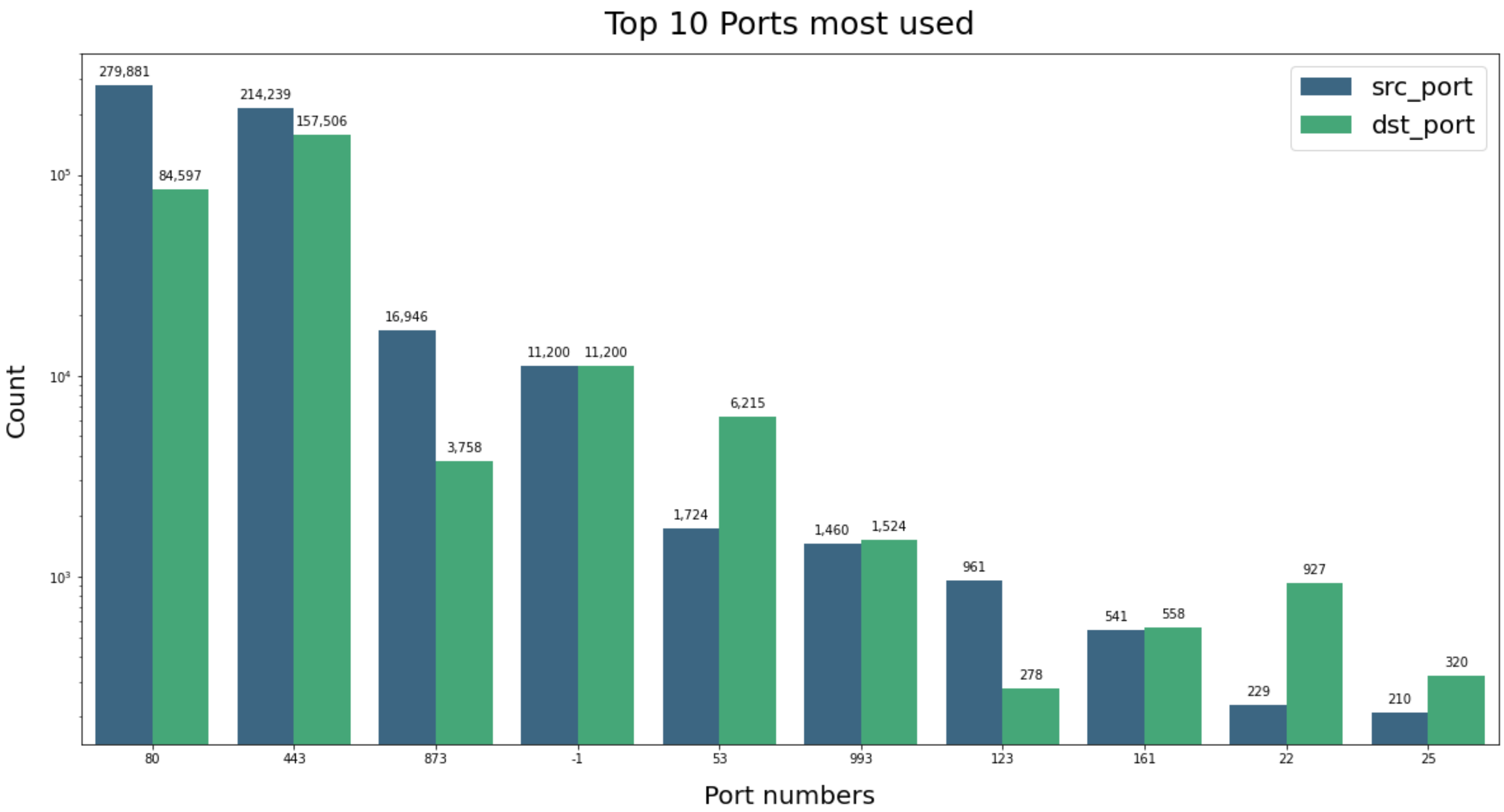
6) GeoLocal Referenciation of the 5 sessions with the highest amount of traffic generated



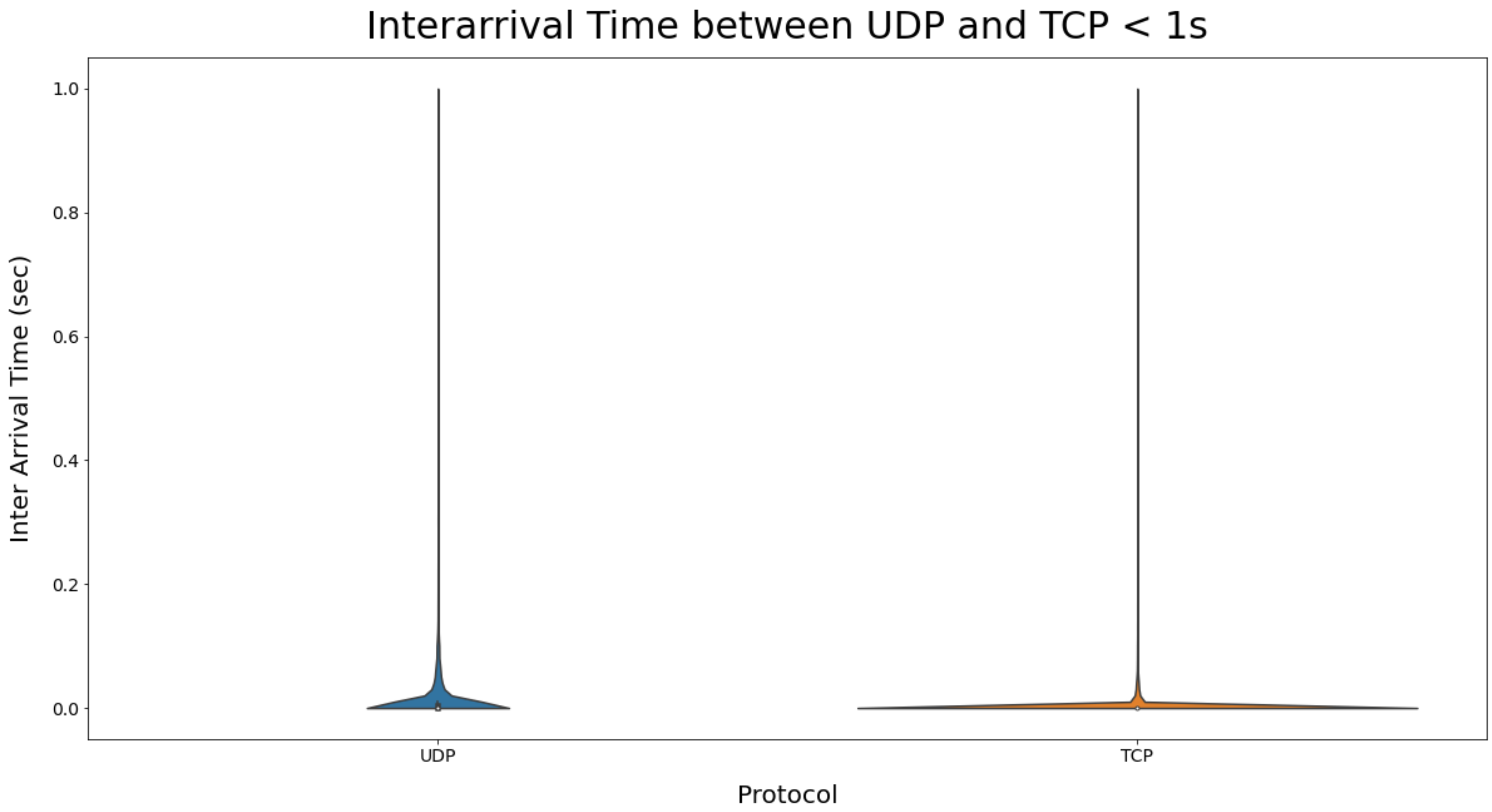
7) 10 Protocols mostly used



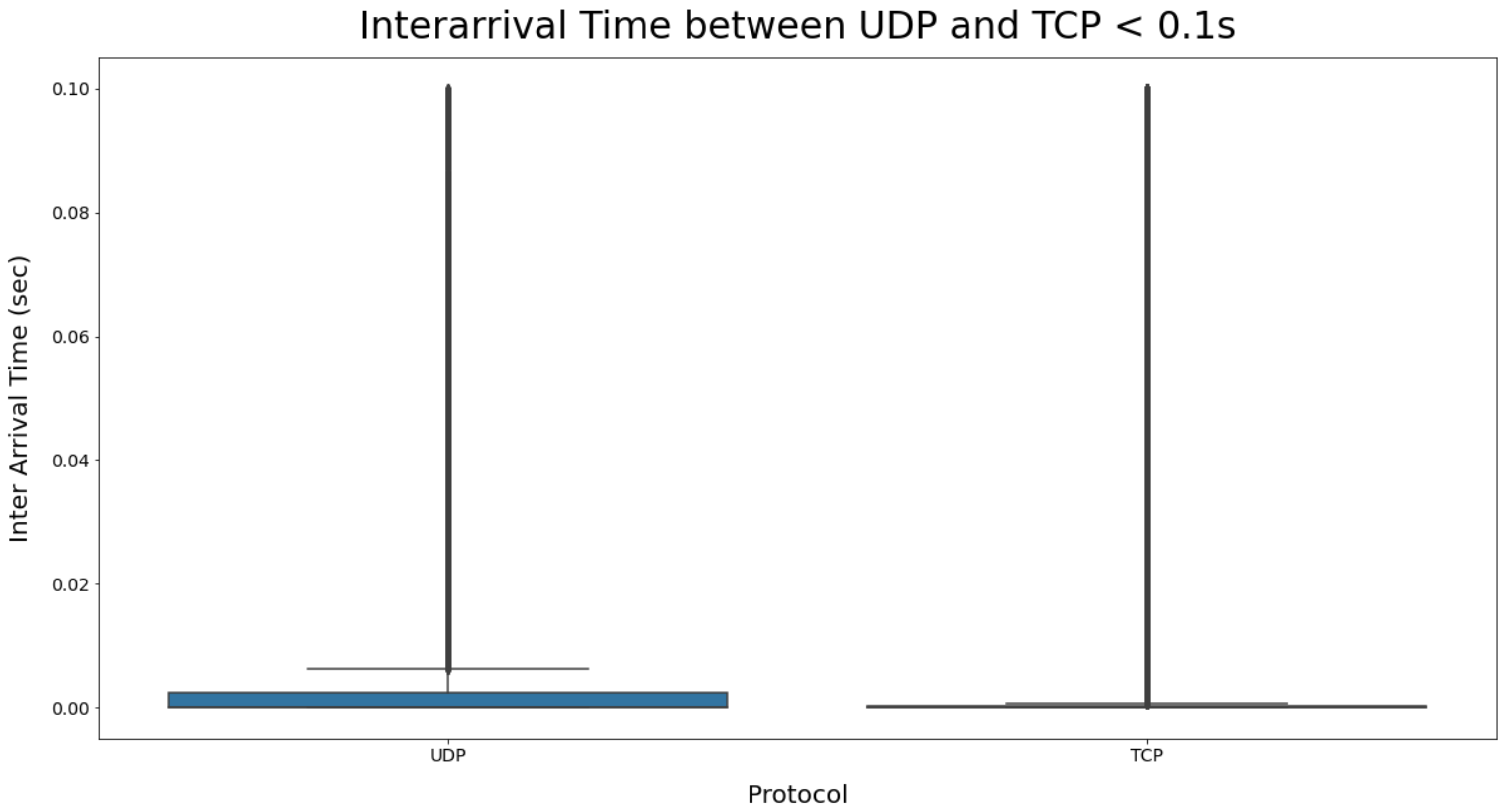
8) Port Scanner evaluation (10 Ports mostly used)



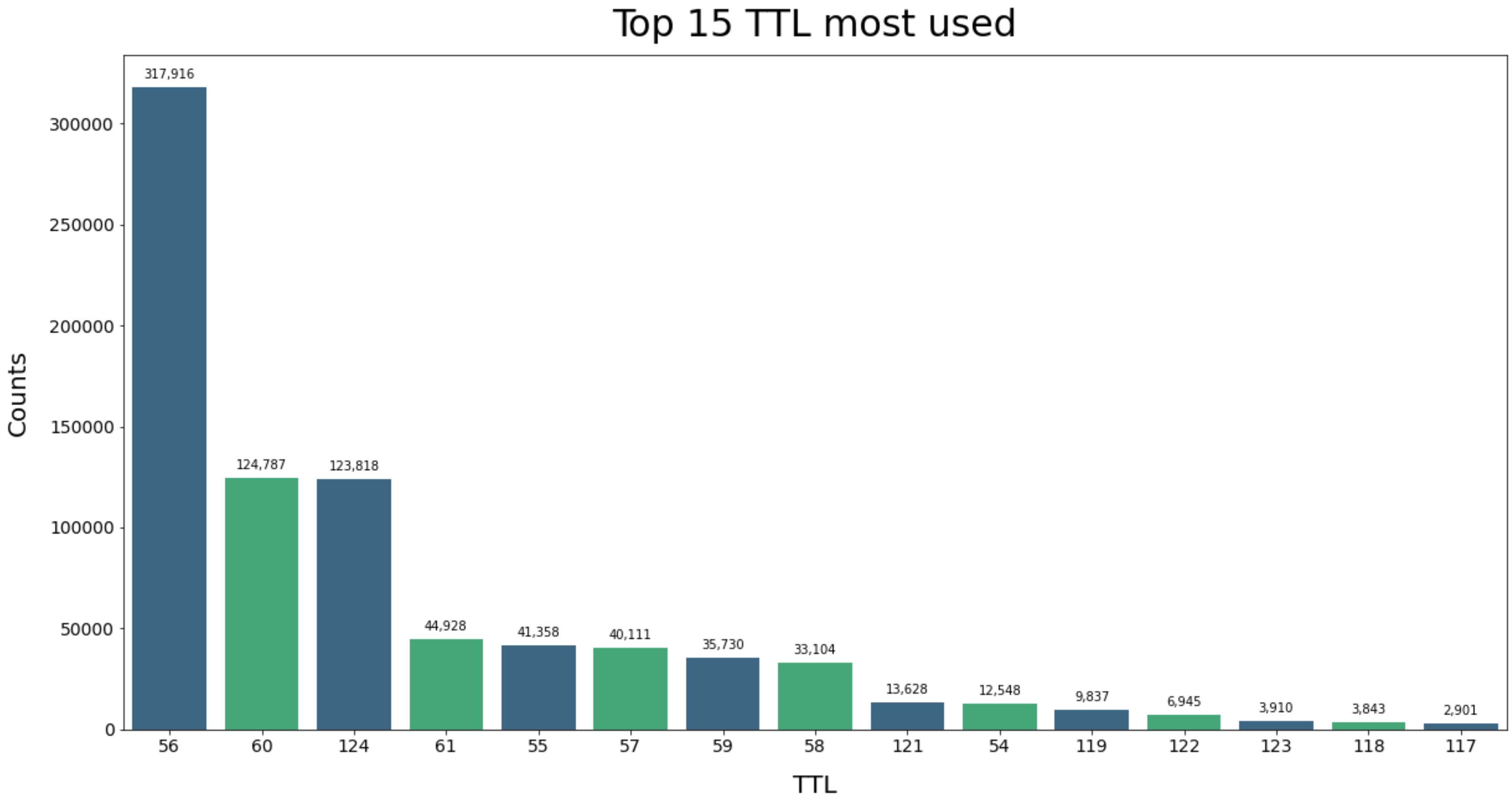
9) InterArrival Time boxplot between TCP and UDP Sessions



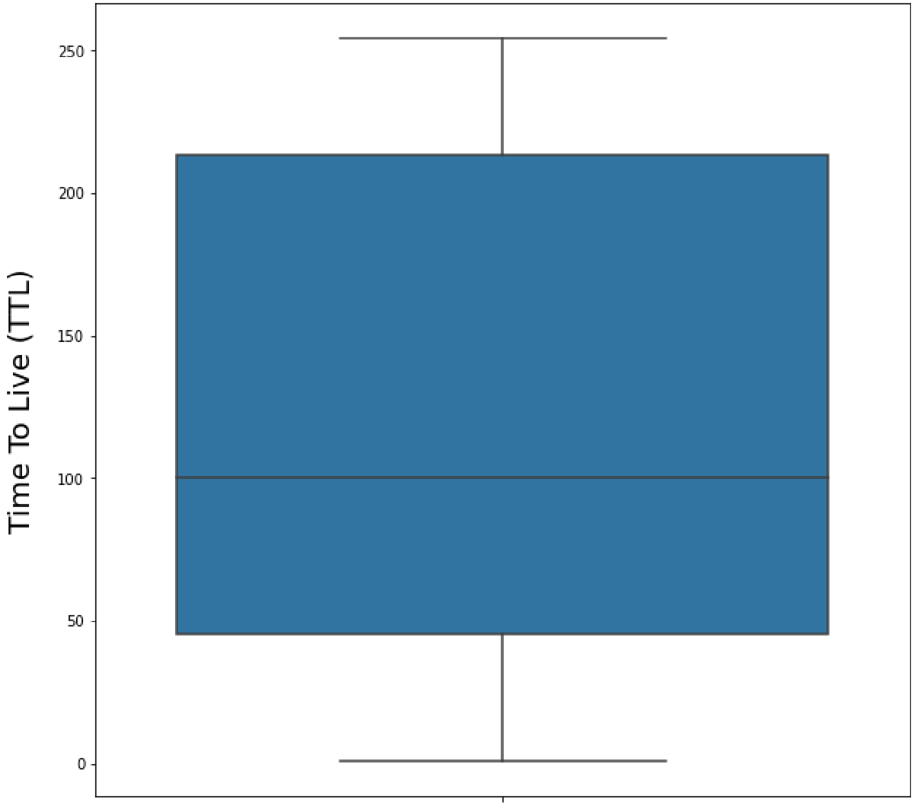
9) InterArrival Time boxplot between TCP and UDP Sessions



10) (Bonus) Top 15 TTL most used



10) (Bonus) TTL



Machine Learning

Problem

Using data given to predict the protocol TCP, UDP and ICMP for each new packet.

Data

832,768 packets of the 1 million of packets from the available capture file.

- Train and Validation data: 75% input data
- Test dataset: 25% input data

- Predictor Variables: {IP Src, IP Dst, Protocol, src-port, dst-port, length, ttl, time}
- Target Variable: {Protocol}

Method

To sovle this classification problem, we’ve used 2 machine learning algorithms SVM and Random Forest.

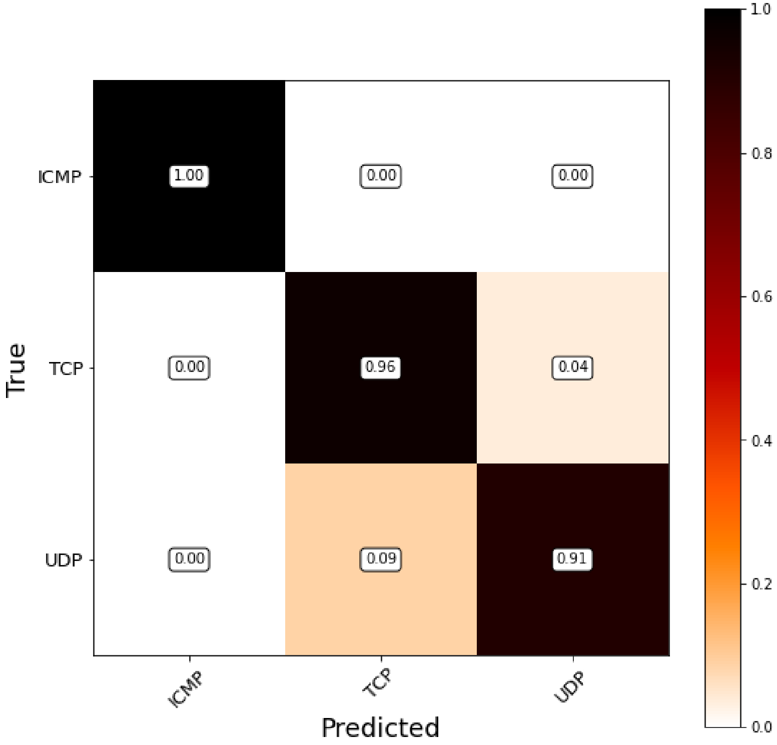
Tuning the hyper-parameters by using RandomSearchCV and GridSearchCV

Data Preprocessing

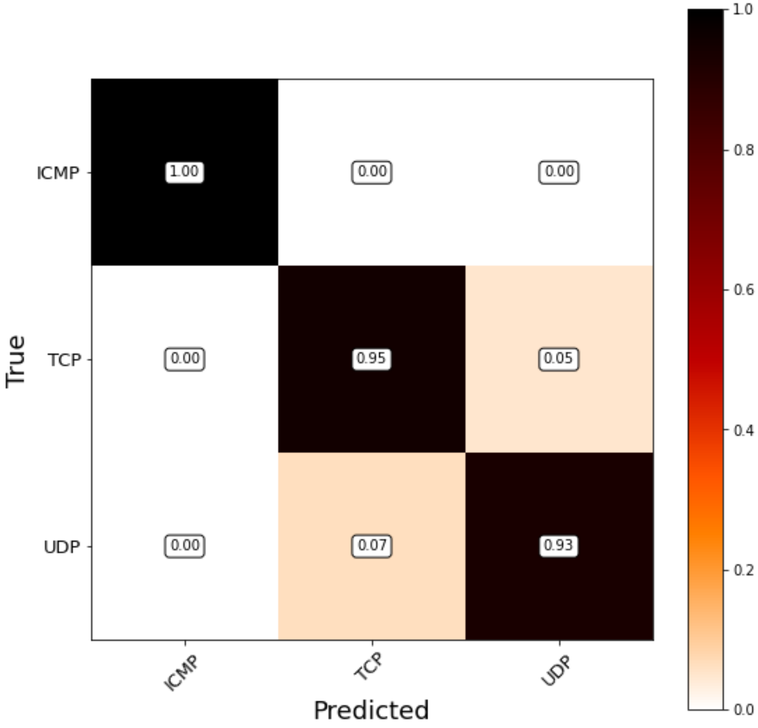
- Missing Values
- Duplicate packets
- One-hot Encode
- Dimentionality Reduction
- Class Imbalance

Evaluation Metrics

Confusion Matrix



Support Vector Machine



Random Forest