# 5G NR DDoS Simulation – Dataset Specification

## 1. Dataset Purpose

Provide ML-ready labeled datasets for training and testing intrusion detection models in 5G NR environments. Labels distinguish between benign vs. attack traffic, and optionally intensity levels (low/medium/high). Usable in scikit-learn, PyTorch, TensorFlow, etc.

## 2. Dataset Structure

Each row = one time window of traffic (e.g., 1 second). Example schema:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| time\_start | time\_end | scenario\_id | ue\_total | attackers | throughput\_bps | pkts\_total | avg\_pkt\_size | jitter\_ms | delay\_ms | unique\_src | unique\_dst | udp\_pkts | tcp\_pkts | pkt\_loss | label\_binary | label\_intensity |
| 0.0 | 1.0 | scenario\_1 | 25 | 5 | 1.2e6 | 1500 | 800 | 2.1 | 10.4 | 5 | 3 | 1400 | 100 | 0 | 0 | 0 |

## 3. Dataset Sources

- FlowMonitor XML → throughput, delay, jitter, packet loss.  
- PCAP files → per-packet sizes, interarrival times, unique IPs, protocol mix.  
- Simulation logs → metadata (attack start/end, attackers, rates).

## 4. Dataset Variations

Freelancer should generate datasets for multiple conditions:  
- UE counts: 25 / 50 / 200  
- Mobility: static vs. mobile  
- Channels: UMa / UMi  
- Attack intensities: low / medium / high  
- Duty cycle: continuous vs. bursty attacks  
  
Each condition ⇒ one dataset, then merge into a master dataset.

## 5. Output Format

- CSV (easy to open in Excel/ML pipelines)  
- Parquet (compact, fast for Python/ML)  
- Organized by folders:  
 datasets/  
 scenario\_25UE\_low.csv  
 scenario\_50UE\_high.csv  
 scenario\_200UE\_mixed.parquet  
 master\_dataset.csv

## 6. Deliverables

- Raw datasets (per scenario & combined).  
- README explaining features, labels, and simulation parameters.  
- Plots (throughput/time, loss/time, benign vs. attack comparison).