**# Model Evaluation Report — CICIDS2018 Dataset**

This report summarizes the accuracy and latency comparison of the DistilBERT model before and after quantization.

---

**##  Accuracy Results**

| Model Type         | Accuracy |

|--------------------|----------|

| FP32 (Original)    | 91.00%   |

| Quantized (PTQ)    | 96.00%   |

---

**## Latency Results**

Average inference time measured over 100 runs on a single example input.

| Model Type         | Average Latency (ms) |

|--------------------|----------------------|

| FP32 (Original)    | 17.03 ms             |

| Quantized (PTQ)    | 7.07 ms              |

---

**##  Summary**

- Quantized model achieved **\*\*higher accuracy\*\*** and significantly **\*\*lower latency\*\***

- PTQ was done using `torch.quantization.quantize\_dynamic`

- Inference used realistic tabular data from `02-14-2018.csv` converted into model input strings

---

**##  Files:**

- `prepare\_data.py` – prepares `cicids2018\_test.csv` from raw CICIDS CSVs

- `eval\_fp32.py` – evaluates accuracy of FP32 model

- `eval\_quantized.py` – evaluates accuracy of PTQ model

- `profile\_latency.py` – benchmarks inference latency

- `cicids2018\_test.csv` – 500 test samples (features + labels)