Author: Pavan Shetty, Northeastern University, Shepherd Lab

* This Document explains the procedure of inference Test of trt models on Jetson Orin Nano utilizing GPU for high speed prediction
* **Folder Information: TRT Test**

Please find **TRT Test** folder at:

<https://github.com/Pavan-r-shetty/trt_test.git>

You would need to specify model input shape same as the one used for your optimization profile in the .hf to .trt conversion

In the file trt\_run\_test.py you can find a lot of low level programming which mainly sets streams between GPU and CPU. Preallocation of memory and setting streams are necessary for high speed prediction. If you don’t, you will get a memory allocation warning and Jetson hangs.  
  
Jetson always needs test runs before you take something meaningful from the prediction. These pretest runs are necessary for Jetson to get to it’s insanely fast speed. (warm-up)

The code takes random inputs of the shape (1,80,8) and predicts 2 values  
80 is my timeseries window

8 is my input channels

1 is the expanded dimension to be considered as a tensor

The script runs for 10 trials and prints prediction speed in ms and corresponding Standard Deviation