Data Cleaning

One of the first things we considered our RNN-LSTM model was that the neural network doesn't train well with a lot of null/Nan values which were abundant in the CMOD dataset.
We first ignored all the samples for shots that were intentionally disrupted since these disruptions are different in nature than the ones we are trying to predict.
We also calculated the percentage null values and correlation for each feature and decided to drop the following features from the train dataset due to large number of missing values (more than 30%) and low correlation:
mirnov
te_width
z_error
z_times_v_z
zcur
v_z

To deal with the remaining missing values, we dropped the corresponding samples from the data. We do not think this will greatly affect our train accuracy by a lot since the percentage of samples droppe is less and the irregularity in time samples is minor. We also dropped the time and intentional_disrupt columns as they are not relevant to measuring the predictability.

Two-fold Model

Our two fold model comprised of an LSTM used to classify if a disruption is going to occur based on a window of initially 10 timesamples. We also used the trained LSTM to extract the essential features in our dataset to use in the next step in our analysis. The LSTM accounted for analysing the temporal causal relations in the data and therefore is a better predictor of disruption occurrence than a simple regressor. The features and predictions extracted from the LSTM are then fed into a Random Forest Regressor to predict the time_until_disruption. We made this choice since Random Forest Regressors are more robust to missing values.

The hyperparameters that can be fine tuned in our models are as follows:

1. Time threshold - This characterizes how far ahead in time can the model predict disruptions. We need to fine tune this parameter so that the model doesn't optimize on random correlations in time which do not cause disruptions.
2. Loop back window- the loop back window for LSTM. It is the number of time samples it trains on at a time.
3. Max depth - The maximum depth of the random forest regressor
4. N_estimators - the number of estimators for the random forest regressor