Task 3 Report Data Selection

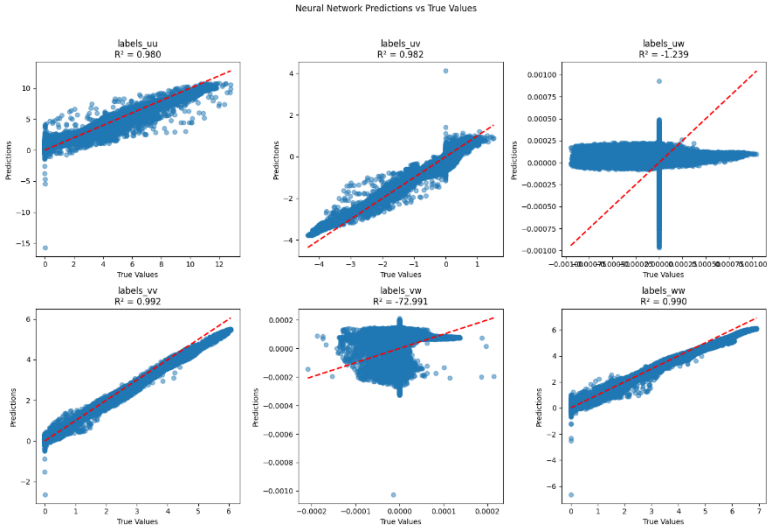
December 18, 2024

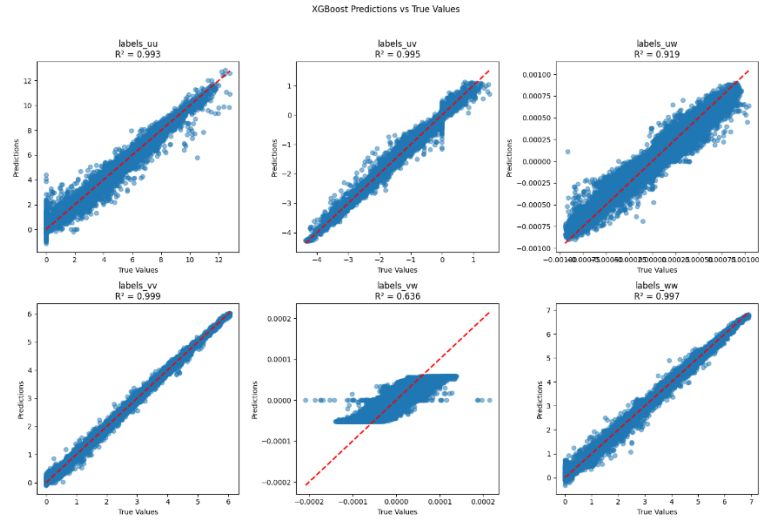
Task 3 Report

After a lot of consideration over data selection, I have had found lots of enormous data that maybe could have been better nut faced accessibility problems and the fact they are enormous one of them was actually a couple of TBs, I have found this dataset on Kaggle: https://www.kaggle.com/datasets/ryleymcconkey/ml-turbulence-dataset?
By taking the kepsilon.csv file, containing RANS simulation data, I have worked on its features having 283 features which is a lot. After checking for nulls and duplicates, I had them reduced by performing PCA on the data to reduce dimensionality, then I performed scaling, split eh data and started training.

Training

Started with implementing DNN with 3 layers having dropout and batch normalization layers too, to avoid overfitting and it occurred that the targets were not all predicted correctly as the loss was very high in two of the 6 labels corresponding to DNS/LES.
Second approach was using XGBoost, that came off as the best model at last and increased the accuracy of the predictions for these two faulty labels of DNN. For the labels_vw its R2 score was R2 -72.990915 0.636258 for DNN and XGBoost respectively and the same for labels_uw R2 -1.239369 0.918971


Neural Network Predictions vs True Values
labels_uu
R² = 0.980
labels_uv
R² = 0.982
labels_uw
R² = -1.239
labels_vv
R² = 0.992
labels_vw
R² = -72.991
labels_ww
R² = 0.990
Predictions
True Values


XGBoost Predictions vs True Values
labels_uu
R² = 0.993
labels_uv
R² = 0.995
labels_uw
R² = 0.919
labels_vv
R² = 0.999
labels_vw
R² = 0.636
labels_ww
R² = 0.997
Predictions
True Values