

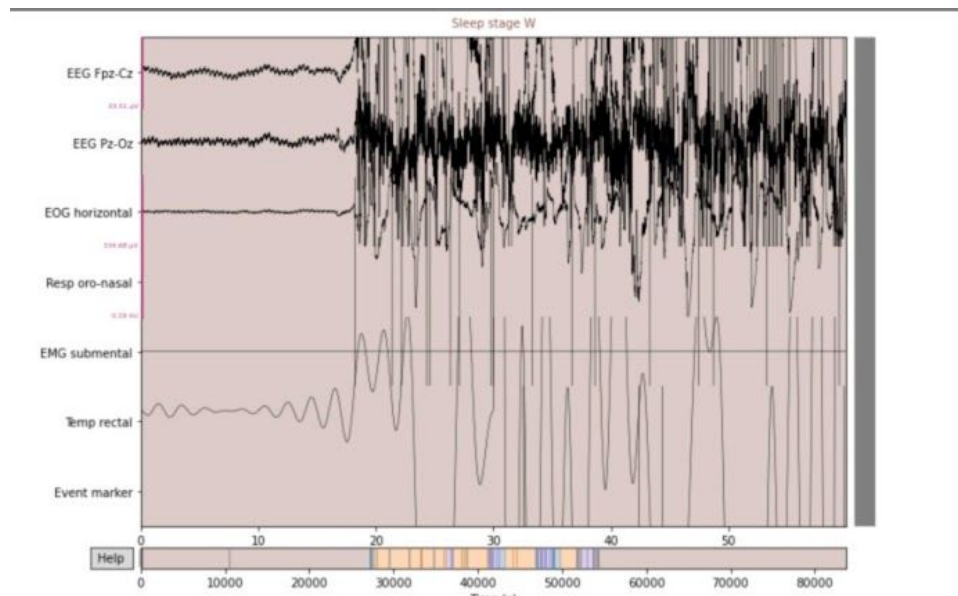
# Preliminary Results

## Machine Learning models

We set up the initial machine learning pipeline to load the data, extract features from the raw data, split the data into training/test sets and performed some early exploration using a few basic classifiers (random forest) and plot the AUC and confusion matrix for the testing results.

### Data Transformation/Feature extraction

The raw sleep data is time series data of different channels with the annotation for different sleep stages. In the initial results, we just used four stages which present in all data: W (wake), Sleep stage 1, Sleep stage 2, REM stage. Please find some example data below:



For each data file, we remove part of the wake stage data (at the beginning and the end of the records) to make the data more balanced. Then the data is filtered to remove background signals. The data is chopped into 30 seconds segments.

In our initial feature extraction, we decompose the signals into five frequency bands: delta, theta, alpha, sigma and beta as features and then extract the power spectral density (PSD) for each band as the feature for training (Ref: Sleep Stage Classification Using EEG Signal Analysis: A Comprehensive Survey and New Investigation).

### Model training and testing

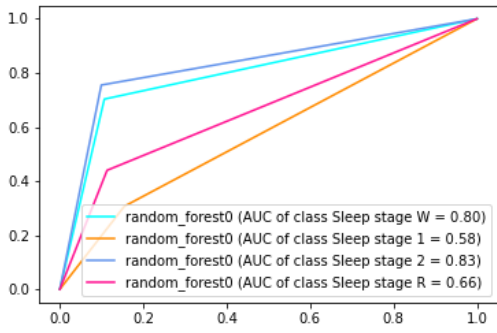
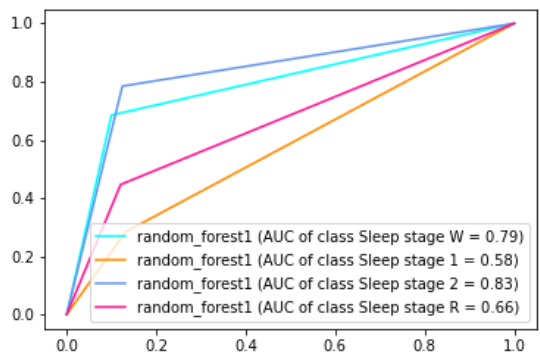
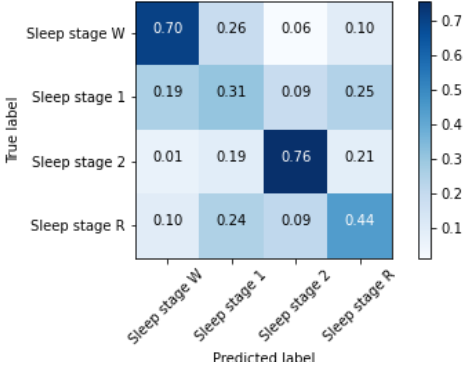
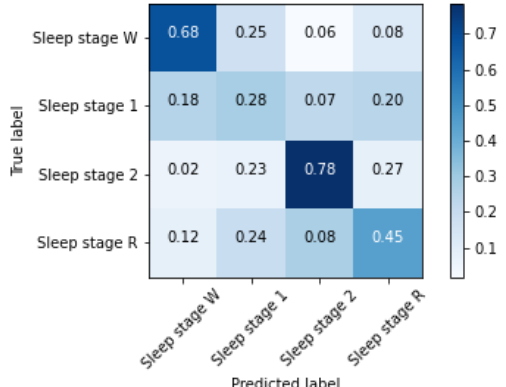
The data is split into two sets: training set (80% of the samples) and testing set (20% of samples). Then we feed the training set to the model. The trained model is evaluated using the

testing set for accuracy, roc-auc and confusion matrix. In the initial exploration, I tried random forest classifier from sklearn.

## Results comparison

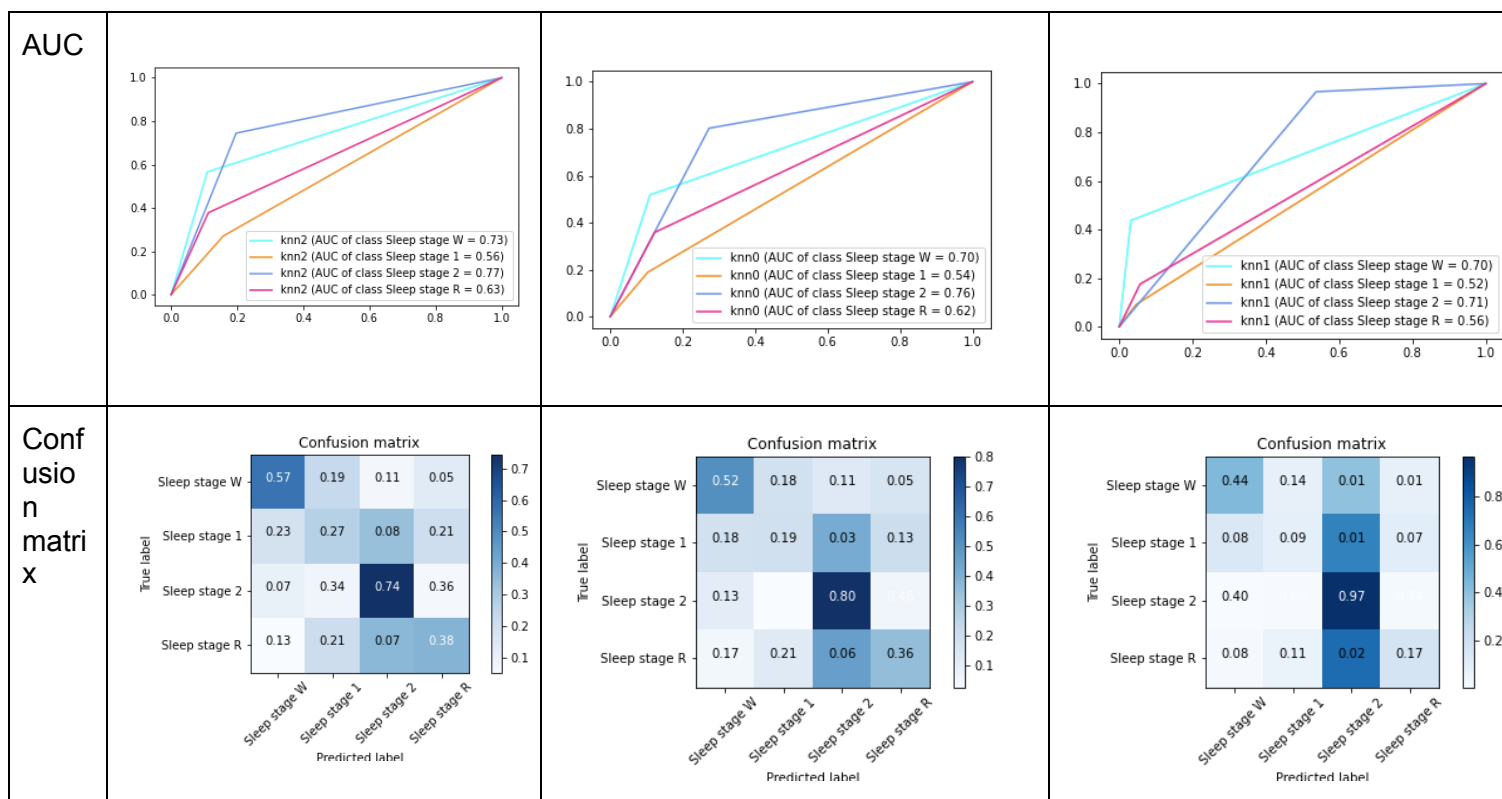
### Random Forest

I tried to limit the max\_depth of the random forest to 8 or no limit. This run without max\_depth seems to have slightly better performance. The results are shown below.

	Random forest with no limit on depth (rf0)	Random forest with max_depth=8 (rf1)
AUC	 <p>random_forest0 (AUC of class Sleep stage W = 0.80)  random_forest0 (AUC of class Sleep stage 1 = 0.58)  random_forest0 (AUC of class Sleep stage 2 = 0.83)  random_forest0 (AUC of class Sleep stage R = 0.66)</p>	 <p>random_forest1 (AUC of class Sleep stage W = 0.79)  random_forest1 (AUC of class Sleep stage 1 = 0.58)  random_forest1 (AUC of class Sleep stage 2 = 0.83)  random_forest1 (AUC of class Sleep stage R = 0.66)</p>
Confusion matrix	<p>Confusion matrix</p>  <p>True label</p> <p>Predicted label</p>	<p>Confusion matrix</p>  <p>True label</p> <p>Predicted label</p>

### Nearest Neighbors

Nearest Neighbors, K = 5 (knn2)	Nearest Neighbors, K = 20 (knn0)	Nearest Neighbors, K = 100 (knn1)
---------------------------------	----------------------------------	-----------------------------------



The performance of current machine learning models are still worse than reported results. Considering that this is still our initial execution, we are focusing more on data exploration and pipeline setup. We will definitely explore more regarding the feature engineer and model hyperparameter tuning.

## Next Steps

1. Load and process data using pyspark
2. Try different features
3. Try different classifiers and compare their performance