Project Checkpoint Report
B11 Group 1
Kevin Chin, Shaheen Daneshvar, Yilan Guo

Introduction:

Obstructive sleep apnea (OSA), the more common form of sleep apnea, is a sleeping disorder where breathing stops and starts intermittently. OSA happens when muscles in the throat get relaxed, narrowing the airway and hampering breathing for 10 seconds or longer, causing blood oxygen concentration to decrease and a buildup of carbon dioxide. Such sudden drops in oxygen levels cause sudden increases in heart rate and blood pressure, resulting in repeated, transient strains on the cardiovascular system. OSA increases the risk of stroke and the risk of irregular heart rhythms or arrhythmias; both stroke and arrhythmias have the potential to cause sudden death.

Sleeping is not uniform and consists of four stages: N1, N2, N3, and REM sleep. The analysis of sleep stages is essential for understanding and diagnosing sleep-related diseases, such as insomnia, narcolepsy, and sleep apnea.; however, there is not enough research on sleep stages and sleep stage classification for sleep apnea. The goal of our project is to identify and classify the sleep stage for people with sleep apnea and understand how it differs from the normal sleep stage.

This report aims to document scientific investigations we have done on EEG classification and exploratory data analysis (EDA) performed on sleep polysomnography data provided by the Sleep Heart Health Study PSG Database under the National Heart Lung & Blood Institute. The methods we have examined and plan to build our model on are YASA classifier and LGBM classifier models. The result section contains the EDA of some essential exploration of the dataset to understand its characteristics and patterns, cleaning the missing values and irregularities such as outliers to improve our results, visualizations representing the data, and some analysis.

Description of Methods:

Because the SHHS consists of two visits, and the number of participants in the second visit is clearly smaller than that in the first visit, to make sure we have a complete record of both visit polysomnography data, we decided to start EDA on the second visit participants. To accord with the epoch period in our later model planning, EEG, ECG, and EOG data are analyzed and visualized in 30-sec periods (One period EDA is shown in the Result section). We also plot an overnight spectrogram of EEG, which demonstrates the relationship between time and Frequency.

We have mainly examined two models: Yasa Sleep Classifier[1] and LGBMClassifier model[2]. Yasa Sleep Classifier is implemented using heterogeneous datasets from the National Sleep Research Resource (NSRR) and the Dreem Open Dataset (DOD) and provides a basic framework for our project. In YASA's preprinted paper, it detailed its preprocessing and feature extraction, which exclude the nights which had poor PSG quality (certain sleep stage is missing) and the nights with abnormal durations and downsample all the EEG, ECG, and EOG data into 100 Hz for computational convenience. Frequency-domain features are extracted using the spectral power in different bands, and other quantitative variables such as age and sex are also added to construct the final model. After that, their model is trained using a LightGBM Classifier which is a well-performed gradient boosting classifier based on decision tree and achieved a median accuracy of 85.9% in testing nights (Vallat, 2021).

The most important feature we plan on investigating are EEG signals. To do so, we will be extracting features from both the time domain and the frequency domain. Each epoch/bin (30 seconds) will have features such as its absolute spectral power, fractal dimension, and standard deviation. Because of the nature of sleep, bins are not independent of each other—an epoch that corresponds to a wake state will likely not be followed by an epoch of N3 sleep—so we will also use some metric that provides information from surrounding bins. An example would be to use a rolling average across time as a feature. Similar metrics will be used for both EOG and EMG signals. Most other classifiers we researched didn't include ECG signals in their model, so we plan to include ECG signals in our classifier. To do so, we will create RR-intervals and use time-domain analysis by taking metrics, such as mean and standard deviation, between interval peaks. Using all of these features, we will pass them into a Light Gradient Boosted Machine to classify each epoch as one of the following stages: W (wake state), R (REM), N1, N2, N3.

Preliminary Results:

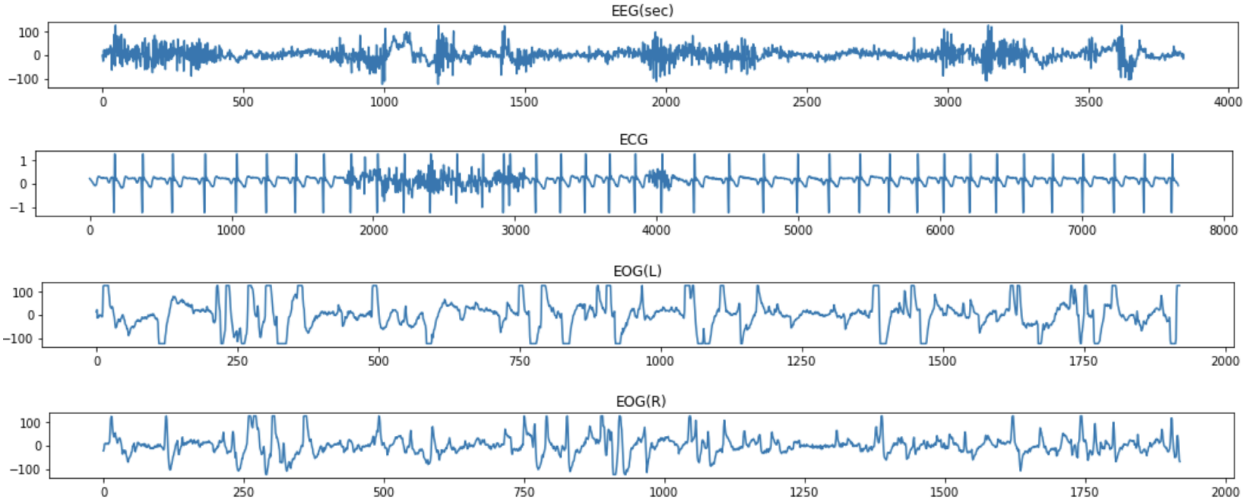
- **EDA:**

The polysomnography data from the SHHS usually consist of about eight hours of data, and they are measured in different frequencies. There are 11 variables which are measurements of the following (units): SaO2 (%), PR (BPM), EEG secs (uV), ECG (mV), EMG (uV), EOG Left (uV), EOG Right (uV), EEG (uV), Airflow (V), Thor Res (V), and ABDO Res (V). The following is a snippet of the dataset:

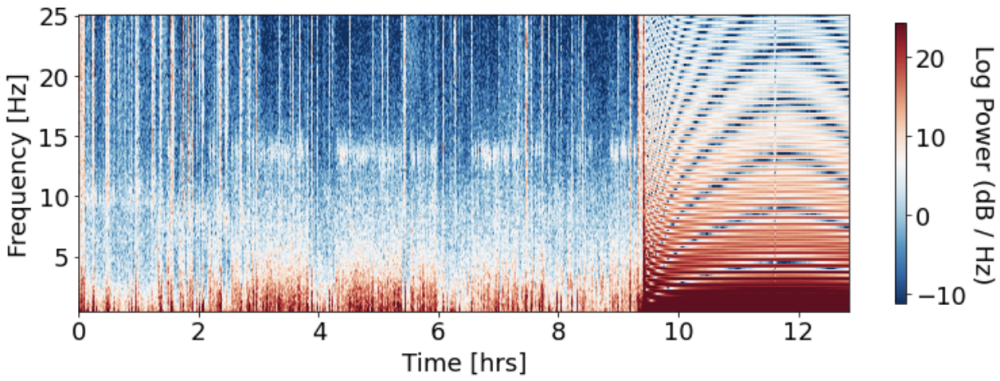
	SaO2	PR	EEG(sec)	ECG	EMG	EOG(L)	EOG(R)	EEG	AIRFLOW	THOR RES	ABDO RES	POSITION	LIGHT	OX STAT
0	95.115587	76.371405	-2.450980	0.210784	-2.594118	11.274510	-23.039216	-6.372549	-0.905882	-0.286275	-0.607843	3.0	0.0	0.0
1	96.092164	77.347982	-8.333333	0.200980	-1.852941	21.078431	-16.176471	16.176471	-0.490196	-0.247059	-0.521569	3.0	0.0	0.0
2	96.092164	79.301137	-16.176471	0.191176	-3.088235	-14.215686	5.392157	19.117647	-0.090196	-0.160784	-0.466667	3.0	0.0	0.0
3	96.092164	81.254292	-26.960784	0.191176	5.311765	-8.333333	8.333333	10.294118	0.239216	-0.090196	-0.294118	3.0	0.0	0.0
4	96.092164	83.207446	-16.176471	0.181373	-8.770588	-10.294118	6.372549	5.392157	0.474510	0.027451	0.278431	3.0	0.0	0.0

[1] https://github.com/raphaelvallat/yasa_classifier
[2] https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html

To better understand and explore the underlying structure of the data, we adopt commonly used 30-s epochs inspection to visualize the EEG, ECG, EOG(L), and EOG(R), the variables we plan to use to construct our model.



The following figure is a spectrogram of EEG, which shows the time-frequency representation of the whole-night recording. From the spectrogram, the participant was awake after about nine hours of sleep. Once we complete our mode, this graph can be paired with different sleep stages.



Appendix

Project Report

Reference

Vallat, Raphael, and Matthew P. Walker. “A Universal, Open-Source, High-Performance Tool for Automated Sleep Staging.” *BioRxiv*, Cold Spring Harbor Laboratory, 1 Jan. 2021, https://www.biorxiv.org/content/10.1101/2021.05.28.446165v1.full.