Sleep Stage Classification for Patients with Sleep Apnea

DSC 180B B011
Group 1
Kevin Chin • Shaheen Daneshvar • Yilan Guo

Abstract

Sleep can be classified into four stages: N1, N2, N3, and REM sleep. While sleep stage classification models do exist, they often do not generalize well to patients with sleep apnea. The goal of our project is to build a sleep stage classifier specifically for people with sleep apnea and understand how it differs from the normal sleep stage. We then explore whether or not the inclusion and featurization of ECG data will improve the performance of our model. Without ECG signals, our model achieves an accuracy rate of 87.14% in the validation set; and surprisingly, our model's accuracy rate decreases to 82.08% in the validation set.

Introduction

Obstructive sleep apnea (OSA), the more common form of sleep apnea, is a sleeping disorder in which breathing stops and starts intermittently. This is caused when muscles in the throat get relaxed, narrowing the airway and hampering the breathing for 10 seconds or longer. This causes 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, sleep stage classification often does not generalize to patients with sleep apnea. The goal of our project is to build a sleep stage classifier specifically for people with sleep apnea and understand how it differs from the normal sleep stage. We will then explore whether or not the inclusion and featurization of ECG data will improve the performance of our model. The thought process behind this is that ECG readings can indicate obstructive events during sleep, which gives information about the patient's sleep state.

Previous Work

Currently, sleep stage scoring is still typically done by hand, requiring a human to label each 30 second epoch based on predefined rules [1]. Because of this, the current method of scoring is somewhat subjective, with human scorers only agreeing 83% of the time [2]. Even individual scorers only have about 90% agreement with themselves when presented with the same signal at different times [3].

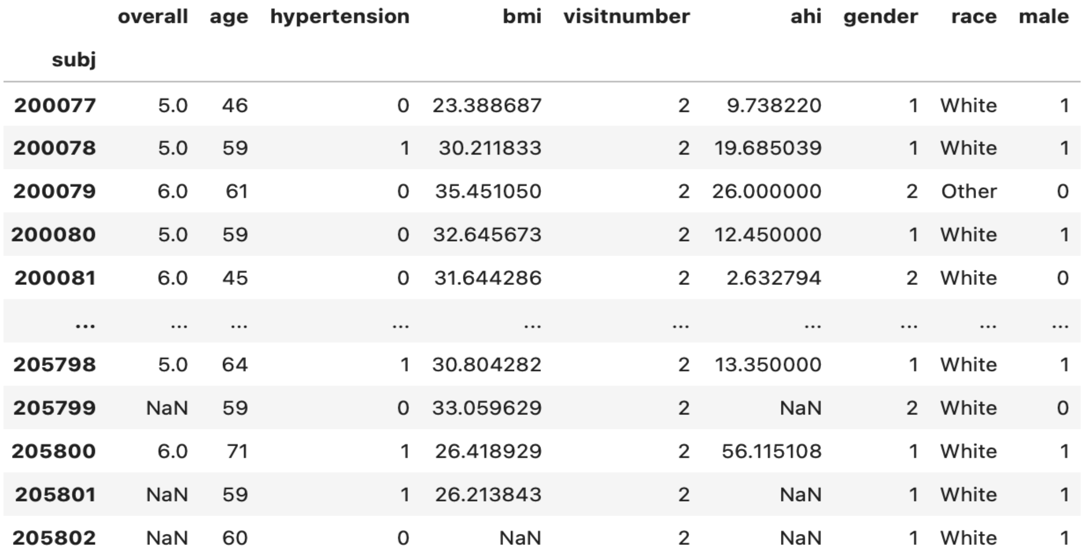
Motivation and Goals

There has been previous work on sleep band classification, but those models often do not generalize well to patients with sleep apnea. In particular we suspect that utilizing ECG data may improve our model's performance; this is because patients with sleep apnea frequently wake during their sleep cycle. These obstructive events can be identified using ECG data and, therefore, considered by the model.

Our project aims to further explore the relationship between sleep stages and OSA by using polysomnography data made available from previous studies, which includes EEG (electroencephalogram), ECG (electrocardiogram), EOG (electrooculography), and EMG (electromyography) signals gathered from a mixture of healthy individuals and individuals with OSA.

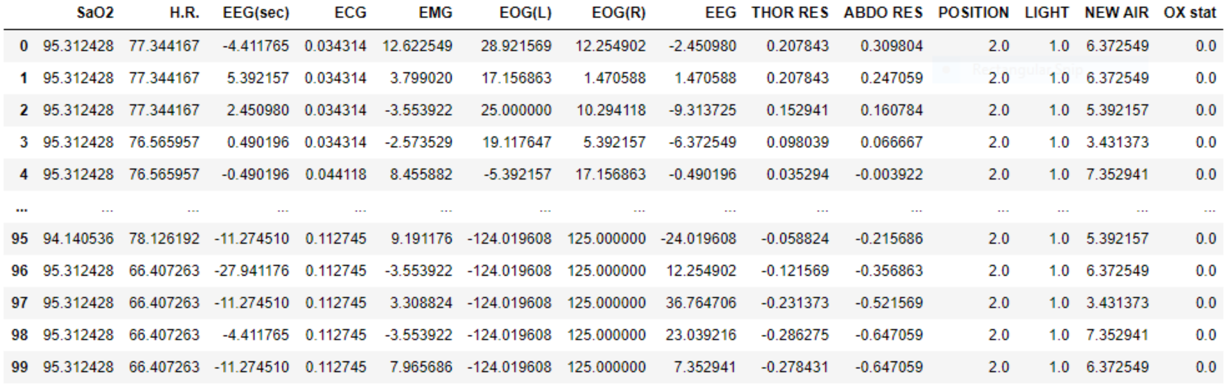
Data

The data we have chosen to work with comes from the National Sleep Research Resource (NSRR). Specifically, we used the Sleep Heart Health Study (SHHS) which consists of two visits and the respective overnight polysomnography data which "records your brain waves, the oxygen level in your blood, heart rate and breathing, as well as eye and leg movements" [4]. Since the second visit has less participants and was done more recently, to make sure we had a more reliable and complete dataset, we continued with the second visit (SHHS 2). This visit consists of 3,295 subjects along with their demographics.

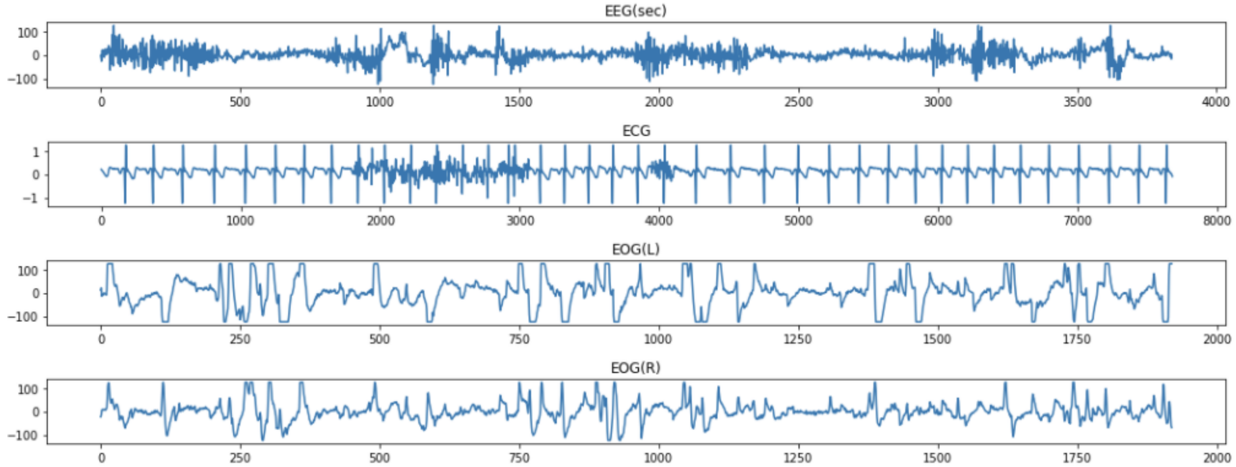


	overall	age	hypertension	bmi	visitnumber	ahi	gender	race	male
subj									
200077	5.0	46	0	23.388687	2	9.738220	1	White	1
200078	5.0	59	1	30.211833	2	19.685039	1	White	1
200079	6.0	61	0	35.451050	2	26.000000	2	Other	0
200080	5.0	59	0	32.645673	2	12.450000	1	White	1
200081	6.0	45	0	31.644286	2	2.632794	2	White	0
...	...	...	...	...	...	...	...	...	...
205798	5.0	64	1	30.804282	2	13.350000	1	White	1
205799	NaN	59	0	33.059629	2	NaN	2	White	0
205800	6.0	71	1	26.418929	2	56.115108	1	White	1
205801	NaN	59	1	26.213843	2	NaN	1	White	1
205802	NaN	60	0	NaN	2	NaN	1	White	1

Each individual has a 9-hour recording of EEG, EOG, EMG, and ECG signals in 30 second periods. The image below represents the raw polysomnography recordings and the following image graphs the signal recordings over time.



	SaO2	H.R.	EEG(sec)	ECG	EMG	EOG(L)	EOG(R)	EEG	THOR RES	ABDO RES	POSITION	LIGHT	NEW AIR	OX stat
0	95.312428	77.344167	-4.411765	0.034314	12.622549	28.921569	12.254902	-2.450980	0.207843	0.309804	2.0	1.0	6.372549	0.0
1	95.312428	77.344167	5.392157	0.034314	3.799020	17.156863	1.470588	1.470588	0.207843	0.247059	2.0	1.0	6.372549	0.0
2	95.312428	77.344167	2.450980	0.034314	-3.553922	25.000000	10.294118	-9.313725	0.152941	0.160784	2.0	1.0	5.392157	0.0
3	95.312428	76.565957	0.490196	0.034314	-2.573529	19.117647	5.392157	-6.372549	0.098039	0.066667	2.0	1.0	3.431373	0.0
4	95.312428	76.565957	-0.490196	0.044118	8.455882	-5.392157	17.156863	-0.490196	0.035294	-0.003922	2.0	1.0	7.352941	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	94.140536	78.126192	-11.274510	0.112745	9.191176	-124.019608	125.000000	-24.019608	-0.058824	-0.215686	2.0	1.0	5.392157	0.0
96	95.312428	66.407263	-27.941176	0.112745	-3.553922	-124.019608	125.000000	12.254902	-0.121569	-0.356863	2.0	1.0	6.372549	0.0
97	95.312428	66.407263	-11.274510	0.112745	3.308824	-124.019608	125.000000	36.764706	-0.231373	-0.521569	2.0	1.0	3.431373	0.0
98	95.312428	66.407263	-4.411765	0.112745	-3.553922	-124.019608	125.000000	23.039216	-0.286275	-0.647059	2.0	1.0	7.352941	0.0
99	95.312428	66.407263	-11.274510	0.112745	7.965686	-124.019608	125.000000	7.352941	-0.278431	-0.647059	2.0	1.0	6.372549	0.0



Feature Extraction

A) ECG, EOG, EMG signals and demographics Feature Extraction

In order to extract features from EEG, EMG, EOG signals and the demographics of individuals, we followed the preprocessing and feature extraction process found in the YASA Github, which is an open source and free Python library. For each night, the models extract a single central EEG, left EOG, chin EMG, then downsamples these signals to 100 Hz and band-pass filtered them between 0.40 Hz and 0.30 Hz.

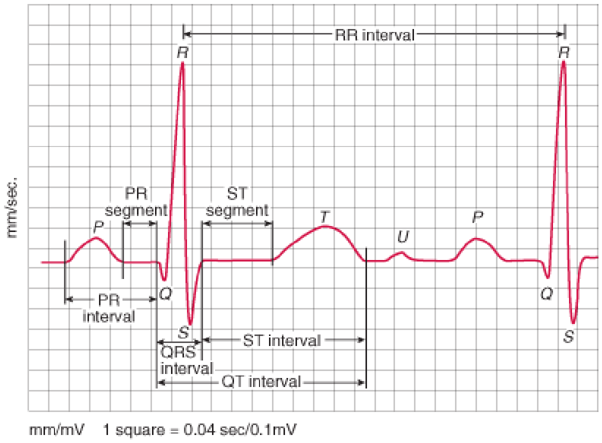
From these signals, features were extracted in two forms: time-domain and frequency-domain. Time-domain features comprise standard descriptive statistics such as standard deviation, interquartile range, skewness, and kurtosis of the signal. Frequency-domain features were extracted from the periodogram of the signal including the relative spectral power in bands, the absolute power of band signals, and power ratios.

While human experts take consecutive epochs into consideration in the scoring sleep process, common machine learning algorithms today rely only on each individual epoch. To seek the contextual temporal information, the YASA model adopts a smoothing approach using 2 rolling windows: “a 7.5 minutes centered and triangular-weighted rolling average and a rolling average of the last 2 minutes prior to the current epoch” [7]. This rolling average method overcomes the previous dilemma by incorporating the past signal information and makes the feature extraction results more reliable.

ECG signal Feature Extraction

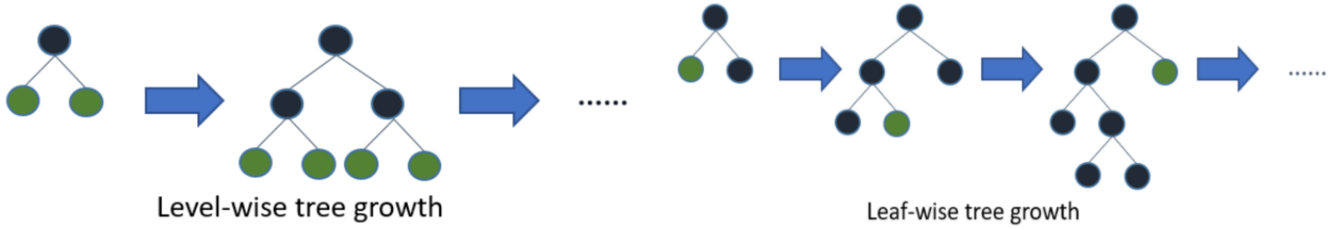
The library we used to extract ECG features is SleepECG. It provides tools for sleep stage classification when EEG signals aren't available in the dataset, which is still the most common way to determine sleep stage. Because this library is half finished and we already have read the data and done the data cleaning part in the previous part, we mainly just use the feature extraction from the module to featurize our ECG data.

The ECG feature extraction process is based on the heart rate variability features. The SleepECG module provides 26 time domain features and 7 frequency domain features. All the time domain features are derived from the normal to normal (NN) intervals, successive differences between NN intervals, and Poincare plot. As the figure shows on the right, the normal to normal interval is the intervals between normal R-peaks. The difference between the NN interval and RR interval is subtle; and an abnormal R-peaks can occur when there is “the presence of recurring arrhythmic events (also known as cardiac dysrhythmia or irregular heartbeats), as well as erroneous beat detection due to low signal quality” [5]. In a time-series, the Poincare plot depicts the correlation between two consecutive data points. In our case, it is a certain NN interval on the *x*-axis versus NN(*n* + 1) (the succeeding NN interval) on the *y*-axis. Derived from the three measures, SleepECG allows us to extract time domain features such as the average, maximum, minimum of NN intervals. SleepECG also provides a method to extract frequency domain features by using Welch’s method to power spectral density [8]. Some examples of ECG frequency-domain features are the variance of NN intervals over the temporal segment, power in very low/low/high frequency range etc.



Model

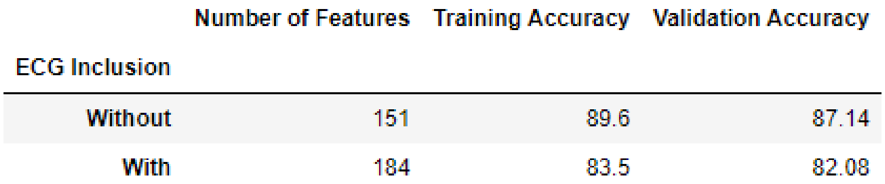
The LGBM algorithm is a fast, distributed, high performance gradient boosted framework based on decision tree algorithms [6]. It is a relatively new algorithm that is challenging and outperforming existing algorithms. Most decision tree algorithms grow by level-wise. That is, they grow horizontally. However, LGBM is unique because it grows leaf-wise or vertically. It will choose the leaf with the max delta loss to grow which can reduce more loss. The visual below helps to explain the difference in growth:



The advantages of this algorithm is that it has fast training speed, higher efficiency, lower memory usage, better accuracy, support of GPU learning, and handles large-scale data. On the other hand, the disadvantages are that it is sensitive to overfitting and doesn't work well with small datasets.

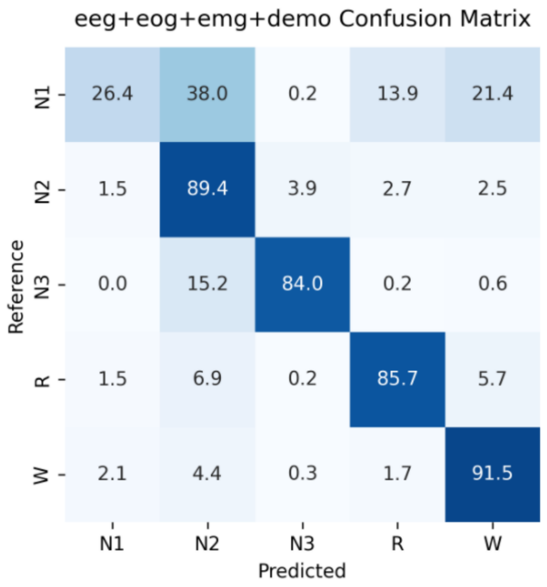
Results

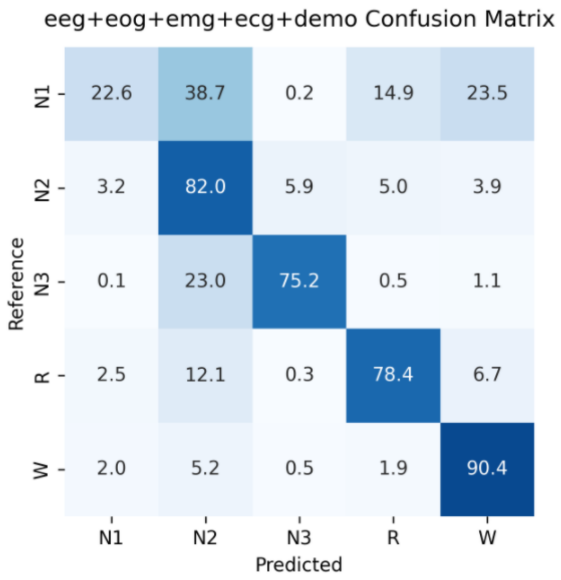
Our first model without ECG data featurized had a total of 151 features while our second model with ECG data had a total 184 features. Our training and validation accuracy in the table below shows that the first model without ECG performed better than the model with ECG features.



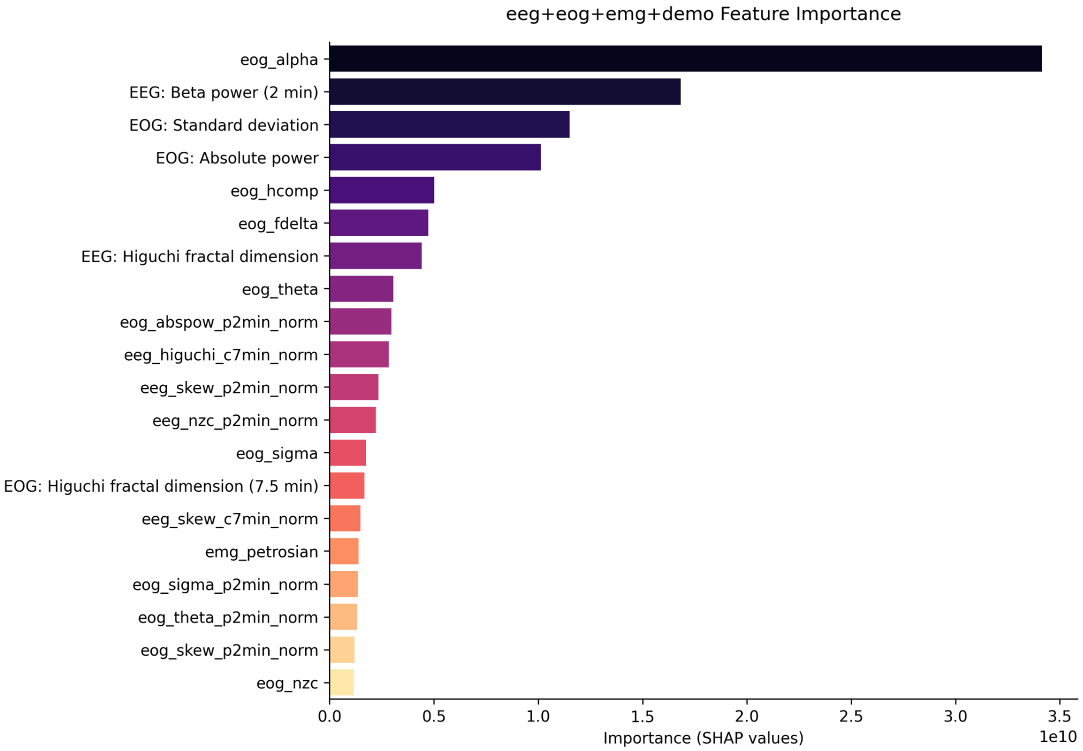
	Number of Features	Training Accuracy	Validation Accuracy
ECG Inclusion			
Without	151	89.6	87.14
With	184	83.5	82.08

Furthermore, a confusion matrix allows us to look more closely into the accuracies within each sleep stage. It shows us the accuracy between the predicted sleep stage and the actual sleep stage. Below, the left matrix represents the first model without ECG signals, and the right matrix represents the model with ECG signals. The scores show that both models performed fairly well for N2 and W (Wake) sleep stage classification. However, it is clear that the model with ECG features performed worse than the model without ECG features particularly in stages N3 and R (REM). It's also important to note that both models performed poorly when classifying sleep stage N1. This is a known issue present in previous sleep stage classification models, and could be a possible future work to discover or improve in the model.

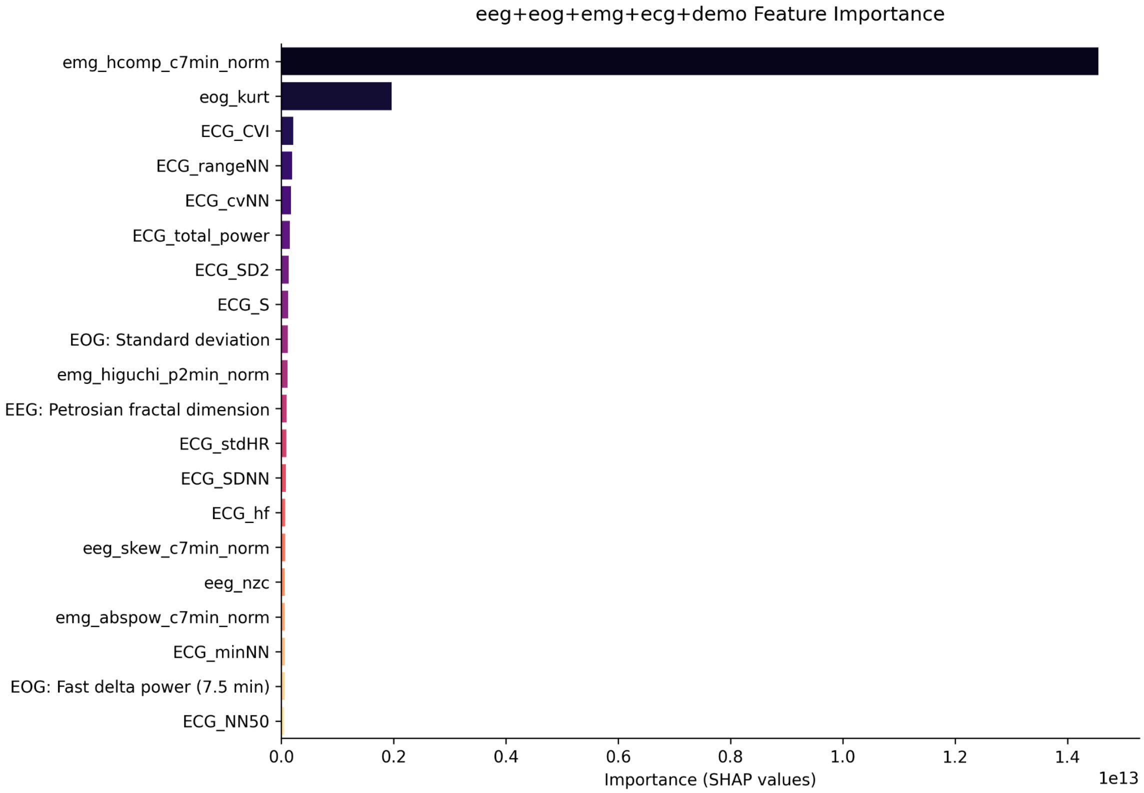




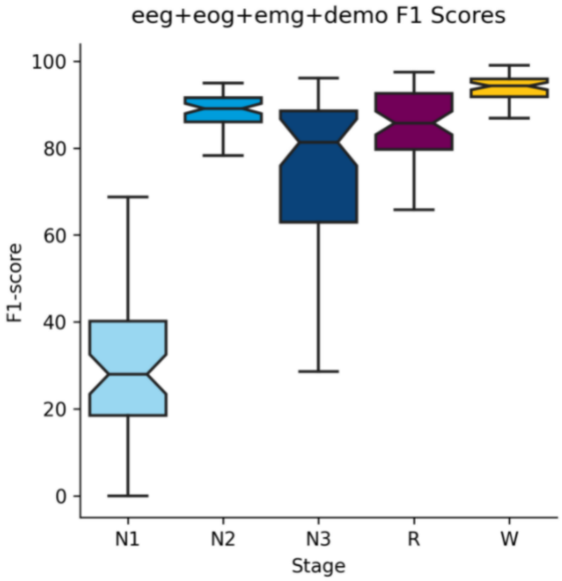
Next, these graphs show the feature importance for the models. This first graph below is for the model without ECG data featurized. It is clear that the most important features come from EOG and EEG signals. It is interesting to note that while time elapsed was an important feature for the existing YASA model, it is not nearly as important in our model (ranked 51 in importance). For normal patients during a night of sleep, the periods of REM sleep towards the beginning are shorter and periods toward the end are longer. This suggests that how long the patient has been asleep affects their current sleep state; hence, its feature importance in the previous model. Since our models only deal with sleep apnea patients, it is possible that because their sleep is frequently interrupted, their sleep doesn't develop over time in the same way that healthy sleepers do. This would explain why the total time elapsed lacks importance in our models.

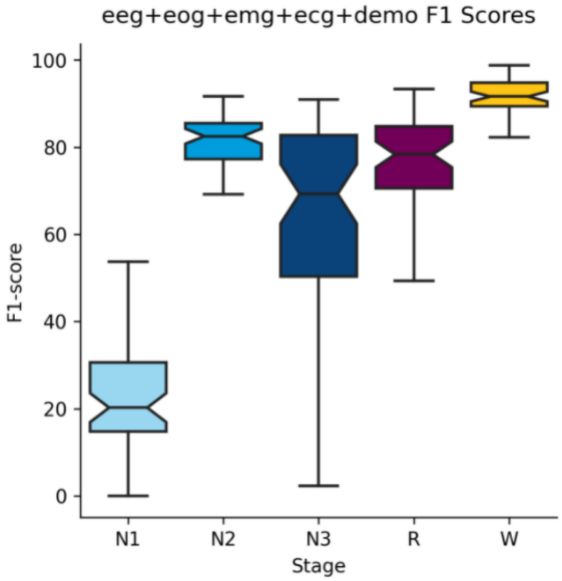


This next graph is for the model that includes ECG features. Although features extracted from ECG data are in the top 3 most important here, its importance is relatively low compared to the first two features which come from EOG and EMG data. Interestingly, the second model seems to rely heavily on one or two features, which can explain its decrease in performance. This may have been caused by the fact that we added too many new features, which increased the search space for our model. As a result, our model only found a very small subset of features to focus on in order to make predictions.



The last figure below shows the F1-scores by sleep stage of both models. The F1-score is the harmonic mean of the precision and recall of a classifier, which allows us to compare the performance of these models. The left plot is for the model without ECG data and the right plot is for the model with ECG data. Once again, the figure shows that both models struggle with N1 sleep stage classification. Noticeably, the model with ECG data struggles to classify the N3 sleep stage much more than the original model without ECG data.





Conclusion

To our surprise, the inclusion of features extracted from ECG data did not improve the model's performance. The model with ECG features performed roughly 5-6% worse in training and validation accuracy compared to the original model without ECG features. This means that our featurization of ECG data did not help the model classify sleep stages for patients with sleep apnea. Even though our findings did not match our hypothesis, our work was still important for future research within the area of sleep analysis or sleep stage classifying. Our findings may help direct others to exclude ECG signals in their models or dive deeper into the reasoning behind why ECG signals may not work so well. Furthermore, a potential future research could be improving N1 sleep stage classification as we mentioned earlier.

Limitations and Discussion

Feature Extraction: While YASA implements a temporal time window to incorporate contextual temporal information, the SleepECG which is the library we used to extract the ECG features does adopt the rolling average method. The feature extraction of ECG still follows an old-school approach by focusing on what happened in each epoch and treating each epoch independently. In further studies, the temporal time window and the rolling average can be applied to extract ECG features, which make them more informative and reliable.

Model Selection: Different models can be implemented in further studies. In our project, we choose to use the LGBM algorithm, which is based on decision trees and supports the computation of large amounts of input data. Our project provides a baseline for model comparison, and further models or deep learning algorithms might improve the performance of classifying sleep stages even more.

References

[1] Berry, R. B., Budhiraja, R., Gottlieb, D. J., Gozal, D., Iber, C., Kapur, V. K., Marcus, C. L., Mehra, R., Parthasarathy, S., Quan, S. F., Redline, S., Strohl, K. P., Davidson Ward, S. L., Tangredi, M. M., & American Academy of Sleep Medicine (2012). Rules for scoring respiratory events in sleep: update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events. Deliberations of the Sleep Apnea Definitions Task Force of the American Academy of Sleep Medicine. Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine, 8(5), 597–619. https://doi.org/10.5664/jcsm.2172

[2] Rosenberg, R. S., & Van Hout, S. (2013). The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine, 9(1), 81–87. https://doi.org/10.5664/jcsm.2350

[3] Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P. L., Favaro, P., Roth, C., Bargiotas, P., Bassetti, C. L., & Faraci, F. D. (2019). Automated sleep scoring: A review of the latest approaches. Sleep medicine reviews, 48, 101204. https://doi.org/10.1016/j.smrv.2019.07.007

[4] “Polysomnography (Sleep Study).” *Mayo Clinic*, Mayo Foundation for Medical Education and Research, 1 Dec. 2020, https://www.mayoclinic.org/tests-procedures/polysomnography/about/pac-20394877#:~:text=Polysomnography%2C%20also%20called%20a%20sleep,leg%20movements%20during%20the%20study

[5] Citi, Luca, et al. “A Real-Time Automated Point-Process Method ... - Researchgate.” *ResearchGate*, https://www.researchgate.net/publication/230638325_A_Real-Time_Automated_Point-Process_Method_for_the_Detection_and_Correction_of_Erroneous_and_Ectopic_Heartbeats.

[6] prashant111. “LightGBM Classifier in Python.” *Kaggle*, Kaggle, 21 July 2020, https://www.kaggle.com/prashant111/lightgbm-classifier-in-python.

[7] Vallat, R., & Walker, M. P. (2021). An open-source, high-performance tool for automated sleep staging. Elife, 10. doi: https://doi.org/10.7554/eLife.70092

[8] Hofer, Florian, and Clemens Brunner. “Feature Extraction¶.” *Feature Extraction - SleepECG Documentation*, https://sleepecg.readthedocs.io/en/stable/feature_extraction.html.