CSE 6250 Big Data for Healthcare Project Final Report: Learning from Sleep Data (Fall 2020)

[1]Tim Wilcox, Candidate - M.S. Analytics
[1]Georgia Institute of Technology, Atlanta, GA, United States

Abstract

The Sleep Heart Health Study collected the sleep patterns along with long term diagnosis of cardiovascular disease for almost 6,000 participants. The process for evaluating the quality of sleep and, by extension, using it to diagnose cardiovascular issues, is very time-intensive and prone to error because it requires expert analysis conducted by a technician or physician. In this project, I seek a time-saving method whereby the presence or onset of cardiovascular disease is derived directly from the polysomnographic data, instead of through human analysis. Beginning with the EEG, ECG, and EEG time series wave forms from the polysomnogram, the power spectral density for each waveform is calculated and all three are concatenate to create a feature vector for every patient. The feature vector is fed to a logistic regression classifier to train. The accuracy and receiver operating curve area under the curve score is calculated on unseen data. The results indicate that after re-weighting each class the model is able to correctly classify CVD positive patients at a rate of 81%*, however there are a significant amount of Type I errors. Finally, a discussion about possible applications for this model and avenues for future work are presented.*

Code & Presentation

The source code for this project is available at: `https://github.gatech.edu/twilcox8/CSE_6250`

Video presentation is available at `https://www.youtube.com/watch?v=xbVMVc72gEg`

Introduction

Getting a good night's sleep is important to many aspects of human health. Poor sleep has been linked to medical conditions such as depression, lower levels of cognition, and many other maladies. There are also many studies that examine the relationship between sleep and cardiovascular health[6,13,16]. According to the National Institutes of Health, Cardiovascular disease (CVD) is the number one cause of death and debilitation in America today[2].

The process for evaluating an individual's own propensity for medical issues which are related to sleep is somewhat complicated[16]. To receive a comprehensive clinical assessment of an individual's sleep patterns, the patient must undergo a polysomnogram (PSG)[6]. During the procedure, several sensors are connected to the patient. Electroencephalographs (EEGs), electrocardiographs (ECGs/EKGs), electrooculographs (EOGs), electromyographs (EMGs), measure electrical signals that are generated in the brain, heart, eyes, and muscles, respectively. In addition to the above sensors, PSGs also routinely include pulse oximetry sensors, temperature sensors, a heart rate monitor, and even a sound sensor to detect snoring or other sleep anomalies. After PSG, the data from all the sensors is divided into 30-second chunks called ”epochs” and a physician or trained sleep analyst annotates the epochs to determine the phase of sleep the individual was in. Apart from “awake”, there are four possible stages of sleep: N1, N2, N3, and REM (rapid-eye movement). The analyst scores the epoch as awake or one of the four sleep stages. This information is then shared with clinicians for use in providing medical diagnoses.

This project sought to examine whether there was a reproducible way to shortcut the process following the data acquisition to the CVD diagnoses. If a clinical diagnosis of elevated risk for CVD could be lifted directly from the sensor data, it would remove the need for the sleep to be scored for that diagnosis. The initial focus was on performing time-series feature extraction on the raw wave forms. The methodology then shifted to spectral density analysis of the wave forms.

Dataset

The data in this project comes from a dataset called the Sleep Heart Health Study (SHHS)[9,15]. SHHS was a multi-cohort study capturing unattended (at-home) PSG data for over $5,800$ participants across 5 different cohorts. The first dataset was captured between 1995 and 1998. There was a follow-up PSG for some of the participants taken between 2001 and 2003. Among many other factors, the study captured long-term CVD outcomes, as recently as 2011.

The research herein focus specifically on EEG, ECG, and EMG wave-forms for each patient. These three signals were found to be the most consistent in terms of rate of appearance and uniform sampling rates. I also focused solely on the CVD outcomes, which were available for 5,037 patients who took part in SHHS1. Of the patients studied in SHHS1, $1,195$ (approximately 76%) of them have a negative CVD diagnosis.

This project found that in a number of cases the stated sampling rate from the study differed from the sampling rate captured in the PSG data file. Because the sampling rate has a direct effect on the power spectral density for a waveform, it was necessary to correct this. This was accomplished by down-sampling to the correct rate. It was discovered that there were several instances where CVD outcomes were not captured for a patient. These data were removed entirely.

Approach

The PSG data were stored in European Data Format (EDF) files, one per sleep study patient. In order to process the data in Apache Spark it was necessary to to create a custom program to parse each file. The EDF files begin with an ASCII header that describes the subject and lists the the name and number of signals which are contained in the file along with the sample rates for each. Every signal has an ASCII signal header that describes the analog and digital maxima and minima and other peculiarities of the data (such as whether scaling was used). Finally, the data itself is encoded in 16-bit signed integer pairs (two's complement) in little-endian format.

At the outset of the project, the proposed method of analyzing the sleep data was to process the multi-sensory time-series waveforms into feature vectors using open-source packages such as tsfresh[3]. However the sheer size of each waveform made the processing time for analysis too computationally expensive to continue attempting this method. Therefore, an alternative procedure was explored involving using fast-Fourier transforms and Welch's method[12] to obtain a power spectral density (PSD) for each waveform. The application of PSD estimates (using modified periodograms) for biomedical data was first explored by Yoganathan et alia[14]. For each sleep study participant, the PSDs of EEG, ECG, and EMG waveforms were computed. The PSDs are independent of time and represent the distribution density of power into frequency components for each signal. Welch's method is computed as follows:

Given a signal vector with length N: $X(j), j \in 0, 1, ..., N-1$, take k segments (which may overlap) of length L, with the segments starting D indexes from one another.

Let

$$X_1(j), \qquad j = 0, ...L-1$$

be the first segment. Then $\forall j = 0, ..., L-1$,

$$X_1(j) = X(j),$$
$$X_2(j) = X(j+D),$$

and finally

$$X_k(j) = X(j + (k-1)D).$$

Now suppose that all K segments span the width of the entire signal such that $(k-1)D + L = N$. For each segment find the modified periodogram for that segment using Godfrey and Tukey's method[1]. This method calculates multiple discrete Fast Fourier Transforms (DFFTs) by selecting the data window $W(j), \forall j = 0, ..., L-1$ and the data $X_1, X_2, ..., X_k$ and form sequences $X_1(j)W(j), ..., X_k(j)W(j)$ using Tukey's tapered cosine window function

$$W(j) = \begin{cases} \frac{1}{2}\left(1 + cos\left(\frac{2\pi}{r}\left[j - \frac{r}{2}\right]\right)\right) & 0 \leq j < \frac{r}{2} \\ 1 & \frac{r}{2} \leq j < 1 - \frac{r}{2} \\ \frac{1}{2}\left(1 + cos\left(\frac{2\pi}{r}\left[j - 1 + \frac{r}{2}\right]\right)\right) & 1 - \frac{r}{2} \leq j \leq 1 \end{cases} \tag{1}$$

Where r is the ratio of the cosine-tapered section length to the entire window length. For example, if $r = 0.5$, then Equation 1 evaluates to a Tukey window where half of the entire window length consists of segments of a phase-shifted cosine with period $2r = 1$. A window size of 250 was chosen for this analysis. This window represents 1 second of data – not coincidentally, this is equal to the sampling rate for each of the chosen 3 sensors. The proportion of tapered cosine was chosen as $r = 0.25$.

Taking the discrete time Fourier transforms (DTFTs) over all K snips of signal like $A_1(n), ..., A_K(n)$like:

$$A_K(n) = \frac{1}{L}\sum_{j=0}^{L-1} X_k(j)W(j)e^{\frac{-2kijn}{L}}$$

where $i = \sqrt{-1}$, obtain K modified periodograms

$$I_k(f_n) - \frac{L}{U}|A_k(n)|^2, k = 1, 2, ..., K$$

where $f_n = \frac{n}{L} \forall n = 0, ..., \frac{L}{2}$ and $U = \frac{1}{L}\sum_{j=0}^{L-1} W^2(j)$. The power spectral density estimate given by these periodograms is the average of all periodograms over K:

$$\hat{P}(f_n) = \frac{1}{K}\sum_{k=1}^{K} I_k(f_n) \tag{2}$$

Furthe, the minimization of data artifacts due to snipping is achieved by taking an overlap of 50% for each window. That is, for all K snips of the waveform, at least half of the data in a single snip is windowed and represented in a subsequent snip.

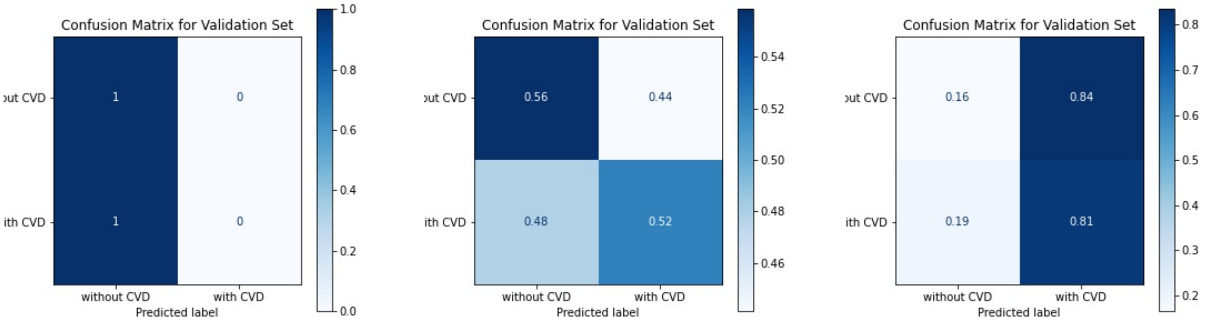
Using Welch's method and Tukey's window above, the time-series data is thus converted into an approximation of the power spectral density. The spectral density can be fed as a feature vector to a classifier. Python along with the Pandas[10] library is used to pre-process the vectors. Before classification, the feature vector is scaled by mean and unit variance (such that $z_i = \frac{x_i - \mu}{s}$ where z_i is the normalized value, x_i is the feature vector for patient i and μ and s are the mean and variance of x over all patients) using the scikit[8] learn library. The feature vector is a concatenation of the spectral densities for three sensors: EEG, ECG, and EMG, respectively. For the three waveforms chosen at a sampling rate of $250Hz$, the real element of spectral density exists between 0 and $62.5Hz$. A resolution of 250 points was selected to represent the data for each PSD estimate. Then, using a logistic regression classifier with 10-fold cross validation in order to keep the model as simple as possible but also preserve reproduceability, the results are tabulated. Graphics are produced on the results using the SciPy[11] library.

Metrics

This being a binary classification task, it was be important to understand how well the classifier correctly classified those patients who had positive CVD outcomes as well as those who had negative outcomes, but also how much the model suffered from Type I or Type II errors. For these reasons, it was chosen to report overall accuracy for general understanding, as well as the receiver operating characteristic area under the curve (ROC-AUC) score. A confusion matrix would be plotted to report the rate of correct and incorrectly classified labels for each true label.

Experimental Results

Using the scikit-learn library, the data was split into training and validation sets, with 25% of the data belonging to the latter. The classifier chosen to evaluate the data was a 10-Fold cross-validated logistic regression classifier because it is a (relatively) simple model to understand and explain. The results of this unweighted model are found in the corresponding line in table 1.



(a) no class weights

(b) balanced class weights

(c) class weights biased toward CVD positive class

Figure 1: Confusion matrices showcasing the classification rates (normalized to the true class) for three weightings of each class. No weights lead to high type II error. Biasing the model to favor CVD positive class leads to greatest accuracy but significant Type I error.

From the confusion matrix (Figure1(a)) above, the initial model suffers from high Type II error. Type II errors are errors that the model makes in classifying patients who actually go on to develop a CVD diagnosis as not having CVD. This was attributed to the class weight imbalance. Over $3/4$ of the data belong to patients that do not have a CVD outcome. The logistic regression classifier will bias the predictions in favor of the majority class. Therefore, a class weight was calculated for both classes to offset the cost of mis-classification using the formula:

$$W(j) = \frac{N}{cN_j} \tag{3}$$

where for each class j, $W(j)$ is the weight, N is the number of samples in the dataset, and N_j is the number of samples in class j and c is the number of classes. As shown in the results (1), the overall accuracy decreases, but the ability to correctly classify CVD positive patients greatly increases. Arguably, the ability to discern CVD positive patients from the larger patient pool is the most important result of the analysis. The results also report a weight selected to favor CVD positive diagnoses.

Table 1: Results of Logistic Regression Classifier on PSD estimate with different sample weights

	Training set		**Validation set**	
Class Weight	**Accuracy**	**ROC-AUC Score**	**Accuracy**	**ROC-AUC Score**
No class weights	76.60%	65.73%	75.31%	57.80%
Balanced class weights[1]	65.37%	72.12%	57.86%	55.37%
Custom class weights	41.36%	72.53%	34.76%	53.35%

[1] See Eq. 3

As shown in Figure 1(b), when using a balanced class weight, the classification rate of CVD positive patients increased to about 46%. With these class rates, the Type II error amount is reduced but not alleviated. Figure 1(c) shows the exploration of a class weighting very biased toward the CVD positive class. Using this weighting, the logistic regression model correctly classifies more than 80% of CVD positive cases. Even though this increases the amount of Type I error (patients who do not go on to develop CVD but are classified as those who do), it is still useful because the goal of the project was to classify CVD positive patients from the overall patient sample.

The reason the model isn't able to better classify CVD patients without misclassifying non CVD patients or vice versa is likely due to the homogeneity of data *across both classes* in the feature vector. Consider Figure 2, and note that the vectors for CVD positive patients look remarkably similar to the feature vectors for non CVD patients. Further, examine Figure 3, and notice that the distribution of spectral densities are near equal for all three signals. To determine whether a statistically significant difference exists across the means of both classes, a one way Analysis of Variance (ANOVA) was performed after calculating the mean of each class. The p-value for the F-statistic for this test is 0.0666, indicating that the mean values for each class *are* statistically significantly different for an $\alpha \geq 0.05$, but only just barely.

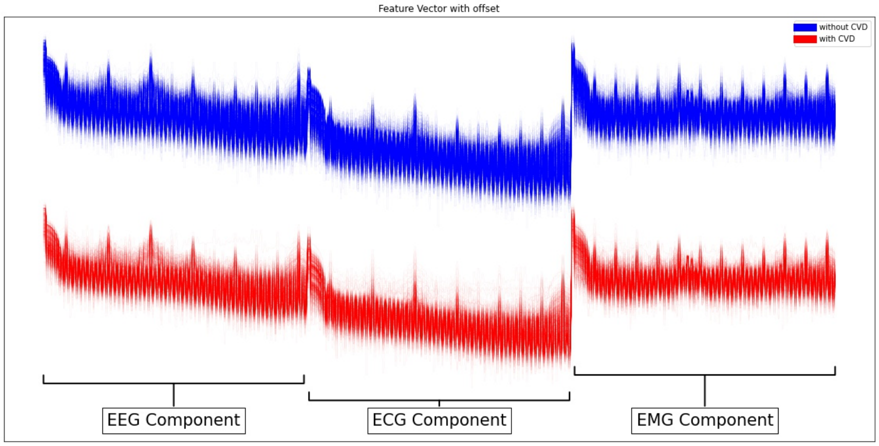


Figure 2: Composite feature vector showing patients with CVD and without CVD

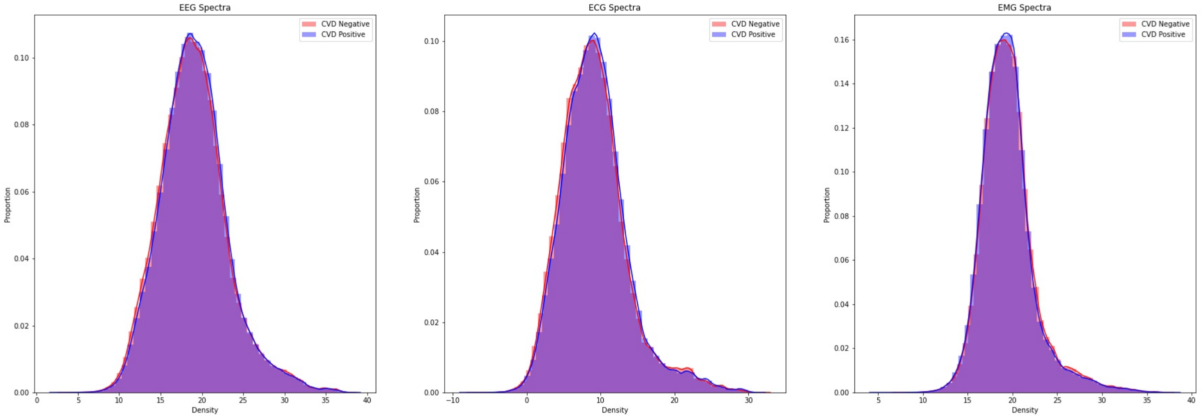


Figure 3: Spectral densities for EEG, ECG, and EMG color coded for patients with CVD and without CVD

Discussion

In this project, the initial difficulties with using a big data tool to process and analyze data were overcome and this resulted in the creation of a novel program. With this program it is possible to parse the EDF files in a custom manner that focuses on creating PSD estimates using Welch's method for each of three signals using Apache Spark (with the Breeze **signal** and **linalg** libraries doing the heavy lifting). The resultant PSDs are subsequently analyzed using Python (with packages sci-kit learn[8], NumPy, Pandas, Keras).

At the outset, the patient feature vectors were computed using open-source libraries in Python, namely pyEDFlib[5] to

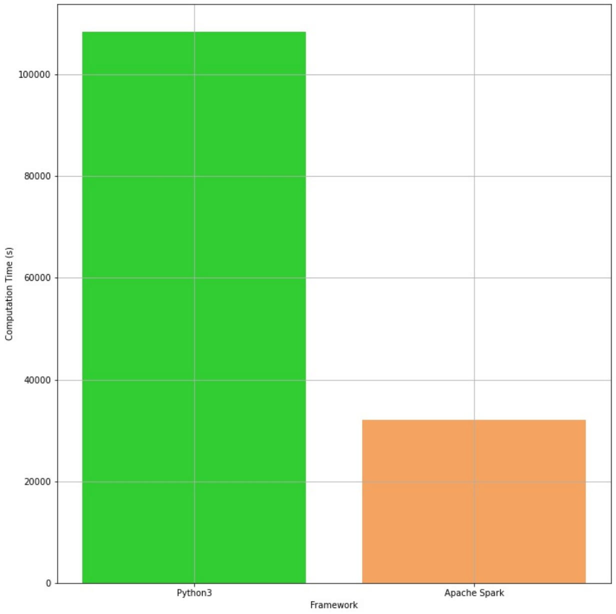


Figure 4: Time to Process 5,793 EDFs and assemble patient feature vector

extract the EDF data and SciPy to create the periodograms. As shown in Figure 4 below, this method took more than 30 hours to process all 5,793 EDF files on a local machine (Intel i9-9900X processor). The use of the custom Apache Spark program cut the processing time down to approximately 21% of the that for Python. As an added bonus, the custom program enabled significantly greater control over the final output.

The results indicate that using power spectral density estimates of EEG, ECG, and EMG time series data can provide limited insight into whether or not a patient may develop cardiovascular disease. Though the model should not be used independently to classify patients as at-risk for CVD (due to its high Type I misclassification error rates), it certainly could be used as one tool that could be part of a broader library of diagnostic tools. Future work toward this end might focus on the combination of PSD estimates and other risk factors such as smoking, hypertension, or other survey responses recorded as a part of the study. Further, PSD estimates for other sensors might be computed such as those from heart rate and pulse oximetry. Lastly, training and evaluating a more complicated model (such as a recurrent convolutional neural network) fit on this data might yield more accurate results.

Conclusion

This project began with a dataset of nearly 5,800 files containing time-series waveforms from a sleep study. Those files were converted to Welch's approximation of power spectral density for each. During the course of this effort, a novel approach to PSD creation was developed using Apache Spark. The individual PSDs were assembled into a feature vector and that was used to train a logistic regression classifier. The method evaluation sought to ascertain whether the data in the composite feature vectors were sufficient to predict outcomes of cardio-vascular disease for patients in the SHHS. The result was a model that penalizes Type II errors, and was able to correctly classify CVD positive patients at a rate of 81% on unseen data. However, this increase in sensitivity came at the expense of specificity. The overall accuracy of the model decreased significantly, and led to an increase in Type I errors. Finally a discussion of avenues for future refinement concluded that further refinement of the feature vector and a more complicated model might yield more accurate results and should be attempted.

Acknowledgments[2]

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, RFP 75N92019R002).

References

[1] Bingham, C., Godfrey, D. and Tukey, J.W. (1967) Modern techniques of power spectrum estimation. *IEEE Transactions on Audio and Electroacoustics*, AU-15(2), 70-73. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1161895

[2] National Institutes of Health (January 2020). Cardiovascular Disease. National Center for Complementary and Integrative Health. https://www.nccih.nih.gov/health/cardiovascular-disease

[3] Christ, M., Braun, N., Neuffer, J., and Kempa-Liehr, A.W. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh - A Python package). (2018). *Neurocomputing*, 307, 72-77. http://www.sciencedirect.com/science/article/pii/S0925231218304843.

[4] Harris,C.R., Millman,K.J., van der Walt,S.J., Gommers,R., Virtanen,P., Cournapeau,D., Wieser,E., Taylor,J., Berg,S., J. Smith, Kern,R., Picus,M., Hoyer,S., H. van Kerkwijk,M.H., Brett,M., Haldane,A., Fernández del Río,J., Wiebe,M., Peterson,P., Gérard-Marchant,P., Sheppard,K., Reddy,T., Weckesser,W., Abbasi,H., Gohlke, C., and Oliphant,T.E. Array programming with NumPy, Nature, 585, 357–362 (2020), DOI:10.1038/s41586-020-2649-2

[5] Nahrstaedt,H., skjerns, Kao,D.T.H., Clarke,S., Zitting,J., Ojeda,D., Miller,C., (2020, February 18). holgern/pyedflib: v0.1.17 (Version v0.1.17). Zenodo. http://doi.org/10.5281/zenodo.3673780

[6] Ibáñez, V., Silva, J., and Cauli, O. (2018). A survey on sleep assessment methods. *PeerJ*, 6, e4849. https://doi.org/10.7717/peerj.4849.

[7] Kemp B, Olivan J. European data format 'plus' (EDF+), an EDF alike standard format for the exchange of physiological data. Clin Neurophysiol. 2003 Sep;114(9):1755-61. doi: 10.1016/s1388-2457(03)00123-8. PMID: 12948806.

[8] Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V., Vanderplas,J., Passos,A., Cournapeau,D., Brucher,M., Perrot,M., Duchesnay,E. 12(85):2825-2830, 2011. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[9] Quan S.F., Howard B.V., Iber C., Kiley J.P., Nieto F.J., O'Connor G.T., Rapoport D.M., Redline S., Robbins J., Samet J.M., Wahl P.W. The Sleep Heart Health Study: design, rationale, and methods. Sleep. 1997 Dec;20(12):1077-85. PMID: 9493915.

[10] Reback,J., McKinney,W., jbrockmendel, Van den Bossche,J., Augspurger,T., Cloud,P., Gorelli,M. (2020, October 30). pandas-dev/pandas: Pandas 1.1.4 (Version v1.1.4). Zenodo. http://doi.org/10.5281/zenodo.4161697

[11] Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W., Bright,J.,van der Walt,S.J., Brett,M., Wilson,J., Millman,K.J., Mayorov,N., Nelson,A.R.J., Jones,E., Kern,R., Larson,E., Carey,CJ, Polat,I., Feng,Y., Moore,E.W., VanderPlas,J., Laxalde,D., Perktold,J., Cimrman,R., Henriksen,I., Quintero,E.A., Harris,C.R., Archibald,A.M., Ribeiro,A.H., Pedregosa,F., van Mulbregt,P., and SciPy 1.0 Contributors. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17(3), 261-272.

[12] Welch, P. D. (1967) The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. *IEEE Transactions on Audio and Electroacoustics*, AU-15(2), 70-73. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1161901.

[13] Wolk, R., Gami A.S., Garcia-Touchard A., and Somers, V.K. (2005). Sleep and Cardiovascular Disease. *Current Problems in Cardiology*, 30(12), 625-662. http://www.sciencedirect.com/science/article/pii/S0146280605001003.

[2] this section required by Sleep Heart Health Study

[14] Yoganathan, A.P., Gupta, R. Corcoran, W.H. (1976). Fast Fourier transform in the analysis of biomedical data. *Medical and Biological Engineering*. 14(2), 239-245. https://pubmed.ncbi.nlm.nih.gov/940380/

[15] Zhang G.Q., Cui L., Mueller R., Tao S., Kim M., Rueschman M., Mariani S., Mobley D., Redline S. The National Sleep Research Resource: towards a sleep data commons. J Am Med Inform Assoc. 2018 Oct 1;25(10):1351-1358. doi: 10.1093/jamia/ocy064. PMID: 29860441; PMCID: PMC6188513.

[16] Zhang, L., Fabbri, D., Upender, R., and Kent, D. (2019). Automated sleep stage scoring of the Sleep Heart Health Study using deep neural networks. Sleep, 42(11), zsz159. https://doi.org/10.1093/sleep/zsz159.

Team Contributions

I was the sole member of Team 24.