Sleep Stage Annotation and Cardiovascular Disease Risk Prediction via Random Forests and Graph Clustering

Rong Cao[1], Daniel Cazzaniga[1], Dylan Kapp[1], Blake Macnair[1]
[1]Georgia Institute of Technology, Atlanta, GA

Abstract

Sleep stage annotation models, while studied extensively on their own, have not seen much application as components of larger model pipelines. At the same time, very few studies have been dedicated to creating models to assist in predicting sleep-related health risks and disorders.

In this paper, we propose a novel machine-learning ensemble that combines sleep stage annotations and predicting cardiovascular disease risk using weighted subject clustering and extracted EEG features based on annotated data.

The sleep stage annotation classifier is trained by using the available annotated EEG data from the SHHS dataset. The health risk classifier is generated by selecting salient health risk markers given in the SHHS dataset to cluster subjects, then extracting high-dimensional features observed in sleep stage data per-cluster. Once trained, the ensemble can take in a single EEG recording and output sleep stage annotations alongside risk probabilities for select sleep-related health risks.

Introduction

Numerous studies have been performed to define automated sleep stage annotation models, with methods involving some combination of recurrent and convolutional neural networks, all achieving accuracies around the range of 80-90% [5,6,7,8,9]. Noticeably fewer studies have been performed to evaluate the efficacy of models to predict cardiovascular diseases and other sleep-related disorders [12,13,141,5].

The data used for this research was taken from Sleep Heart Health Study (SHHS). The SHHS dataset is derived from a multi-center cohort study to determine cardiovascular and related consequences of sleep-disordered breathing [18]. The data comprises two main cohorts: SHHS1 and SHHS2. 5,804 polysomnograms were recorded in SHHS1 from baseline clinical visits between 1995 and 1998. A subset of 4,080 subjects from SHHS1 were involved in recording polysomnograms in SHHS2 in follow-ups between 2001 and 2003. The dataset also provides information each subject's medical history, medication, demographics, and cardiovascular disease outcomes.

The first component of our study will use EEG data and associated sleep stage annotations to train an automated sleep stage classifier, using the accuracies from previous studies - around 80-90% - as a baseline. The second component will take biometric and medical data for subjects, alongside their sleep stage recordings, to generate health risk predictors, associating target diseases such as myocardial infarctions, coronary surgical interventions, and strokes, to distinct sleep stage patterns like abnormal CAP-rate or abnormal REM sleep. These two components will then be combined to create a single classifier that can take in EEG data from a new subject and output sleep stage annotations and health risk probabilities.

Sleep-related disorders impact millions of adults and present significant risk factors for more serious disorders and diseases, such as Parkinson's, Alzheimer's, and other neurodegenerative diseases [16,17]. By creating a novel sleep stage annotation classifier and health risk prediction classifier, we hope to enable more early-intervention health risk predictions for subjects with minimal data or resources and to enable these subjects to seek out professional assistance corresponding to the identified risks.

Experimental Setup

Sleep Stage Annotations

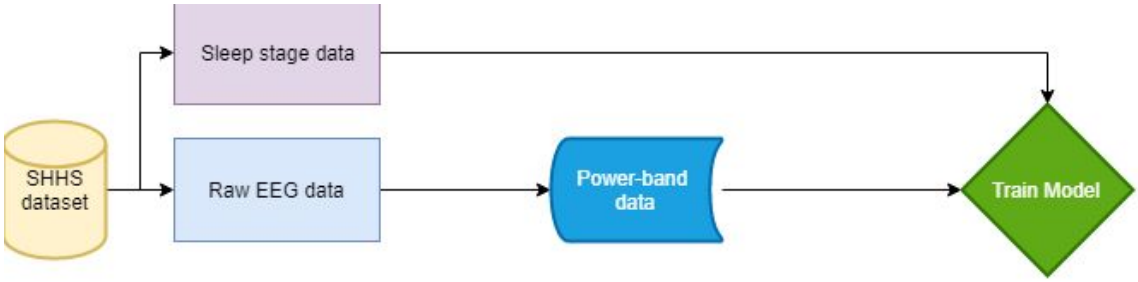

Figure 1. *Sleep stage annotation training model*

To train a classifier for labeling sequential sleep stage markers on raw EEG data, our model follows the pipeline seen in **Figure 1**. The first step takes all the raw EEG data and transforms each sample into a set of frequency power spectral densities (PSD), one for each 30-second epoch (corresponding to sleep stage markers given for every 30 seconds). The power spectral density sets are then matched with their corresponding sleep stage markers and are used to perform supervised training on our sleep stage classifier model.

For the initial model development, 2% of samples from SHHS1 were randomly selected. Among them, 50% were selected for training, 20% for validation and 30% for testing (114 subjects total , 57 for training, 22 for validation and 35 for testing). Features were constructed using Python's MNE library[11] from PSD of 6 different bands: 0.75-4.50 Hz (Delta), 4.5-8.5 (Theta), 8.5-11.5 (Alpha), 11.50-15.50 (Sigma), 15.5-30 (Beta), 30.00-50 (Gamma) from two EEG channels (EEG(sec) and EEG) in the raw data. Two slightly different approaches were used. The first used the average PSD within the band, which resulted in 2*6=12 features for each epoch; the second one used the full PSD, which had 2*6*202=2424 features. In the second approach only one channel (EEG) was kept in the initial test to reduce feature size. Once the feature of a given epoch was constructed, the features from one preceding epoch and one following epoch was added for temporal context. The resulting data was then standardized before feeding into the model.

Five different models were tested in the initial phase, MLP, CNN, RNN, LSTM and Random Forest. All neural network models used a batch size of 200 and cross entropy loss function was used for optimization. For modeling, PyTorch and scikit-learn were used. For data visualization, matplotlib was used.

Model	Model spec
MLP	4-layer MLP, with 16 hidden units
CNN	2 convolution layer, 2 pool layer and 2 fully connected layer
RNN	Many to one RNN architecture, with 3 hidden layer, hidden size 16
LSTM	A basic LSTM model with 20 LSTM cell and one linear layer for output
Random Forest	Random forest with maximum 500 trees, gini criterion, balanced class weights

Table 1: *Initial model specs*

Grid search was then carried out for both random forest and neural network models. For random forest, it was carried using RandomSearchCV function within sklearn's model selection package. The best model has n_estimators (number of trees) of 500, max depth of 80 and minimum sample split of 5. For Grid search on neural

network models, the models were first implemented in Keras and custom codes were used to test different hyperparameters like number of epochs, batch size, drop-out rates, input layer size, etc.

Serving as a benchmark, an alternative neural network model was built in Keras, using the model specifications described in towardsdatascience.com by Youness Mansar[22]. A CNN model was first used to extract the features from raw data, which was fed into a second CNN model to predict sleep stages. Note although Youness's model specification was used, the data preparation still used existing data preparation pipeline used for other models. For feature construction, both raw EEG signal and transformed PSD data were tested. The one using PSD signal has 202 features per subject and the EEG signal has 3750 features (time series).

Subject Graph Clustering

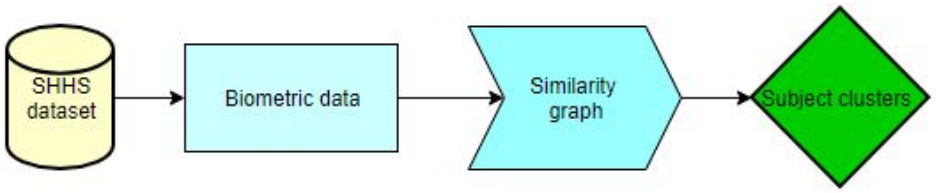

Figure 2. *Subject graph clustering*

From the raw data list of 1,991 subject variables, a salient subset of variables was chosen from subject's medical history, medication usage, and demographics. These variables specifically did not include subject's EEG measurements or attributes from these readings. These biometric and medical traits were then modeled into four graph vertex classes - Subject, Medical History, Medication, and Demographics. Bidirectional graph edges were defined as Subject-MedicalHistory, Subject-Medication, and Subject-Demographics.

Using Scala, the graph was processed using Jaccard similarity coefficient, or intersection over union, for all subjects. The Jaccard results were then passed to a Kmeans spectral clustering algorithm in python's Scikit-learn library to identify a variable number of clusters of similar subjects[23]. Kmeans clustering is an efficient algorithm designed to operate on Big Data[19]. The output of this method was the cluster ID of each subject ID.

Subject CVD Outcomes via Clustering

Once clustered, the risk of various CVD outcomes were calculated. The SHHS1 CVD summary dataset provided medical outcomes of subjects after initial participation in the study. From this, each cluster of similar subjects produced an average risk factor for each salient CVD condition. The health risks marked for analysis by the output include: 'Any Coronary Heart Disease', 'Any Cardiovascular Disease', 'Coronary Artery Bypass Graft Surgeries', 'Congestive Heart Failure', 'Myocardial Infarctions', 'Stroke', and 'Alive'. The number of each events were not considered (i.e. working with binary variables as opposed to dealing with experiencing X number of strokes).

EEG Patterns per Cluster

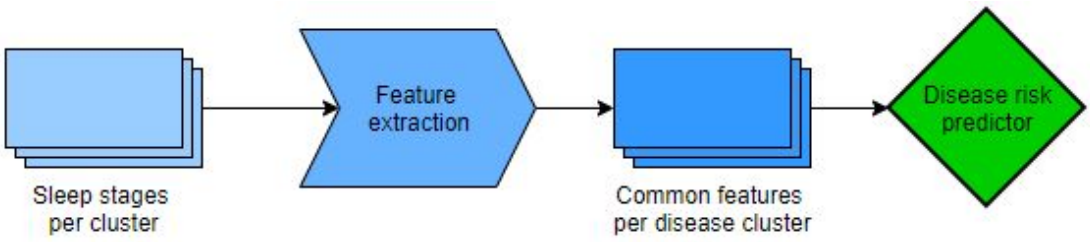

Figure 3. *EEG feature extraction per cluster*

Within each cluster, all sleep stage data will be accumulated and analyzed to extract general trends and common features (ex: cluster A having higher observed ratio of NREM CAP sleep to total CAP sleep (CAP-rate) compared to cluster B)[19]. That way, new EEG data can match each cluster to some weighted value between 0-1 where 1 is the sum of the matching correlation of all clusters.

New Subject Health Risk Factor Predictions

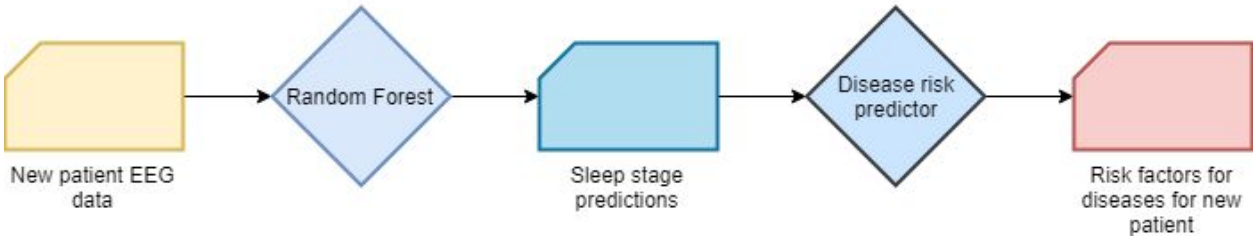

Figure 4. *Risk factor prediction for new patient EEG data*

Once a fully trained sleep stage classifier and health risk predictor has been generated, the pipeline can take a single full-night EEG recording of a new subject as input and feed it into the sleep stage classifier model to generate sleep stage annotations. The annotations are then used as input for the disease risk predictor by comparing EEG features of each subject cluster. Finalized disease risk probabilities are determined based on the average CVD metrics within each weighted cluster.

Experimental Results

Sleep stage classification

In our initial test, we have tested 5 different model specs with two feature construction method. The result is shown below

Model	Average PSD		Full PSD	
	Accuracy	F1 Score	Accuracy	F1 Score
MLP	0.73	0.75	0.73	0.73
CNN	0.75	0.76	0.77	0.78
RNN	0.72	0.74	0.72	0.74
LSTM	0.71	0.73	0.56	0.66
Random Forest	0.75	0.77	0.77	0.79

Table 2: *In initial model testing, two metrics were used, accuracy and weighted average F1 score.*

In the initial test, Random Forest labels sleep stages with the highest accuracy with CNN as a close second. It was also shown that using full PSD used significantly more features but only showed marginal improvement.

Grid search did not yield better results. Random Forest still performs best with an overall accuracy of 0.76 and F1-score of 0.77. All grid search carried out for Neural network did not improve the model.

The alternative CNN-CNN model showed slight improvement over existing random forest model in terms of overall accuracy, but they were computationally more intensive (Table 3). In a desktop machine (AMD Ryzen 7 2700

Eight-core processor, 16GB memory), the train time for a CNN-CNN model using raw EEG (3750 data points for each PPG) took more than 24 hrs and the one using PSD took less time but still significant more than random forest.

Model	Accuracy	F1-score	Speed
Random forest after Grid search	0.76	0.77	<10min
CNN-CNN use raw EEG	0.81	0.81	>24hrs
CNN-CNN use full PSD	0.77	0.78	2-3hrs

Table 3: *Final model results.*

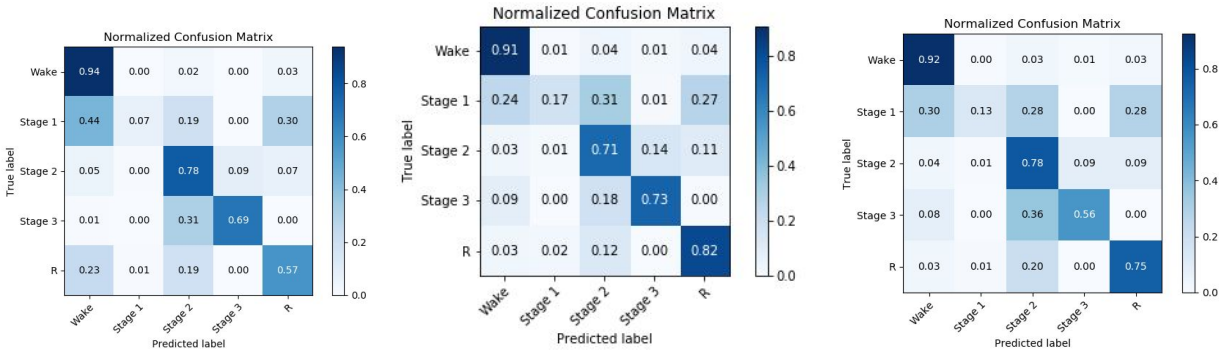

Figure 5. Confusion Matrix from **left:** Random Forest Classifier Middle: CNN-CNN using full EEG signal **right:** CNN-CNN using PSD

In the final pipeline, considering speed, overall accuracy, random forest model was used. But in the future, if further testing on CNN-CNN using PSD can result in accuracy over 80%, it can serve as a better alternative.

Subject clustering

Subjects were clustered based on their similarity on demographics, medical history and medications. 5 clusters were produced with their health risks shown below.

ClusterID	Coronary Heart Disease	Cardiovascular Disease	Coronary Artery Bypass Graft Surgeries	Congestive Heart Failure	Myocardial Infarctions	Stroke	Mortality Rate
1	1.49%	2.54%	.27%	1.35%	.87%	.68%	6.93%
2	.1%	.3%	.1%	.1%	0%	0%	16.06%
3	3.74%	4.96%	1.6%	2.82%	1.13%	.94%	21.36%
4	.36%	.36%	.12%	0%	6%	0%	4.4%
5	7.36%	8.96%	2.55%	3.78%	2.26%	1.81%	51.95%

Table 4: *Health risks of each cluster.*

Each cluster has different cardiovascular risk factors, for example, Cluster 5 has the highest probability of coronary heart disease, cardiovascular disease and second highest myocardial infarctions (heart attack). It is worth noting that this group also had the highest mortality rate. This provides some level of indication that high chances of disease have higher correlation to death. Clusters 1 and 4 have several minimal risk factors which has a logical correlation to low mortality rate.

Discussion

Our final model performance is comparable to overall accuracy by Tsinalis et.al [5]in Automatic sleep stage scoring (overall accuracy 78%) but falls behind other state-of-the-art models (overall 80%, mean 88)[21]. The final model showed variations of the misclassification rates among sleep stages. Stage 1 has the highest misclassification error with only 5% of sleep stage 1 being correctly classified, which is probably related to its low occurrence rate. In the training dataset, only 1597 out of 45043 epochs are Stage 1. Stage Wake is the most correctly classified stage with more than 90% being correctly classified. Use class weights can improve the accuracy of stage 1 but at the same time decreased the overall accuracy.

Improving the model quality turned out to be one of the biggest challenges. In the initial round of model development, to better capture the temporal information, models were tested by adding additional preceding and following epochs and use difference of preceding/following epochs and current epochs as the features. None of them yield better results. Different constructs of neural network model and hyperparameter grid search did not yield significant difference.

The biggest improvement was obtained by extracting features from raw EEG data as tested in the alternative CNN-CNN model. Using power spectral density was also a good alternative way to improve the model accuracy, as demonstrated in **Table 3**, which takes significant less time than using full EEG signal.

Conclusion

With this research, we demonstrated a big data pipeline that can predict a new subject's sleep stages using raw EEG readings and compare them with pre-trained subject clusters to predict risk probabilities for cardiovascular diseases. The pipeline provides a quick and effective way to monitor and assess a subject's CVD risk, without knowing medical history, medication or demographics.

This research could potentially assist in efforts where human annotation is not feasible or too time consuming and expensive. High accuracy means an EEG annotator would need to focus on validating smaller parts of the sleep stage classifications rather than generate the labels manually. Furthermore, by comparing a new subject's sleep data with CVD attributes of other subject clusters, preventive action could be taken. A subject who receives a high risk factor of stroke or heart attack may be inclined to visit a doctor or take medications to help fight the onset of such conditions.

The sleep stage prediction was developed using simple model based on power spectral density of EEG data. The subject clusters were built using salient medical and demographic variables leveraging modern big data tools like scala and python. In the future, we anticipate greater performance through refinement of salient subject variables, variations in cluster sizes and/or clustering methods, and enhanced neural network algorithms possibly including ensemble frameworks. Given the complexity of sleep data and its associations with cardiovascular diseases, we are confident ample pathways will continue to exist to explore risk factor classifications and predictions.

References

1. St-Onge MP, Grandner MA, Brown D, Conroy MB, Jean-Louis G, Coons M, et al. Sleep duration and quality: Impact on lifestyle behaviors and cardiometabolic health: A scientific statement from the American Heart Association. Circulation. 2016;134(18):e367-e86.
2. Merai R, Siegel C, Rakotz M, Basch P, Wright J, Wong B, et al. CDC grand rounds: a public health approach to detect and control hypertension. MMWR Morb Mortal Wkly Rep. 2016;65(45):1261-4.
3. CDC Heart Disease Facts. Available from: https://www.cdc.gov/heartdisease/facts.htm.
4. Ibanez V, Silva J, Cauli O. A survey on sleep assessment methods. PeerJ. 2018;6:e4849.
5. Tsinalis O, M. Matthews P, Guo Y, Zafeiriou S. Automatic sleep stage scoring with single-channel EEG using convolutional neural networks. arXiv:161001683. 2016.
6. Biswal S, Kulas J, Sun H, Goparaju B, Westover MB, Bianchi M, et al. SLEEPNET: automated sleep staging system via deep learning. arXiv:170708262. 2017.
7. Supratak A, Dong H, Wu C, Guo Y. DeepSleepNet: a model for automatic sleep stage scoring based on raw single-channel EEG. 2017;PP.
8. Sors A, Bonnet S, Mirek S, Vercueil L, Payen J-F. A convolutional neural network for sleep stage scoring from raw single-channel EEG. 2018;42:107-14.
9. Phan H, Andreotti F, Cooray N, Chén OY, Vos MD. SeqSleepNet: end-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2019:1.
10. TensorLy. Available from: http://tensorly.org/stable/modules/api.html#module-tensorly.decomposition.
11. MNE - MEG + EEG analysis & visualization. Available from: https://martinos.org/mne/dev/index.html.
12. Biswal S, Sun H, Goparaju B, Westover MB, Sun J, Bianchi MT. Expert-level sleep scoring with deep neural networks. J Am Med Inform Assoc. 2018;25(12):1643-1650.
13. Dreyer HC, Owen EC, Strycker LA, et al. Essential amino acid supplementation mitigates muscle atrophy after total knee arthroplasty: a randomized, double-blind, placebo-controlled trial. JBJS Open Access. 2018;(3)2:e0006. https://doi.org/10.2106/jbjs.oa.1800008128. AORN J. 2019;109(1):127-132.
14. Fulda S. Idiopathic REM sleep behavior disorder as a long-term predictor of neurodegenerative disorders. EPMA J. 2011;2(4):451–458. doi:10.1007/s13167-011-0096-8
15. Postuma RB, Gagnon JF, Rompré S, Montplaisir JY. Severity of REM atonia loss in idiopathic REM sleep behavior disorder predicts Parkinson disease. Neurology. 2010;74(3):239-44.
16. Skaer TL, Sclar DA. Economic implications of sleep disorders. Pharmacoeconomics. 2010;28(11):1015-23.
17. Pietzsch JB, Garner A, Cipriano LE, Linehan JH. An integrated health-economic analysis of diagnostic and therapeutic strategies in the treatment of moderate-to-severe obstructive sleep apnea. Sleep. 2011;34(6):695-709.
18. The National Sleep Research Resource, Sleep Heart Health Study. Available from: https://sleepdata.org/datasets/shhs
19. Lin and Cohen. Power Iteration Clustering (PIC). http://www.icml2010.org/papers/387.pdf.
20. MG Terzano, L Parrino, A Sherieri, R Chervin, S Chokroverty, C Guilleminault, M Hirshkowitz, M Mahowald, H Moldofsky, A Rosa, R Thomas, A Walters. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. Sleep Med 2001 Nov; 2(6):537-553.
21. Tsinalis O, M. Matthews P, Guo Y. Automatic Sleep Stage Scoring Using Time-Frequency Analysis and Stacked Sparse Autoencoders. Ann Biomed Eng 2016 May; 44(5):1587-1597.
22. Youness M. Sleep Stage Classification from Single Channel EEG using Convolutional Neural Networks. https://towardsdatascience.com/sleep-stage-classification-from-single-channel-eeg-using-convolutional-neural-networks-5c710d92d38e
23. Scikit-learn Python library available from: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html