Automated Sleep stage Scoring using Deep learning methods.

Richard L Albright(ralbright7@gatech.edu**), Karthick Govindaraju (**karthick1288@gatech.edu**), Jay Sumners(**jsumners3@gatech.edu**)**

Abstract

We have implemented a 6-layer convolutional neural network (CNNs) with a 1-dimensional input, three convolution/pooling combinations, and three fully-connected layers for automatic sleep stage scoring based on EEG channels Fpz-Cz and Pz-Oz. The EEG signals were split into 30-second epochs and decomposed using Welch power spectral density along a number of overlapping frequency bandwidths. The CNN was trained on the spectral density as the features of each epoch.. During initial training, we were able to achieve an overall validation accuracy of 89%; however the mean accuracy across different sleep stages were relatively lower at 66%. An imbalanced class structure favoring the "wake" stage and difficulty differentiating "wake" from "light sleep" was found and is consistent with the literature.

Introduction

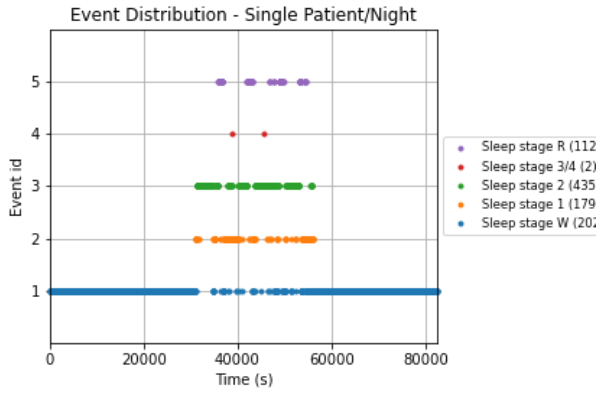
Sleep is one of the most important aspects of the wellbeing of human beings. Healthy sleep is a prerequisite for a healthy individual. Sleep disorders have been linked to 7 of the 15 leading causes of death in the U.S., including cardiovascular disease, malignant neoplasm, cerebrovascular disease, accidents, diabetes, septicemia, and hypertension [7]. Hence detecting and understanding the sleep disorders is an important step in improving the overall public health. Multiple studies in the past have proved that sleep stages can be good indicators of sleep disorders and accurately classifying the sleep stages plays a vital role in the diagnosis and treatment of these sleep disorders. Historically these sleep stage scorings have been performed by trained experts according to Rechtschaffen and Kales sleep staging criteria. Unsurprisingly, it would take experts several hours to annotate the sleep stages accurately, hence the manual process of sleep staging is very expensive. This leads to the need for automated sleep staging algorithms that have comparable accuracy to the manual scoring.

There have been several attempts to automate the sleep stage scoring process using deep learning algorithms, but they are mostly sequence-to-sequence models that attempt to score sepochs in real-time. Our project's aim was to build a post-hoc model rather than a sequence-to-sequence model that can assist human sleep-stage assignment after data collection.

Data Description

The *sleep-cassette* portion of the Sleep-EDF Database[10] was used for model development. Obtained from a 1987-1991 study, the data includes EEG signals in Fpz-Cz and Pz-Oz channels from 78 patients over two nights each. The population included only Caucasian patients not taking sleep-related medications at the time of the study.

Total Patients	Age Range Patient Counts				Sex	
	25-39	40-59	60-79	80-101	Male	Female
78	20	21	21	16	37	41

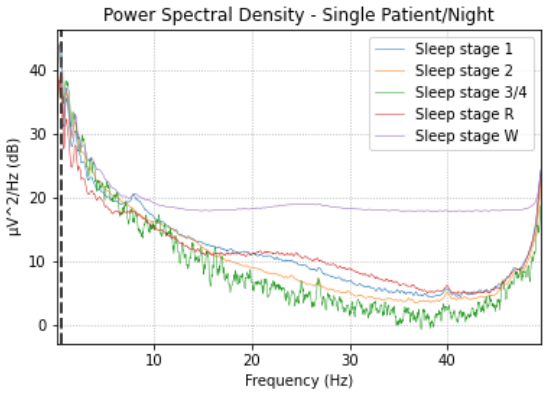


The data consisted of approximately 20 hours (split over the two nights) of EEG data and sleep-stage annotations at 30-second epochs for each patient recorded in the patient's home via cassette tape. There were a few known issues, including night 1 for two patients and night 2 for one patient. Bad epochs were dropped, although there were few. While the data was captured up to 100Hz, EEG data above 50 Hz was inconsistent across patients.

While the initial models were built without addressing the imbalance in responses, the team is currently testing solutions for the final report; such as clipping the PSG..

Total Epochs	Epochs per Night			Count of Responses (Sleep Stage)				
	Min	Mean	Max	Wake	1	2	3/4	REM
414,961	2,040	2,712	2,880	285,433	21,522	69,132	13,039	25,835

Data Pre-Processing



Power spectral density transformations were selected as a way to reduce the noise and convolution of the raw EEG signal (although some studies used the raw signal[2]). Using *pyspark* to load and transform the EEG signals in a distributed manner significantly reduced the processing time. Additionally, the *mne* package in python provided tools for loading and processing data from EDF files.

Welch transformed EEG signals clipped to 0.5-49.5 Hz (the most consistent range between patients) at 0.5 or 1.0 Hz intervals were used in initial model development. The Welch transformation was chosen since it allows, natively, mean or median averaging over the bandwidth. The expectation was that since EEG signals lead into the next epoch, that the median may provide a more predictive central tendency--being less susceptible to skewness. Additional transformations (e.g. multitaper, morlet) have also been tested or are being tested for the final report. A summarization of the ETL process follows:

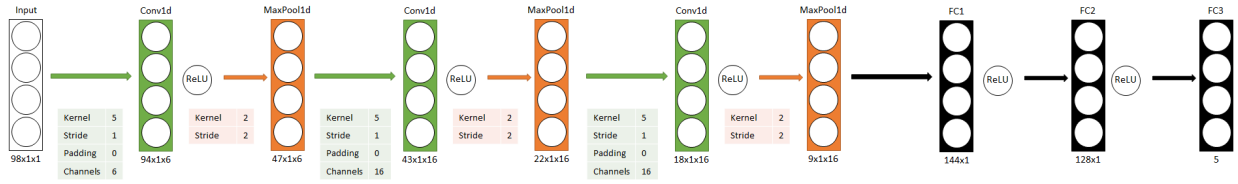
1. Group PSG and Hypnogram paths by patient and night.
2. Using *pyspark* to work distributed and the toolset from *mne*, for each pair of PSG and Hypnogram files:
    a. Load raw signal from PSG and annotation from Hypnogram.
    b. Apply annotations to raw and set channels.
    c. Extract events from annotated raw signals.
    d. Set max time (*tmax*) as 30.0 -1 / sfreq of the raw object (sets length of epoch).
    e. Convert raw to epochs using events, event ids, tmax, etc.

f. Drop bad epochs.
g. Apply indicated power spectral transformation.
h. Separate PSDs at each bandwidth.

Convolutional Neural Network

A CNN typically consists of successive convolutional layers and pooling layers followed by some number of fully connected layers. The optimization algorithm that we are using to train the network is the popular Stochastic Gradient Descent (SGD) . We used ReLU activation function between the convolutional layers due to the fact that we have 3 convolutional layers and ReLU will reduce the computational intensity without compromising the performance of the CNN. The last pooling layer is followed by 3 fully connected layers to slowly reduce the number of features.

The input data consists of a (# epochs, # frequency bins) and passed to the model in mini-batches of 32 epochs. Since we are utilizing SGD as an optimization algorithm, relatively small mini-batches may permit the model to converge faster (although SGD never truly converges). The input dataset is split into 60% training, 20% validation, and 20% testing. On larger datasets and under further development, training may take up a larger percentage of the overall data.

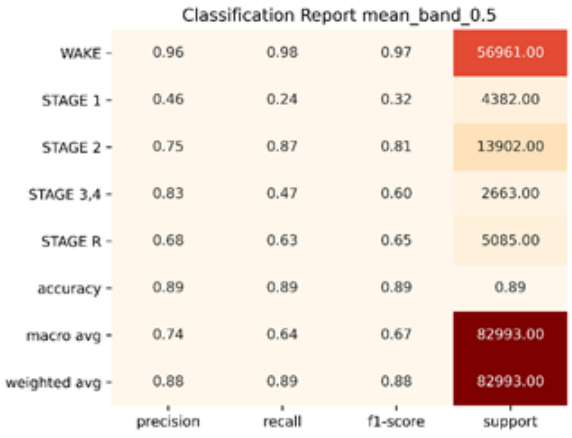


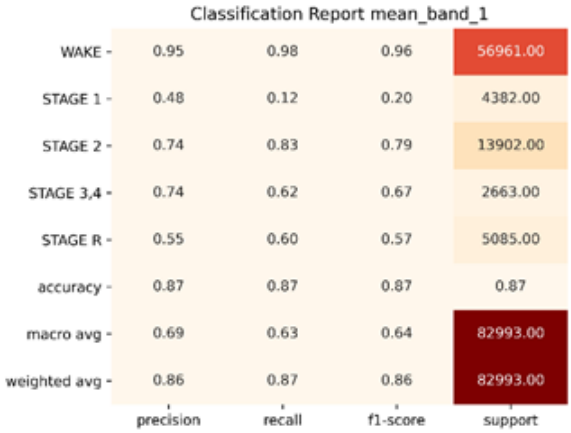
Experimental Results

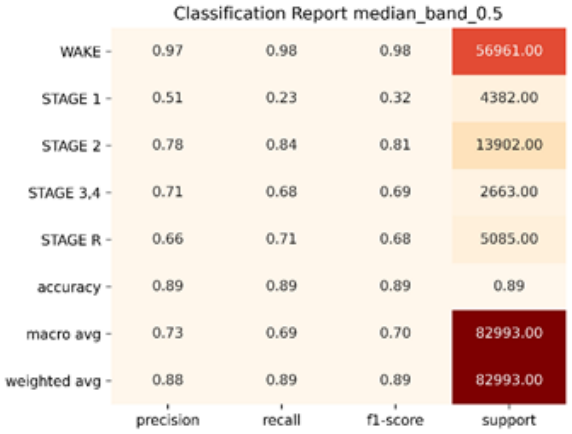
The CNN was trained with four different transformations of the raw data to determine the best transformation for this CNN architecture.

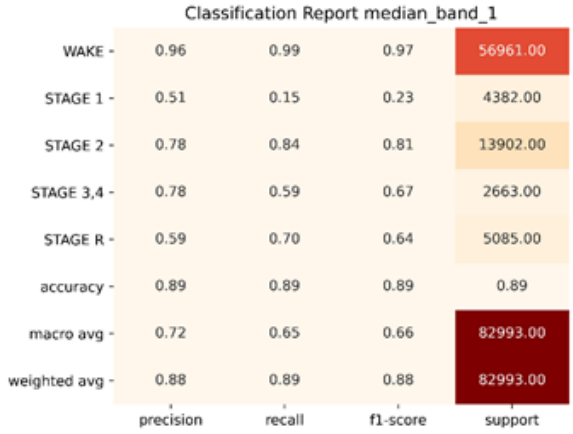
1. Welch power spectral density using the mean to average over a bandwidth
    a. Bandwidth of 0.5 Hz.
    b. Bandwidth of 1 Hz.
2. Welch power spectral density using the median to average over a bandwidth
    a. Bandwidth of 0.5 Hz.
    b. Bandwidth of 1 Hz.

We trained the first model using all of the input sleep stages despite the fact that there was a class imbalance in the training data. The results are shown below.

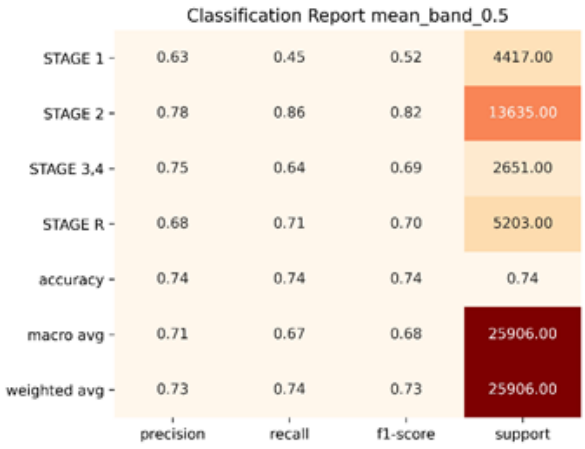


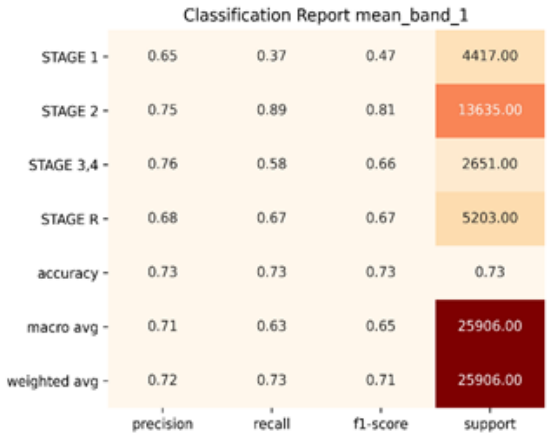


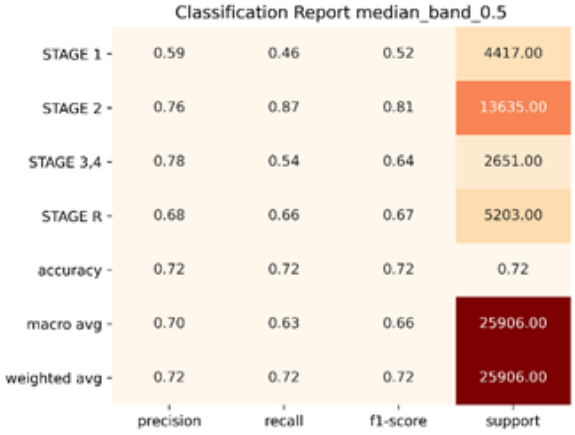


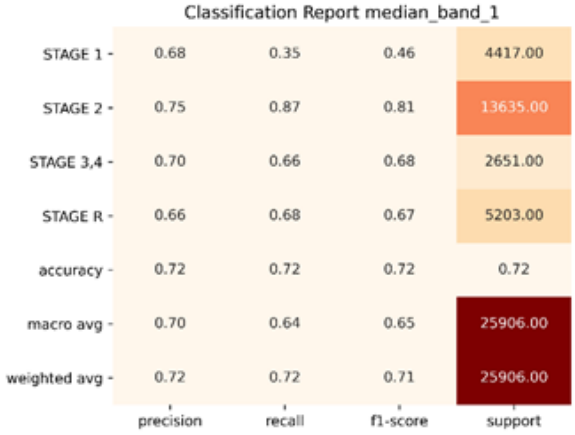


As it can be seen here, due to the imbalance in the data favoring the ‘wake’ sleep stage, the model was able to classify the “wake” stage more readily at the expense of the other classes. In another round of testing, we attempted omitting the “wake” stage to baseline the performance of the model among the other classes and obtained the following results.









Discussion

As the removal of the ‘wake’ stage significantly improved the performance of the model on the other stages, it appears that trimming the “wake” data appropriately may improve the overall accuracy of the model or, at least, improve the balance of accuracy in the model.

Conclusion/Optimization

We are at a point of fast iterations through hyperparameters and model input data structures for the final report. The findings herein have been the most telling on our journey to a final model. From our work thus far we’ve concluded that:

1. It is possible to score high levels of accuracy if “wake” stage imbalance is not corrected. Literature in the field should be read with a keen eye to this detail.
2. Spectral densities have some correlation with sleep-stages and sufficient predictive power.
3. Traditional bandwidths, such as *delta*, *theta*, and *alpha*, may be too large to provide significant information when reduced to a central tendency and that significantly small, non-traditional, bandwidth may be more appropriate for machine learning models.

The next steps include

1. Determining a consistent method for “wake” stage clipping.
2. Transforming the data with other power spectral density transformations and comparing results on a relatively simple CNN (including multitaper and morlet wave transformations)..
3. Final hyperparameter and depth tuning to optimize the final model.

References

1. O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou. Automatic sleep stage scoring with single-channel eeg using convolutional neural networks. arXiv preprint arXiv:1610.01683, 2016.
2. S. Biswal, J. Kulas, H. Sun, B. Goparaju, M. Brandon Westover, M. T. Bianchi, and J. Sun. SLEEPNET: Automated sleep staging system via deep learning. 26 July 2017
3. M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi. Learning sleep stages from radio signals: A conditional adversarial architecture. In International Conference on Machine Learning, pages 4100–4109, 2017
4. R. U. D. K. Linda Zhang, Daniel Fabbri. Automated sleep stage scoring of the sleep heart health study using deep neural networks. 2019.
5. I. Al-Hussaini, C. Xiao, M. B. Westover, and J. Sun. Sleeper: interpretable sleep staging via prototypes from expert rules. In Machine Learning for Healthcare Conference, pages 721–739, 2019.
6. O. Tsinalis, P. Matthews, and Y. Guo. Automatic Sleep Stage Scoring Using Time-Frequency Analysis and Stacked Sparse Autoencoders: Annals of Biomedical Engineering, Vol. 44, No. 5, May 2016 pp. 1587–1597
7. Chattu VK, Manzar MD, Kumary S, Burman D, Spence DW, Pandi-Perumal SR. The Global Problem of Insufficient Sleep and Its Serious Public Health Implications. *Healthcare (Basel)*. 2018;7(1):1. Published 2018 Dec 20. doi:10.3390/healthcare7010001
8. Zhang GQ, 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.
9. Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, Rapoport DM, Redline S, Robbins J, Samet JM, Wahl PW. The Sleep Heart Health Study: design, rationale, and methods. Sleep. 1997 Dec;20(12):1077-85. PMID: 9493915.
10. Goldberger, A., L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [https://physionet.org/content/sleep-edfx/1.0.0]. 101 (23), pp. e215–e220." (2000).