Sleep Staging from raw photoplethysmography signal using deep learning

CS236781: Daniel Hadromi (302632625), Kevin Kotzen (345011423)

Abstract

Accurate sleep staging is in an important step in the identification and diagnosis of sleep disorders. Sleep is known to be reflected in cardiac and respiratory signals and hence it should be feasible to stage sleep by analyzing these signals. In this work we demonstrate that fingertip photoplesmography can be used to predict sleep stage with reasonable accuracy. A dataset of over 1000 patients full overnight sleep studies is used to train a convolutional neural network to stage sleep from the photoplesmography signal. The model trained achieves a weighted accuracy of 56% which is comparable with published results.

Intro

Sleep is essential for human health, well-being and longevity [1]. Insufficient sleep and poor sleep quality are known to cause a myriad of physical and mental diseases such as cardiovascular disease, obesity and depression [1]–[3]. Sleep disorders are highly prevalent, affecting up to one sixth of the global adult population [2]. Despite the impact on quality of life and ubiquity of sleep disorders, many people with sleep disorders are unaware of their problem and remain undiagnosed [3]. Sleep disorders are traditionally diagnosed during a sleep study called Polysomnography (PSG). During a PSG study, the patient is monitored and observed overnight in a sleep laboratory. The patient is connected to devices that measure and record a number of neurophysiological and cardiorespiratory variables [4]. This process is uncomfortable for the patient, who has to spend a night out of home in a clinical environment, labor intensive, requiring a technician to monitor the patient overnight and another technician to perform manual sleep staging, and expensive with high laboratory and labor costs, making PSG an imperfect tool for sleep monitoring [5]. With the recent increase in proliferation of wearable sensors and mobile health applications there has been a rapid increase in the number of tools that aim to assess sleep quality and disorders in a more objective and frequent way, particularly targeting the monitoring of the individual in their home environment i.e. outside of the traditional clinical setting [4]–[9].

Sleep Staging: Human sleep is non-uniform in nature and consists of a number of distinct phases that are present in a cyclic nature [2]. These phases, defined by the American Academy of Sleep Medicine (AASM) [10] as Wake (W), Non-REM-1 (N1), Non-REM-2 (N2) and Non-REM-3 (N3) and Rapid Eye Movement (REM), relate to changes in the brain waves and other physiological processes [2]. A typical night's sleep and the associated stages are shown in Figure 1.

*Photoplesmography***:** Sleep stage is known to modulate respiratory and cardiovascular systems and hence measurement of these signals is expected to allow sleep staging. Photoplesmography (PPG), a measurement of the changes in the volume of blood flowing through the microvascular bed of tissue using light [11], can be used to measure cardiorespiratory signals and is therefore a

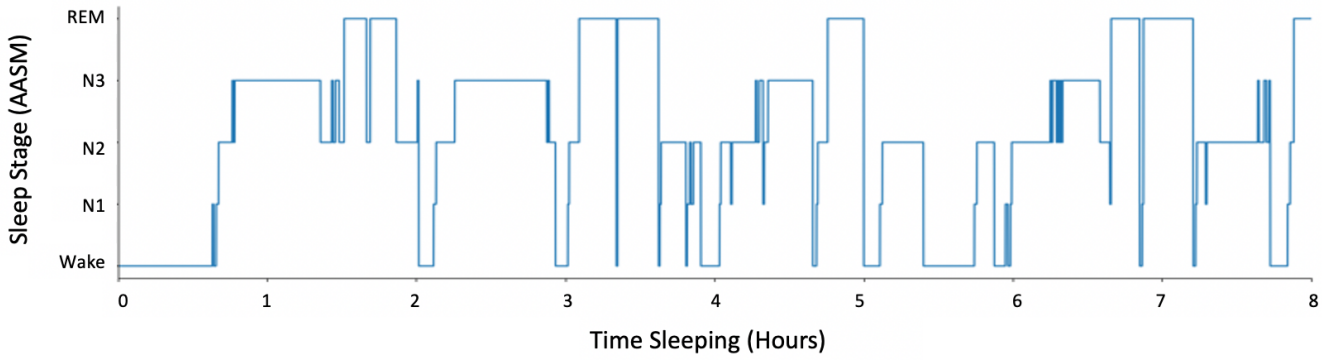

Figure 1: Somewhat typical Sleep Staging Hypnogram showing the cyclic nature of sleep over a single 8-hour period.

promising signal for sleep diagnosis. PPG sensors have seen widespread adoption and are now a standard feature in leading fitness trackers and smart watches where they are used to monitor heart rate [12].

Literature Review: A number of prior studies [13]–[16] have attempted to stage sleep from PPG using features, such as heart rate, respiration rate, or heart rate variability, engineered from the raw PPG signal with varying levels of success. A few key works are discussed below.

- In 2015 *Fonseca et al* described the process of automatic sleep stage classification with cardiorespiratory signals. For each 30s of sleep, they extract 142 features from the cardiorespiratory signals. The features are then used to train a multi-class Bayesian linear discriminator with time varying properties. This work was state-of-art for automated sleep staging at publication, achieving a 10k-fold accuracy of 69% in 4 stage sleep classification. Limitations include the use of a small dataset (48 patients), an extensive feature engineering and feature selection process requiring significant domain knowledge, and the use of a nonstandard/non-reproducible test set.
- In 2018 *Li et al* published a paper in which a time-series like signal, a cross-time frequency-domain representation cardiorespiratory coupling signal (CRC), was used for sleep staging. In this work, Li looks to classify continuous five-minute segments of PPG-derived frequency-domain images using a CNN as a feature extractor. This work claimed to achieve an accuracy of 75% on 10k fold cross validation. Limitations in this work include: the five-minute (vs thirty-second) sleep segments, the use of a very limited set of data - as only segments where a full five minutes belonging to the same sleep stage are used, an unspecified signal quality threshold and a nonstandard/non-reproducible test set.
- In an August 2020 paper, *Sridhar et al of Virily Life Sciences* (formerly Google Health) develop a state-of-art approach achieved using the instantaneous heart rate (IHR) derived from the ECG. They use a CNN used to extract features from the IHR and then use a dilated CNN to extract temporal information. While not fully explored, they mention that the same approach is theoretically possible using PPG as the input signal.
- Finally, while engineered features have shown good results, modern advances in representation learning (e.g. deep learning) combined with the availability of "big databases" suggest that yet better performances may be achieved [17], [18]. In May 2020, *Korkalainen et al* demonstrated that the raw PPG signal could be used for sleep staging. They build a model that is based on predicting sleep stages for PPG sequences of 50 minutes duration. They make use of a CNN and RNN network.

Methods

This project covers a single component of a larger MSc project. Our goal is to explore the limitations of classifying stand-alone 30s segments of PPG without the use of long-term temporal-information or memory. As such, we will not use recurrent neural networks or time dilated convolutional neural networks. Our approach is hence closest to *Li et al* who extract a five-minute CRC spectrogram from the ECG signal. We intend to improve on this work by:

1) Using the raw PPG signal which is easy to obtain from modern smart watched. The ECG signal used by Fonesca et al requires electrodes to be attached to the patient's skin.
2) Use the raw PPG signal without any engineered features. Fonesca et al extracted a CRC spectrogram from the ECG.
3) Classify 30 second segments as opposed to 5 minutes as done by Fonesca. This will increase our dataset size and create a more realistic representation of real-world data.
4) Clearly describe our train and test sets such that further works can make a meaningful comparison.

Motivation: As shown in our literature review, it is possible to carry out sleep staging from cardiorespiratory signals. We wish to see whether sleep staging can be improved using the raw PPG signal as opposed to using engineered features as was done in prior work. Motivation for this is based on the fact that many state of art approaches in machine learning have been achieved using the original signal as the input as opposed to engineered features. With regards to the PPG signal and sleep staging, it is expected that components of the PPG signal such as the morphology of each pulse, respiration modulation, heart rate variability, patient movement, and other unknown features may contain information that relate to

sleep staging. Thus, by using a CNN to extract these features we may improve results.

Dataset: We obtained access to the Multi-Ethnic Study of Atherosclerosis (MESA) Sleep database (MESA Sleep) [19], [20] following the Technion institutional review board (IRB) number 62-2019. MESA Sleep contains 16300 hours of full overnight polysomnography measured from 2237 patients aged 54-95 years old. MESA Sleep is maintained by and available from (once permission has been granted) the National Sleep Research Resource (NSRR) [19], [20]. MESA Sleep measurements include EEG, respiratory signals, electrocardiography (ECG), actigraphy and finger pulse oximeter. The PPG signal comes from the finger pulse oximeter and is sampled at 256Hz. Sleep annotations in the form of a hypnogram (labeled sleep stage per 30s epoch) are also provided in the dataset. In order to reduce computation time, we down sample PPG data to 16Hz.

Signal Quality: PPG signals are highly susceptible to noise caused by movement and hence signal quality assessment is a necessary step to remove data that is too noisy. To estimate signal quality, we use a technique similar to bSQI, a signal quality index used in ECG [21]. We compare the similarity between PPG peaks detected by two different PPG peak detectors (where one peak detector is better than the other). The correlation between these peaks, in the form of the F1-score is used as the signal quality metric. The percentage of patients whose full overnight recording meets a specified F1-score is shown in the Figure 2. Following visual evaluation of patient PPGs, we choose a signal quality threshold of 60% which resulted in the inclusion of 1378 patients in our dataset.

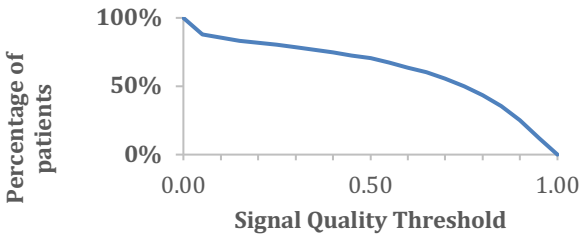

Figure 2: Signal Quality Assessment per patient

Sleep Stages: In order to simplify our problem and produce results that are comparable to other works in this field, we reduced the number of sleep stage from five to four by distinguishing only Wake, Light Sleep (N1 + N2), Deep Sleep (N3) and REM. It must be noted that the sleep stages are highly imbalanced as shown in Figure 3.

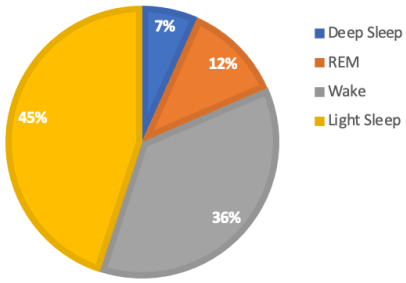

Figure 3: Pie chart depicting the breakdown of dataset into sleep stages.

Building Sequences: In order to capture longer term cardiorespiratory dynamics we use sleep segments of two and a half minutes long, but only the sleep stage for the center segment is used as the label for each segment. Segments may actually belong to multiple sleep stages. We overlap segments by two minutes as displayed in Figure 4.

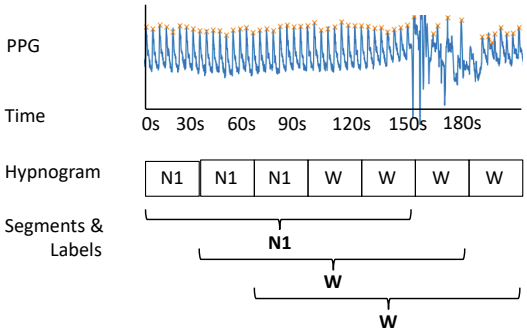

Figure 4:Diagram describing the building of 150s overlapping sequences.

Evaluation data: A train-test split will be carried out at a patient level ensuring that data from a single patient appears in either the train set or test set. Cross validation will be carried out on the train set in order to optimize hyperparameters and tune the network configuration. The last 100 patients in the MESA dataset are used as the test set.

Implementation and Experiments

Data Preloading: In line with Biomedical Engineering faculty requirements, all medical data used in this project is stored on a dedicated secure networked hard drive. Access to this drive is slow and reading data directly from this drive is a speed bottleneck during model training. In order to get around this, the data is preloaded and stored in RAM. During preloading, we optimize memory by

recasting data from 32-bit floating point to 16-bit floating point and down sampling it from 256Hz to 16Hz.

Data Loader/Generator: Batches of overlapping sequences are efficiently generated using a data loader. The data loader keeps track of a shuffled set of sequences and generates batches of the desired size on the fly. Data loading speed is maintained by using 8 CPU cores to prepare data. The data loader is built such that it can be easily extended to carry out transformations (such as Fourier Transform or Wavelet transform) in the future should this be desirable.

Deep Learning: Deep learning was implemented using Keras v2.4.3 with Tensorflow backend v2.3.0. We initially used pyTorch to implement our model, however after experiencing some issues (which we later discovered were not related to pyTorch, but rather our data loader) we rewrote our codebase and switched to Keras. All deep learning was carried on a single Nvidia RTX 2080 GPU at the Biomedical Engineering Faculty at Technion.

Loss Function Formulation: This is a simple classification problem and hence the multiclass-cross-entropy loss function is used. This is a simple average of the cross-entropy-logarithmic loss for each class. A perfect model will have a loss of 0.

Optimization Evaluation Metric: We evaluate model performance based on a modified weighted accuracy. We noticed that a main weakness of the models developed in literature was a poor minimum accuracy. While the model was good at differentiating between some classes, it struggled with others. In order to address this; we developed our own evaluation metric. The weighted accuracy is formulated in order to maximize the lowest accuracy while still maintaining a higher overall accuracy. In order to achieve this, we average the lowest accuracy with the median of all the accuracy. This results in a robust evaluation metric.

Hyperparameter Optimization: We make use of Bayesian optimization in order to find the best hyperparameters for our model. Bayesian optimization relies on gaussian processes and is an efficient means of tuning hyperparameters. We used Bayesian optimization to tune the hyperparameter shown in Table 1.

Table 1: Hyperparameter Optimization Ranges

Hyperparameter	Min	Max
Epochs	5	50
Learning Rate	1e-4	1e-3
Dropout	0	0.5
Kernel Size	3	13
Batch Size	8	128
Channels	4	7

Model Architecture: 1D CNNs are known to be effective in extracting features from time series signals. We based our initial CNN design on the feature extraction stage of *Korkalainen et al.* Korkalainen uses three blocks of Conv1D-Conv1D-Pool as shown in Figure 5. The output of this network is expected to contain features that are relevant for sleep staging.

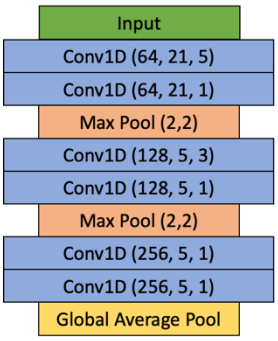

Figure 5:Starting point CNN network design

We ran experiments to see if additional network layers, different kernel sizes, different pooling strategies and normalization techniques including dropout and batch norm were effective at improving generalized model performance. The models we ran are shown in Figure 6.

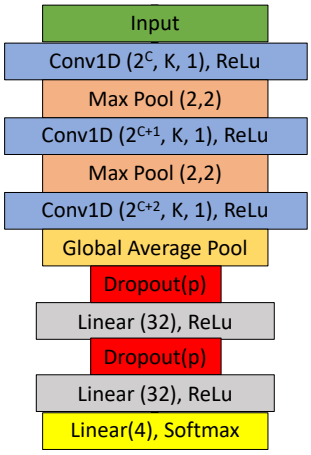

Figure 6: Model used during hyperparameter optimization. Hyperparameters include C=channels, K=Kernel Size, p=Dropout.

Results and Discussion

The results of the hyperparameter optimization are shown in Figures 7-9. Each Figure and the parameter is explained briefly.

Epochs: Increasing the number of epochs increases weighted accuracy to a given point. Thereafter, the model begins to overfit and the weighted-F1 score decreases. The optimal results were achieved with 30 epochs.

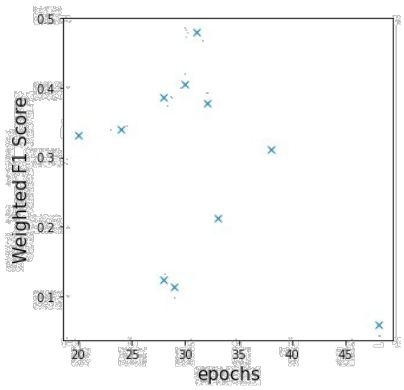

Figure 7: Effect of Epochs on Weighted F1-Score

Dropout: Dropout is a highly effective means of reducing overfitting by turning off a given percentage of random neurons during training. This forces the network to find robust features and connections. Increasing the amount of Dropout results in an increase in weighted accuracy until around 0.35 after which it degrades performance. The model with the highest weighted accuracy had a dropout of 0.375.

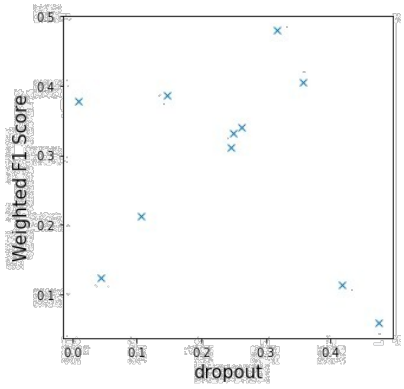

Figure 8:Effect of Dropout on Weighted F1-Score

Kernel Size: Unlike 2D-CNN design where a kernel size of 3x3 or 5x5 are standard, kernel size in 1D CNN's are highly variable and can have a major impact on performance. We used a single kernel size for all of our convolutions; however it is sometimes suggested to use a larger kernel on the first convolution to capture longer term features. The impact of kernel size on performance is not clear from the experiments we ran and should be explored further. The model achieves the highest weighted accuracy with a kernel size of 4.

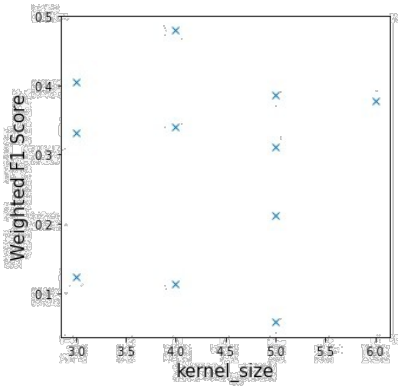

Figure 9:Effect of Kernel Size on Weighted F1-Score

Finally, we use the hyperparameters from the best model built by the Bayesian optimizer and train a network on the full training set. The model was then tested on the test set and the results are displayed in the confusion matrix shown in Figure 10. This model has a weighted accuracy of 56%.

		Ground Truth			
		Wake	**Light**	**Deep**	**REM**
Prediction	**Wake**	62%	22%	9%	7%
	Light	12%	57%	16%	16%
	Deep	7%	29%	53%	12%
	REM	6%	17%	15%	61%

Figure 10: Confusion matrix of after training on the complete training set and evaluating on the test set.

These results are within the range of those achieved by Q Li et al [9], shown in Figure 11, who extracted a feature based transformation of the dataset. It is expected that with more time and computing power the results achieved with the raw PPG can exceed those of Q Li et al.

		Ground Truth			
		Wake	**Light**	**Deep**	**REM**
Prediction	**Wake**	75%	13%	2%	10%
	Light	12%	62%	15%	11%
	Deep	3%	38%	59%	0%
	REM	18%	20%	1%	61%

Figure 11: Confusion matrix from Q Li et al.

We attempted to develop a model to predict sleep stages by the raw PPG signal without using memory in our system. The problem was more challenging than initially expected and state-of-art results could not be matched. We however gained many insights into this work and have a lot of suggestions for future work. These are listed below.

- *Data Quality:* Our PPG signals were very noisy. We removed approximately 50% of our data in order for it to meet quality requirements. Further work into the data to include in the training set should be carried out. It is even suggested to consider the signal quality threshold as a hyperparameter.
- *Data Labeling:* Interrater agreement is known to be low for sleep staging. In order to ensure that all data is labeled according to the same standards the state-of-art EEG sleep staging algorithm could be used to relabel all hypnograms in a uniform format.
- *Sampling Rates:* In order to speed up training times, we down sampled our PPG dataset to 16Hz. Any information contained in the 8Hz and above range will be lost. In other works, the researches have down sampled their data to 64Hz and 100Hz. The impact of the sample rate should be experimented on by comparing two similar models where the only variable that changes is the sample rate.
- *Feature Extraction:* As the literature review detailed, sleep staging using feature extraction from the PPG has yielded reasonable results. The state of art is currently held by Sridhar et al who use the instantaneous heart rate signal. This signal in addition to HRV metrics and transforms such as the wavelet transform should also be explored.
- *Dataset Diversity:* In this work we use a single dataset. It is suggested to increase dataset diversity by including additional datasets. The Cleveland Family Health Study and the SOMNIA datasets are potential candidates.
- *Sequence Models and Memory:* Sleep stages are highly cyclic and hence sequence models are expected to greatly improve results. For example, a person who in WAKE must progress through NREM1 to NREM2 to reach REM. Someone in REM may be disturbed and rapidly and reach WAKE. These stages can be statically model in memory and this is the approach that all state-of-art automated sleep staging has used.
- *Personalization:* In order to improve results even further the learning process can be supplemented with a per patient personalized model. A small subset of the given patient can be annotated and then used as a reference for the remaining data. In real world context the data collected in a sleep laboratory can be used to retrain a personalized model (with transfer learning). The personalized model will then be expected to perform even better than the general model.

To conclude, this field is evolving rapidly these days and we should expect some breakthroughs to come, having said that to process such big data without losing information and trying/combining all of the above will require a large cluster to perform in reasonable times, despite that, we showed that with a single simple CNN network, one can definitely get close to SOTA results.

References

[1] M. Walker, *Why we sleep*. London: Penguin Books, 2018.

[2] H. Reuveni, A. Tarasiuk, T. Wainstock, A. Ziv, A. Elhayany, and A. Tal, "Awareness Level of Obstructive Sleep Apnea Syndrome During Routine Unstructured Interviews of a Standardized Patient by Primary Care Physicians," *Sleep*, vol. 27, no. 8, pp. 1518–1524, Dec. 2004, doi: 10.1093/sleep/27.8.1518.

[3] V. Kapur *et al.*, "The Medical Cost of Undiagnosed Sleep Apnea," *Sleep*, vol. 22, no. 6, pp. 749–755, Sep. 1999, doi: 10.1093/sleep/22.6.749.

[4] K. E. Bloch, "Polysomnography: a systematic review," *Technol. Health Care*, vol. 5, no. 4, pp. 285–305, Oct. 1997, doi: 10.3233/THC-1997-5403.

[5] R. D. Chervin, D. L. Murman, B. A. Malow, and V. Totten, "Cost-Utility of Three Approaches to the Diagnosis of Sleep Apnea: Polysomnography, Home Testing, and Empirical Therapy," *Ann. Intern. Med.*, vol. 130, no. 6, p. 496, Mar. 1999, doi: 10.7326/0003-4819-130-6-199903160-00006.

[6] E. Dafna, A. Tarasiuk, and Y. Zigel, "Sleep staging using nocturnal sound analysis," *Sci. Rep.*, vol. 8, no. 1, pp. 13414–13474, Sep. 2018, doi: 10.1038/s41598-018-31748-0.

[7] J. M. Siegel, "Sleep viewed as a state of adaptive inactivity," *Nat. Rev. Neurosci.*, vol. 10, no. 10, pp. 747–753, Oct. 2009, doi: 10.1038/nrn2697.

[8] R. B. Berry *et al.*, "Rules for Scoring Respiratory Events in Sleep: Update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events," *J. Clin. Sleep Med.*, Oct. 2012, doi: 10.5664/jcsm.2172.

[9] Q. Q. Q. Li, Q. Q. Q. Li, C. Liu, S. P. Shashikumar, S. Nemati, and G. D. Clifford, "Deep learning in the cross-time frequency domain for sleep staging from a single-lead electrocardiogram," *Physiol. Meas.*, vol. 39, no. 12, p. 124005, Dec. 2018, doi: 10.1088/1361-6579/aaf339.

[10] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. Marcus, and B. V Vaughn, "The AASM manual for the scoring of sleep and associated events," *Rules, Terminol. Tech. Specif. Darien, Illinois, Am. Acad. Sleep Med.*, vol. 176, 2012.

[11] A. A. R. Kamal, J. B. Harness, G. Irving, and A. J. Mearns, "Skin photoplethysmography — a review," *Comput. Methods Programs Biomed.*, vol. 28, no. 4, pp. 257–269, Apr. 1989, doi: 10.1016/0169-2607(89)90159-4.

[12] A. Carpenter and A. Frontera, "Smart-watches: a potential challenger to the implantable loop recorder?," *EP Eur.*, vol. 18, no. 6, pp. 791–793, 2016, doi: 10.1093/europace/euv427.

[13] P. Fonseca *et al.*, "Validation of Photoplethysmography-Based Sleep Staging Compared With Polysomnography in Healthy Middle-Aged Adults," *Sleep*, vol. 40, no. 7, Jul. 2017, doi: 10.1093/sleep/zsx097.

[14] Z. Beattie *et al.*, "Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals," *Physiol. Meas.*, vol. 38, no. 11, pp. 1968–1979, Oct. 2017, doi: 10.1088/1361-6579/aa9047.

[15] Z. Shinar, A. Baharav, Y. Dagan, and S. Akselrod, "Automatic detection of slow-wave-sleep using heart rate variability," in *Computers in Cardiology 2001. Vol. 28 (Cat. No. 01CH37287)*, 2001, pp. 593–596.

[16] P. Dehkordi, A. Garde, W. Karlen, D. Wensley, J. M. Ansermino, and G. A. Dumont, "Sleep stage classification in children using photoplethysmogram pulse rate variability," in *Computing in Cardiology 2014*, 2014, pp. 297–300.

[17] A. Patanaik, J. L. Ong, J. J. Gooley, S. Ancoli-Israel, and M. W. L. Chee, "An end-to-end framework for real-time automatic sleep stage classification," *Sleep*, vol. 41, no. 5, May 2018, doi: 10.1093/sleep/zsy041.

[18] H. Phan, F. Andreotti, N. Cooray, O. Y. Chen, and M. De Vos, "SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 3, pp. 400–410, Mar. 2019, doi: 10.1109/TNSRE.2019.2896659.

[19] D. A. A. Dean *et al.*, "Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource," *Sleep*, vol. 39, no. 5, pp. 1151–1164, May 2016, doi: 10.5665/sleep.5774.

[20] G.-Q. Zhang *et al.*, "The National Sleep Research Resource: towards a sleep data commons," *J. Am. Med. Informatics Assoc. JAMIA*, vol. 25, no. 10, pp. 1351–1358, Oct. 2018, doi: 10.1093/jamia/ocy064.

[21] F. Liu *et al.*, "Dynamic ECG Signal Quality Evaluation Based on the Generalized bSQI Index," *IEEE Access*, vol. 6, pp. 41892–41902, 2018, doi: 10.1109/ACCESS.2018.2860056.