Sleep Staging with Convolutional Recurrent Neural Networks

J. Van Zaen, F. Braun, P. Renevey, M. Lemay

The correct identification of sleep stages is of the utmost importance in the diagnosis of sleep disorders. The gold standard for sleep staging is polysomnography which is laborious and expensive as it requires several sensors and manual annotation by a trained specialist. To mitigate these issues, a neural network architecture was developed to classify sleep stages from a single photoplethysmogram signal. Such an architecture could be fed data collected with a wearable sensor allowing for the automatic and unobtrusive estimation of sleep stages over several nights in home settings.

Accurately identifying sleep stages is critical to diagnose sleep disorders. The current gold standard for this task is polysomnography (PSG) which records several signals over one night including electroencephalogram (EEG), electrocardiogram (ECG), electromyogram (EMG), respiratory airflow, and oxygen saturation via photoplethysmogram (PPG). These signals are then visually inspected by a trained specialist that annotates 30-second windows as wakefulness (WAKE), light sleep (N1 and N2), deep sleep (N3), or rapid-eye movement (REM). This approach has several drawbacks. First, PSG is expensive and time-consuming as it required specialized equipment and multiple trained professionals. Consequently, it is usually performed for a single night and often in a sleep laboratory which might not be representative of usual sleep patterns. It is also cumbersome for the subjects as they need to wear many sensors and electrodes during the night which might affect their sleep. Finally, since the data is annotated by visual inspection, annotations from different specialists can significantly differ.

To alleviate the issues of PSG, several approaches that combine a reduced set of sensors with machine learning have been proposed. A recent study [1] used a neural network to classify sleep stages from a single PPG signal. This network combines convolutional layers to automatically extract high-level features from the input signal and recurrent layers to consider time dependencies across 30-second windows. Using only PPG data enables cost-efficient, long-term monitoring of sleep.

In this work, a new neural network architecture is derived from the abovementioned study [1] and applied to the MESA dataset [2,3] which comprises 1698 PSG records. Each record includes a PPG signal recorded at the finger with a sampling frequency of 256 Hz and is annotated over 30-second windows with the five sleep stages (WAKE, N1, N2, N3, or REM). The dataset was split into three subsets: 60% for training (n = 1018), 20% for validation (n = 340), and 20% for testing (n = 340). To prepare the data for training, the PPG signals were resampled at 64 Hz, filtered with a high-pass filter with a cut-off at 0.1 Hz, and scaled by the standard deviation estimated over the training subset. The proposed architecture takes 100 30-seconds windows of PPG data as input and yields the estimated sleep stage for each window as output. It is divided in three parts: a convolutional part to extract features, a recurrent part to consider dependencies between successive windows, and a fully connected part for classification. The first part consists of six convolutional layers with max pooling followed by a global average pooling layer. The second part is composed of two bidirectional gated recurrent unit (GRU) layers. Finally, the third part includes two fully connected layers for classification. The network parameters were trained by minimizing the categorical cross-entropy with the Adam optimizer and an initial learning rate of 0.001. The learning rate was divided by two if the loss on the validation subsets did not decrease for five consecutive epochs. To limit overfitting, Gaussian dropout before and regular dropout after the recurrent layers.

The performance of the neural network was evaluated in terms of overall accuracy for 3, 4, and 5 classes:

- 3 classes: WAKE, NREM (N1, N2, and N3), REM
- 4 classes: WAKE, LIGHT (N1 and N2), DEEP (N3), REM
- 5 classes: WAKE, N1, N2, N3, REM

The accuracy values for the training, validation, and test subsets are reported in Table 1. There was some overfitting as shown by the difference between the training subset and the validation and test subsets. Dropout seems insufficient and additional regularization might help to bridge the gap between training and validation performance. However, the values obtained on the validation subset are close to the ones obtained on unseen data.

Table 1: Accuracy metrics.

Subset	3-class accuracy	4-class accuracy	5-class accuracy
Training	91.3%	86.3%	81.2%
Validation	86.7%	80.8%	76.1%
Test	76.1%	80.8%	75.7%

These results suggest that reliable automatic sleep staging with a single PPG signal is possible. This could pave the way towards the unobtrusive and long-term sleep monitoring using wearables.

[1] H. Korkalainen, J. Aakko, B. Duce, S. Kainulainen, A. Leino, S. Nikkonen, I. O. Afara, S. Myllymaa, J. Töyräs, T. Leppänen, “Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea,” Sleep 43 (2020).

[2] X. Chen, R. Wang, P. Zee, P. L. Lutsey, S. Javaheri, C. Alcántara, C. L. Jackson, M. A. Williams, S. Redline, "Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (MESA)", Sleep 38 (2015) 877.

[3] G. Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, S. Redline, "The National Sleep Research Resource: towards a sleep data commons", J. Am. Med. Inform. Assoc. 25 (2018) 1351.