HHS Public Access

Author manuscript

Conf Proc IEEE Eng Med Biol Soc. Author manuscript; available in PMC 2019 July 01.

Published in final edited form as:

Conf Proc IEEE Eng Med Biol Soc. 2018 July ; 2018: 3240–3243. doi:10.1109/EMBC.2018.8513039.

A Novel Sleep Stage Scoring System: Combining Expert-Based Rules with a Decision Tree Classifier

Kristin M. Gunnarsdottir,
Johns Hopkins University, Baltimore, MD, USA

Charlene E. Gamaldo,
Johns Hopkins School of Medicine, Baltimore, MD, USA.

Rachel M. E. Salas,
Johns Hopkins School of Medicine, Baltimore, MD, USA.

Joshua B. Ewen,
Johns Hopkins School of Medicine, Baltimore, MD, USA.

Richard P. Allen, and
Johns Hopkins School of Medicine, Baltimore, MD, USA.

Sridevi V. Sarma [Member IEEE]
Johns Hopkins University, Baltimore, MD, USA

Abstract

Overnight polysomnography (PSG) is the gold standard tool used to characterize sleep and for diagnosing sleep disorders. PSG is a non-invasive procedure that collects various physiological data which is then scored by sleep specialists who assign a sleep stage to every 30-second window of the data according to predefined scoring rules. In this study, we aimed to automate the process of sleep stage scoring of overnight PSG data while adhering to expert-based rules. We developed an algorithm utilizing a likelihood ratio decision tree classifier and extracted features from EEG, EMG and EOG signals based on predefined rules of the American Academy of Sleep Medicine Manual. Specifically, features were computed in 30-second epochs in the time and the frequency domains of the signals and used as inputs to the classifier which assigned each epoch to one of five possible stages: N3, N2, N1, REM or Wake. The algorithm was trained and tested on PSG data from 38 healthy individuals with no reported sleep disturbances. The overall scoring accuracy was 80.70% on the test set, which was comparable to the training set. Our results imply that the automatic classification is highly robust, fast, consistent with visual scoring and is highly interpretable.

I. INTRODUCTION

Sleep is a basic human need and is strongly related to the quality of life [1]. An estimated 50–70 million Americans suffer from a chronic sleep disorder adversely affecting daily

Correspondence: kgunnar1@jhu.edu.

functioning and overall health. However, the majority of individuals who meet the diagnostic criteria for a sleep disorder are underdiagnosed [2], undiagnosed and/or untreated.

One important tool used to characterize sleep and evaluate certain sleep disorders is overnight polysomnography (PSG), which involves noninvasively collecting multiple physiological recordings. The data is then scored and reviewed in 30-second windows (epochs) by a registered sleep technologist who uses guidelines established by the American Academy of Sleep Medicine (AASM). Besides being both laborious and extremely time-consuming, this scoring process is also very subjective. Although accredited labs may vary on their degree of “software assisted” staging programs, the current industry standard still mandates visual scoring [3].

Numerous research studies have been conducted to establish the reliability of a completely automated scoring process. One limitation of many existing methods is that either the exact decision procedure is uninterpretable (e.g. as in [4]–[6]), or the features computed are not familiar by the visual experts. Additionally, most studies use only a small subset of the PSG recordings which may not provide sufficient information to discriminate between certain sleep stages [7], [8]. In order to build a more clinically adoptable system we designed an algorithm that closely follows the rationale and logic of the expert scorer who collectively considers a number of the PSG recordings when determining the appropriate sleep stage. Specifically, we developed an automated sleep stage scoring algorithm utilizing a likelihood ratio decision tree classifier that uses features extracted in the time and frequency domains of EEG, EMG and EOG signals based on predefined rules of the AASM Manual [3]. The performance of our classifier was evaluated by measuring the agreement between this automated program and a visual scoring expert.

II. Methods

A. Study Population

A total of 38 healthy adults participated in the study (Table I). The data was randomly split into a training set and a test set with 19 participants in each group. The training set was used to determine the optimal thresholds for each feature as well as the scoring order of sleep stages that yielded the maximum classification accuracy. Then, the performance of the algorithm was evaluated using the test set.

This study was approved by the Johns Hopkins Medicine Institutional Review Boards and all participants provided informed consent prior to enrollment. All participants were evaluated by a board-certified sleep specialist who conducted a clinical interview that also included a number of validated sleep surveys. Both “good sleepers” (PSQI ≤ 5) and “poor sleepers” (PSQI > 5) as defined by the Pittsburgh Sleep Quality Index (PSQI) [9] were included in the study. Participants were excluded if they endorsed symptoms consistent with Restless Leg Syndrome, circadian rhythm disorder or had evidence of clinically significant sleep apnea (Apnea Hypopnea Index > 5) on PSG.

B. Data Acquisition

The polysomnography was conducted in the Johns Hopkins Clinical Research unit. The recordings were performed utilizing the same sleep laboratory software, equipment model and procedural protocol [3] for all participants. The collected PSG data included six EEG channels, recorded from left and right frontal, central and occipital locations, two EOG channels (left and right eye), three EMG channels (chin, right leg and left leg), one ECG channel, respiratory flow and effort, oximeter, thermistor and cannula.

C. Data Analysis

1) Human Expert Visual Scoring—All PSG recordings were analyzed by a seasoned licensed and registered sleep technician using the Embla RemLogic sleep diagnostic software. The recordings were visually scored according to the AASM Manual for Scoring Sleep and Associated Events [3] by assigning one of five possible stages (N3, N2, N1, REM or Wake) to every 30-second epoch of the data. A board-certified sleep specialist reviewed and finalized all recordings.

2) Automated Sleep Stage Scoring Algorithm—The proposed automated sleep scoring system consists of five main steps. Prior to analysis, the data were preprocessed in accordance with AASM criteria and RemLogic settings (described below). Then, features based on the AASM scoring rules were extracted from the PSG signals. The third step entailed choosing an optimal threshold for each feature. A likelihood ratio decision tree classifier was then utilized to perform the classification and finally a set of temporal contextual smoothing rules was applied on the annotated data. All data analysis and scoring algorithm implementation was performed using Mathworks MATLAB R2015b.

3) Data Preprocessing—The PSG recordings used include four EEG channels (F3-A2, F4-A1, C4-A1 and O1-A2), both EOG channels and all three EMG channels. In fact, the EEG features were computed from all six channels, but only the channels giving the best separation of sleep stages were used in the final classifier. The signals were preprocessed and formatted based on the AASM guidelines [3]. Epochs that contained major movements and muscle artifacts obscuring the signals for more than half an epoch were manually identified and excluded when estimating the algorithm performance in accordance with AASM recommendations.

4) Feature Extraction—The continuous recordings were divided into non-overlapping 30-second epochs for feature extraction. The foundations for the decision rules were inspired by the AASM criteria and the features were extracted based on the corresponding characteristics of PSG data in the time and frequency domains. In addition, the algorithm included consensus input by an interdisciplinary team of certified sleep experts (Allen RP, Gamaldo CE, Salas RME, personal communication 2015–2016) and a biomedical engineering expert (Sarma SV, personal communication 2015–2016) with experience in neurophysiological signal processing. Thus, the final algorithm represents logic that reflects the features that quantify the AASM scoring rules and translate the sleep expert knowledge into metrics that can be used for automated processing. Table II lists the complete set of features used in the model and their corresponding sleep stages as well as the physiological

meaning of each feature. The first feature on the list (EOG_1) was used to split the epochs into two groups of possible stages (N3/N2/N1/REM/Wake vs. N1/REM/Wake) before assigning each epoch a sleep stage using the other features. All features were normalized to reduce the effects of individual differences on classifier performance. Features that were not inherently normalized by their computation method were normalized using:

$$z_{epoch} = \frac{x_{epoch} - \mu}{\mu} \quad (1)$$

where z_{epoch} is the normalized value for a particular epoch, x_{epoch} is the feature value at that epoch and μ is the average feature value across all epochs over the entire night. μ was calculated excluding the highest 5% and lowest 5% of values.

5) Threshold Determination—The decision thresholds were chosen using patient data from the training cohort. For each feature, we used the expert annotations to draw five probability distributions (one for each stage) and then chose a threshold attempting to optimize the number of epochs detected for the conditioned sleep stage with minimum decision error. The classification of sleep stages was performed in a hierarchical manner and thus the probability distributions were drawn using only the remaining epochs after scoring each stage. Two features and their sleep stage probability distributions are shown in Fig. 1.

6) Automatic Sleep Stage Classification—The classification process (Fig. 2) comprised five steps. First, epochs were divided into two groups based on the auto- and cross-correlations of the EOG signals. If little to no eye movement was present in an epoch, the epoch was assigned to group 1, else it was assigned to group 2.

In the second step, all epochs belonging to group 1 were assigned a sleep stage. In this group, all five sleep stages were possible and the stages were scored in the following order: N3, Wake, N1, N2 and REM. The scoring order was selected based on the discriminating ability of the features with stages that were easier to detect scored first. Once an epoch was assigned a sleep stage it was excluded from the scoring of the remaining stages. In our analysis, no quantitative feature was found to be informative enough to discriminate stage REM from the other stages. Consequently, after scoring the first four stages, all remaining unscored epochs were assigned to stage REM.

Since sleep is a continuous process, alternating between different sleep stages every 30 seconds is highly unlikely. The AASM Manual has a number of rules that recommend considering the neighboring epochs for the scoring of a current epoch [3]. Therefore, a smoothing process utilizing temporal contextual information was applied after scoring the epochs in group 1. These smoothing rules refer to the relationship between epochs prior to and after the current epoch. Specifically, let A, B and C represent the possible stages. Then three consecutive epochs of A, B, A were replaced with A, A, A and four consecutive epochs of A, B, B, A or A, B, C, A were replaced with A, A, A, A.

The fourth step consisted of scoring epochs in group 2. Eye movement activity was high in group 2 and thus epochs were only scored as N1, REM or Wake. In this group, stage Wake was scored first, followed by N1 and as for group 1, the remaining unscored epochs were set as stage REM. Lastly, the same set of contextual smoothing rules were applied to the epochs in group 2.

III. Results

The performance of the proposed algorithm was evaluated by comparing the agreement between the automatic classification and the human expert scoring which served as the gold standard. The overall scoring accuracy, after removing epochs containing major muscle or movement artifacts, was 79.87%, with 11,115 epochs out of 13,916 correctly classified. The results of the test set were highly comparable to the training set results. The overall scoring accuracy of the test set was slightly higher than for the training set. The test set scoring accuracy using all the available data was 77.00% and increased to 80.70%, with 11,035 epochs out of 13,674 correctly classified, after excluding major movement epochs. The highest agreement with the human scorer for a single test subject was 91.48% (Fig. 3(a)) and the lowest agreement was 67.51% (Fig. 3(b)). As Fig. 4(a) shows, the highest scoring accuracy was obtained for stage N3, and the highest percentage of captured epochs was obtained for stage N2, followed closely by Wake. N1 turned out to be the hardest stage to score with a scoring accuracy below 50% and far less epochs captured compared to the other stages. Fig. 4(b) shows the confusion matrix after scoring the test set. For N1, most misclassifications occurred between the N1-REM pair, followed by N1-N2 and N1-Wake. Other commonly misclassified pairs were N3-N2 and REM-N2. All remaining pairs had misclassification rates below 5%.

IV. Discussion

At present, the standard procedure of PSG data analysis is heavily dependent upon human visual scoring. Here, a new system for automatic sleep stage scoring of PSG data has been proposed, yielding an overall scoring accuracy of 80.70%. The agreement between the algorithm and the human expert is highly comparable to reported inter-scorer reliability amongst sleep experts which has been found to be around 82% [18]. An important limiting factor of visual scoring is the time it takes to score each study, which typically requires around 2.5–4 man-hours (Johns Hopkins Center for Sleep, personal communication, April 28 2016). Not only does it contribute to high operating costs of sleep centers, but is also expensive in terms of valuable expert time. In comparison, the average run-time of our algorithm was 32.5±1.9 seconds for feature extraction and scoring of a full night’s sleep recording.

The results of our algorithm are similar to reported performances of existing sleep stage classification systems, with stage N1 recurrently being the most misclassified sleep stage [5], [8], [11]–[17], [19]. However, the proposed algorithm represents several desirable and superlative features over these methods. Our algorithm was driven by rules established by sleep experts, features that they quantify through visual inspection, and decision rules that are explicit and interpretable. We conclude that an automatic classification algorithm based

on a likelihood ratio classifier, and importantly using features extracted from the AASM Manual, can to a large extent reproduce the judgment of a visual scoring expert. Therefore, we see this tool as assisting visual scorers to speed up their process and providing a way to diagnose sleeping disorders in a more robust, quantitative and ultimately cost-effective manner.

Acknowledgments

Research supported by The Johns Hopkins Catalyst Award; JHU CFAR NIH/NIAID fund 1P30AI094189-01A1; NSF 1609038.

References

[1]. Zammit GK et al., "Quality of life in people with insomnia," Sleep, vol. 22, Suppl 2, pp. S379–385, 5 1999. [PubMed: 10394611]

[2]. Institute of Medicine Committee on Sleep Medicine and Research; Colten HR, Altevogt BM, editors. Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem. Washington, DC, USA: National Academies Press [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK19960/

[3]. Berry RB et al., The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, version 2.3, American Academy of Sleep Medicine, Darien Illinois, 2016.

[4]. Ebrahimi F. Automatic sleep stage classification based on EEG signals by using neural networks and wavelet packet coefficients. Conf. Proc. IEEE Eng. Med. Biol. Soc.; Vancouver, BC. 2008. 1151–1154.

[5]. Tsinalis O et al., "Automatic sleep stage scoring using time-frequency analysis and stacked sparse autoencoders," Ann. Biomed. Eng, vol. 44, no. 5, pp. 1–11, 2015.

[6]. Hassan AR and Bhuiyan MIH, "Automatic sleep stage classification," in Proc. 2nd Int. Conf. Elect. Inform. Commun. Technol., Khulna, Bangladesh, 2015, pp. 211–216.

[7]. Bódizs R et al., "Wakefulness–sleep transition: Emerging electroencephalographic similarities with the rapid eye movement phase," Brain Res. Bull, vol. 76, no. 1–2, pp. 85–89, 5 2008. [PubMed: 18395615]

[8]. Zoubek L et al., "Feature selection for sleep/wake stages classification using data driven methods," Biomed. Signal Process. Control, vol. 2, no. 3, pp. 171–179, 7 2007.

[9]. Buysse DJ et al., "The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research," Psychiatry Res., vol. 28, no. 2, pp. 193–213, 5 1989. [PubMed: 2748771]

[10]. Wendt SL. Validation of a novel automatic sleep spindle detector with high performance during sleep in middle aged subjects. Conf. Proc. IEEE Eng. Med. Biol. Soc.; San Diego, CA. 2012. 4250–4253.

[11]. Virkkala J et al., "Automatic sleep stage classification using two-channel electro-oculography," J. Neurosci. Methods, vol. 166, no. 1, pp. 109–115, 10 2007. [PubMed: 17681382]

[12]. Fraiwan L et al., "Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier," Comput. Methods Programs Biomed, vol. 108, no. 1, pp. 10–19, 10 2012. [PubMed: 22178068]

[13]. Hsu Y-L et al., "Automatic sleep stage recurrent neural classifier using energy features of EEG signals", Neurocomput, vol. 104, pp. 105–114, 3 2013.

[14]. Liang SF et al., "Automatic stage scoring of single-channel sleep EEG by using multiscale entropy and autoregressive models," IEEE Trans. Instrum. Meas, vol. 61, no. 6, pp. 1649–1657, 6 2012.

[15]. Figueroa Helland VC et al., "Investigation of an automatic sleep stage classification by means of multiscorer hypnogram," Methods Inf. Med, vol. 49, no. 5, pp. 467–472, 2010. [PubMed: 20644896]

[16]. Fraiwan LA et al. “Automatic sleep stage scoring with wavelet packets based on single EEG recording,” Int. J. Med. Health Biomed. Bioeng. Pharm. Eng, vol. 3, no. 6, pp. 485–488, 2009.

[17]. Güneş S et al., “Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting,” Expert Syst. Appl, vol. 37, no. 12, pp. 7922–7928, 12 2010.

[18]. Danker-Hopfe H et al., “Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard,” J. Sleep Res, vol. 18, no. 1, pp. 74–84, 3 2009. [PubMed: 19250176]

[19]. Doroshenkov LG et al., “Classification of human sleep stages based on EEG processing using hidden Markov models,” Biomed. Eng, vol. 41, no. 1, pp. 25–28, 2007.

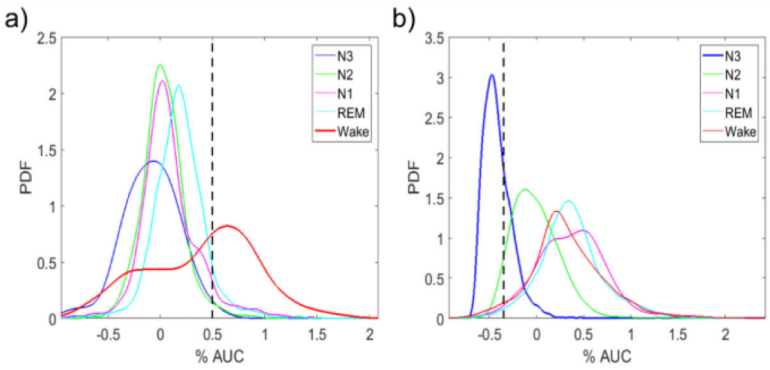

Fig. 1.

Probability density functions and the corresponding decision thresholds for (a) relative Alpha power, a feature for stage Wake, and (b) relative Beta power, a feature for stage N3. The decision thresholds are marked with a black dashed line.

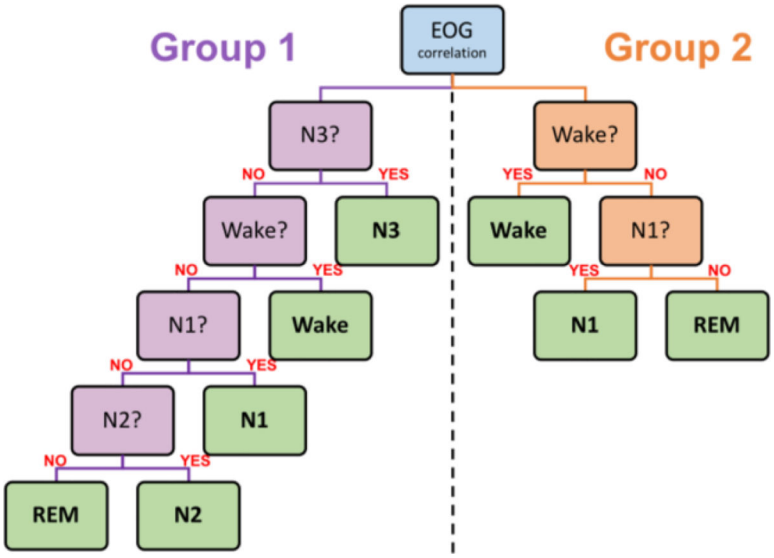

Fig. 2.

A flowchart showing the automatic scoring process of the likelihood ratio decision tree classifier.

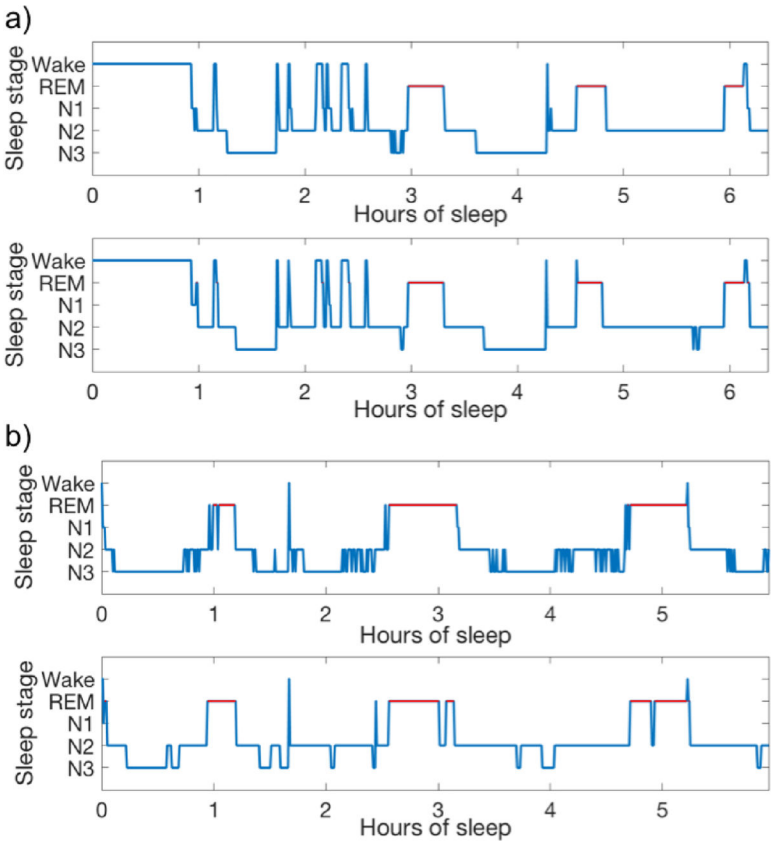

Fig. 3.

Comparison of hypnograms scored by the human expert (top) and by the algorithm (bottom) for (a) the subject with the highest scoring accuracy and (b) the subject with the lowest scoring accuracy.

a)

Sleep stage	Scoring accuracy (%)	Epochs captured (%)
Wake	86.02	93.70
N1	43.24	15.24
N2	77.94	90.44
N3	91.85	67.87
REM	75.73	81.23
Overall scoring accuracy (%)	**80.70**	

b)

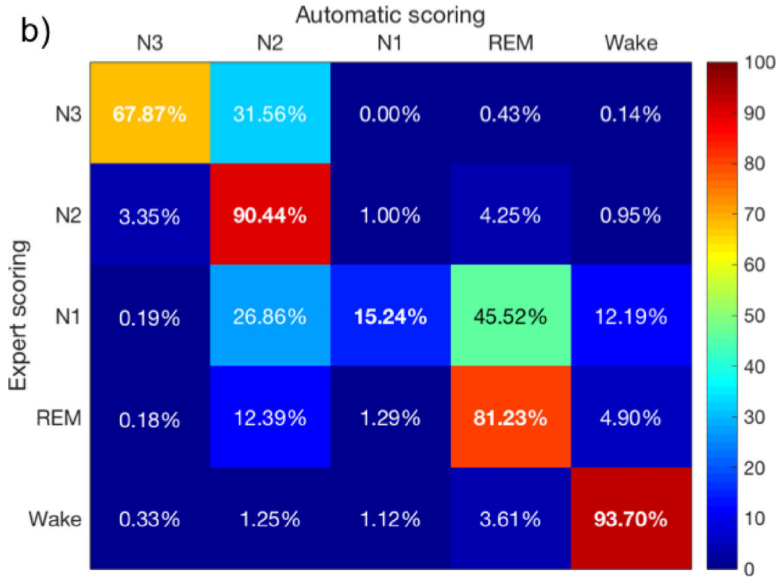

Fig. 4.

Scoring results of the test set. (a) Scoring accuracy and epochs captured are reported for each sleep stage along with the overall scoring accuracy. (b) Confusion matrix for the automatic scoring algorithm. The values are the percentage of epochs belonging to the stage scored by the expert (indicated by the rows) that were classified by our algorithm as the stage indicated by the columns. The diagonal elements represent the percentage of epochs where the automatic classifier was in agreement with the human expert for each sleep stage.

TABLE I.

Population Statistics

Subject statistics	Training Set	Test Set
Gender (M / F)	7/12	9/10
Age (years)	23.8±3.0	24.6±3.4
Ethnicity (Caucasian / Asian /African American)	9/5/5	8/8/3
PSQI	2.2±1.7	3.7±2.1
Time in bed (hours)	6.2±0.1	6.2±0.1

TABLE II.

The features used in the classifier

Sleep stage	Quantitative feature	Signal	AASM feature
Group 1 vs. Group 2	$EOG_1^{0.3-35}$	Right EOG and Left EOG	Eye movements present/absent
Wake	EMG Energy	EMG Chin, Left Leg, Right Leg	Increased EMG activity
	Alpha Power	O1-A2 EEG	Alpha rhythm observed
	Theta Power	O1-A2 EEG	Low theta activity
N1	$EOG_2^{0.1-0.45}$	Right EOG and Left EOG	Eye movements present
N2	Maximum Spindle Duration	F3-A2 EEG	Spindles present
	Number of Spindles	C4-A1 EEG	Spindles present
	Delta Power	F3-A2 EEG	Moderate to high delta activity
	$EOG_3^{0.3-0.45}$	Right EOG and Left EOG	Little to no rapid eye movements
N3	Delta Power	F4-A1 EEG	High delta activity
	Beta Power	F3-A2 EEG	Low beta activity

Group 1 = N3/N2/N1/REM/Wake, Group 2 = N1/REM/Wake. The EOG features are combinations of the cross- and autocorrelations of the two EOG signals, and the EMG energy is strongly linked to muscular activity. Spindles were detected using the Wendt algorithm [10] and the length of a single spindle was restricted to 0.5–2 seconds.