HHS Public Access

Author manuscript

Clin Neurophysiol. Author manuscript; available in PMC 2021 September 16.

Published in final edited form as:

Clin Neurophysiol. 2020 June ; 131(6): 1187–1203. doi:10.1016/j.clinph.2020.02.027.

Automatic Detection of Cortical Arousals in Sleep and their Contribution to Daytime Sleepiness

Andreas Brink-Kjaer[a,b,c], **Alexander Neergaard Olesen**[a,b,c], **Paul E. Peppard**[d], **Katie L. Stone**[e,f], **Poul Jennum**[c,*], **Emmanuel Mignot**[a,*], **Helge B.D. Sorensen**[b,*]

[a]Center for Sleep Sciences and Medicine, Stanford University, California, USA

[b]Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark

[c]Danish Center for Sleep Medicine, Glostrup University Hospital, Glostrup, Denmark

[d]Department of Population Health Sciences, University of Wisconsin-Madison, Madison, Wisconsin, USA

[e]Research Institute, California Pacific Medical Center, San Francisco, CA

[f]Department of Epidemiology and Biostatistics, University of California, San Francisco

Abstract

Objective: Significant interscorer variability is found in manual scoring of arousals in polysomnographic recordings (PSGs). We propose a fully automatic method, the Multimodal Arousal Detector (MAD), for detecting arousals.

Methods: A deep neural network was trained on 2,889 PSGs to detect cortical arousals and wakefulness in 1-second intervals. Furthermore, the relationship between MAD-predicted labels on PSGs and next day mean sleep latency (MSL) on a multiple sleep latency test (MSLT), a reflection of daytime sleepiness, was analyzed in 1447 MSLT instances in 873 subjects.

Results: In a dataset of 1,026 PSGs, the MAD achieved an F1 score of 0.76 for arousal detection, while wakefulness was predicted with an accuracy of 0.95. In 60 PSGs scored by nine expert technicians, the MAD performed comparable to four and significantly outperformed five expert technicians for arousal detection. After controlling for known covariates, a doubling of the arousal index was associated with an average decrease in MSL of 40 seconds ($p = 0.0075$).

Conclusions: The MAD performed better or comparable to human expert scorers. The MAD-predicted arousals were shown to be significant predictors of MSL.

Significance: This study validates a fully automatic method for scoring arousals in PSGs.

Corresponding author: Name: Andreas Brink-Kjaer, andbri@dtu.dk, Postal address: Technical University of Denmark, Department of Health Technology, Ørsteds Plads, Building 349, 2800 Kgs. Lyngby, Denmark.

*Shared last author

Author Contributions

A.B.K., H.B.D.S., P.J., E.M. contributed to the conception and design of the study. A.B.K., A.N.O., H.B.D.S. contributed to developing the model. A.B.K., E.M. contributed to the statistical analysis. E.M., A.N.O., P.E.P., K.L.S. contributed to data acquisition. A.B.K., A.N.O., P.E.P., K.L.S., P.J., E.M., H.B.D.S. contributed to drafting the manuscript.

Keywords

Arousal; Polysomnography; Automatic detection; Deep neural networks; Daytime sleepiness; MSLT

1 Introduction

Sleep dysregulation and sleep disorders are associated with cardiovascular, metabolic, and psychiatric disorders (Saper et al. 2005; Kryger et al. 2011). Subjects with complaints of sleep problems are usually examined in sleep clinics, through analysis of nocturnal polysomnography (PSG) and, less frequently using multiple sleep latency tests (MSLTs), an objective measure of daytime sleepiness where latencies to sleep are measured during 4–5 daytime naps. A PSG recording involves measuring electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), electrocardiography (ECG), airflow, respiratory effort and blood oxygen saturation during a night's sleep. Analysis of PSGs include scoring of Sleep Disordered Breathing (SDB) events (reported number of apneas or hypopneas per hour of sleep as the Apnea Hypopnea Index, or AHI), periodic leg movements during sleep (number of periodic leg movement (PLM) event per hour of sleep with and without associated arousals, PLMI and PLMAI respectively), and sleep stages [wakefulness (W), non REM sleep (stage one N1, two N2, or three N3) or REM sleep (R)], reported as percent of total sleep time. Hypopneas in the AHI can be defined as events with a drop of ≥30% peak respiratory signal excursions lasting ≥10 seconds, which is associated with either a 3% oxygen desaturation and/or an arousal; or only as events associated solely with 4% oxygen desaturation. The latter definition is used by Medicare, which is the main insurance provider for older adults in the United States, while the former is the most commonly used definition in sleep clinics and the one used in this work. A typical sleep study also reports sleep latency, latency for sleep onset to the first epoch of REM sleep, and sleep efficiency (SE), the percent of time asleep when in bed. A slight variation of sleep efficiency is Wake After Sleep Onset (WASO), which, unlike SE, only considers wake after sleep onset has occurred. In the context of sleep stage scoring, sleep stages are attributed to successive 30 second epochs using a majority rule, an arbitrary decision historically justified by the use of paper printing in sleep studies (Stephansen et al. 2018).

In addition to traditional sleep scoring every 30 seconds, sleep is often disturbed by transient arousal microevents between 3 and 15 seconds, or smaller wake segments that disturb the EEG, but that are not long enough to be scored as a full epoch of wakefulness (>15 sec of wake during a 30 sec epoch). These events, called microarousals, are also often associated with brief increases in muscle tone in the EMG, another important feature allowing proper scoring of arousals notably during REM sleep. Although microarousals can occur naturally as part of normal sleep-wake physiology, these are often the result of external stimuli (e.g. disturbing sound) or internal sleep disorder events such as SDB (i.e. sleep apnea) or PLM events. In traditional sleep study scoring, arousals are not always systematically scored, although in most instances arousals, whether of short duration or resulting in a full-blown epoch of wakefulness, are generally scored as part of the AHI and the PLMAI. Scoring of cortical arousals in this context is carried out according to AASM guidelines (Berry et

al. 2018). The scoring rules for arousal events, according to the AASM guidelines, do not depend on context of SDB or PLM events.

As mentioned above, wake events disturbing sleep longer than 15 seconds are reported as part of SE or WASO, whereas microarousals are only systematically scored by technicians when following SDB or PLM events. An excessive number of arousals, whether in the form of brief epochs of wakefulness (integrated in SE and WASO measures), or as microarousals (reported as part of the AHI or PLMI), is associated with sleep fragmentation and poor sleep, which in turn is linked to daytime sleepiness (Halasz et al. 2004). Subjects with excessive daytime sleepiness have a sevenfold greater rate of automobile accidents (Findley et al. 1988) and decreased quality of life.

As of today, the gold standard for detecting arousals is through visual inspection of PSG recordings [5]. This approach is both time-consuming, expensive, and has a low intra- and interscorer reliability due to subjective interpretation of used scoring guidelines (Bonnet et al. 2007; Magalang et al. 2013). Furthermore, as mentioned above, many shorter-duration arousals are often only scored in the context of SDB or PLMs, which is strongly limiting as spontaneous arousals are also likely affecting physiology.

The limitations of current arousal detection approaches have motivated the development of algorithms to automatically detect arousals, notably microarousals. Several computerized methods in the literature have been proposed to detect arousals using machine learning on features derived from PSG signals (De Carli et al. 1999; Cho et al. 2005; Shmiel et al. 2009; Sugi et al. 2009; Sorensen et al. 2012; Popovic et al. 2013; Shahrbabaki et al. 2015; Wallant et al. 2016; Fernández-Varela et al. 2017a, 2017b).

These studies provide working arousal detection systems; however, they are only validated in small datasets (6 – 60 PSGs) that have been manually scored by only a few human scorers, which makes comparison and generalization difficult to assess. Recently, two groups (Alvarez-Estevez and Fernández-Varela 2019; Olesen et al. 2019) validated automatic arousal detectors in larger cohorts. Olesen et al. (2019) used a deep learning framework to predict arousals and validated the approach in 1000 PSGs with an F1 score of 0.75 (for an explanation of the F1, see section 2.7). Alvarez-Estevez and Fernández-Varela (2019) validated an existing simple model (Fernández-Varela et al. 2017a) in 2768 PSGs with an F1 score of 0.64. This simple model was initially reported with an F1 score of 0.79 in a smaller cohort of 22 PSGs, which shows a lower performance for unseen data and emphasizes the importance of validating using large independent cohorts.

Another limitation is that the gold standard for scoring arousals through visual inspection of PSG is based on rules of duration that distinguishes microarousals (3–15 seconds) and wake (>15 seconds), a distinction which is arguably arbitrary.

In this study, we aimed at developing a fully automatic system, the Multimodal Arousal Detector (MAD), for the detection of all microarousals and wake events using recent advances in machine learning such convolutional and recurrent neural networks. The proposed approach challenges the gold standard by combining automatic scoring of arousals and wake with a 1-second resolution as a single measure. Furthermore, our study also differs

from previously reported state-of-the-art methods given the much larger sample size, as 5,362 PSGs gathered at multiple locations are used.

Our method merges brief and long arousals scored by the MAD as a single measure, which we have clinically validated by as predictors of daytime sleepiness, as assessed by the MSLT. The rationale of this analysis was to show that scoring arousals and wake in 1-second intervals and combining these provide a meaningful measure that correlated with next day sleepiness.

2 Method

2.1 Data Description

Diversity in datasets is a prerequisite for developing and validating deep learning detectors (LeCun et al. 2015). Sufficient data diversity was ensured by using data from thousands of subjects from four cohorts, scored by different sleep technicians based at various sleep centers. These include the MrOS Sleep Study (Blank et al. 2005; Orwoll et al. 2005; Blackwell et al. 2011; Dean et al. 2016), the Cleveland Family Study (CFS) (Redline et al. 1995, 1999; Dean et al. 2016), the Wisconsin Sleep Cohort (WSC) (Young et al. 2008), and the Stanford Sleep Cohort (SSC) (Andlauer et al. 2013). All PSGs included at least a central (C3/A2) EEG derivation, left and right EOG, chin EMG, and lead II ECG. The cohorts contain in-lab recordings (WSC, SSC), as well as Home Sleep Testing (HST) (MrOS, CFS). Signals of electrophysiological data were sampled at frequencies between 100 and 512 Hz. We note that not all cohorts had consistent and systematic scoring of all microarousals. To test interscorer reliability in comparison to detector performance, PSGs from 30 SSC and 30 WSC subjects were annotated five times each using a pool of nine sleep technicians. Annotations included microarousal scoring as defined in the AASM manual (Berry et al. 2018). The subset of 30 SSC PSGs comprised patients with various sleep disorders such as SDB (n=24), insomnia (n=4), delayed sleep phase syndrome (n=1), and others (n=4). Informed consent has been obtained by all participants of the parent epidemiological studies, and the study approved by parent cohorts steering committees and the Stanford Institutional Review Board.

2.1.1 MrOS Sleep Study—The MrOS Sleep Study is a multi-center community-based cohort designed to study the relationships between sleep disorders and vascular disease, falls, fractures, and mortality in older men. A total of 2,909 men age 67 years or older underwent a full unattended PSG at six clinical sites in the United States (Blank et al. 2005; Orwoll et al. 2005; Blackwell et al. 2011; Dean et al. 2016). PSG recordings were acquired with Compumedics Safiro Sleep Monitoring System that used a high pass pre-filter with a cut-off frequency of 0.16 Hz. A total of 2,888 PSGs were included from this study. All arousals (not just those associated with SDB or PLMs) were scored according to an older ASDA (American Sleep Disorders Association's) definition (Bonnet et al. 1992; Blackwell et al. 2011), precursor of the current AASM definition. Sleep stages were scored based on Rechtschaffen and Kales (RK) rules, however rules were slightly modified (e.g. deep sleep was scored as N3 sleep) (Hobson 1969). In this study, we consider the differences in sleep

stage scoring guidelines negligible for the purpose of modelling wake/sleep in the included cohorts.

2.1.2 Cleveland Family Study—The CFS is a large family study of sleep apnea conducted in 2,284 subjects of age between 6 and 88 from 361 families (Redline et al. 1995, 1999; Dean et al. 2016). PSGs in this study were recorded using a Compumedics E-Series System and a band-pass pre-filter with cut-off frequencies of 0.16 Hz and 105 Hz. This study included 726 PSGs from the CFS. Certified sleep technicians scored all PSGs with rules similar to those used in the MrOS Sleep Study. All arousals were scored in the CFS sample.

2.1.3 Wisconsin Sleep Cohort—Participants of the WSC study were randomly sampled from Wisconsin state agencies (Young et al. 2008), the age of the participants that was included ranged from 37 to 85. PSGs in this study were measured with a Grass Comet Lab based system using a pre-filter with cut-off frequencies 0.3 Hz and 35 Hz for EEG and EOG, while cut-off frequencies of 10 Hz and 70 Hz was used for EMG (Young et al. 2008). Among the large WSC sample, a subset of 271 PSGs had all arousals scored, although only the onset of arousals was annotated. In addition, 1,447 PSGs with an associated MSLT, but without independent scoring of arousals were included. In this study, these data were used to examine the clinical significance of the proposed system by comparing statistics of model predictions to daytime sleepiness.

2.1.4 Stanford Sleep Cohort—PSGs in the SSC were recorded in patients with a wide range of sleep disorders at the Stanford Sleep Clinic (Andlauer et al. 2013). PSGs in the SSC were recorded using a band-pass pre-filter with cut-off frequencies of 0.1 Hz and 0.45 times the sampling frequency. Thirty subjects of age between 20 and 90 had all arousals scored.

2.1.5 Data Usage and Summary—In developing the MAD, one dataset was used as a training set while a separate test dataset was set aside to provide unbiased estimates of model's performance. PSGs from MrOS and CFS were the only datasets with accurate and consistent scoring of all arousals. We therefore used 80 % of randomly selected subjects from both cohorts as the training set, while keeping the remaining 20 % for testing. Also included in the validation set were 271 PSGs from the WSC which had arousal scoring consistent with MrOS and CFS.

The remaining PSGs from WSC (n=1447) did not have the necessary arousal scoring for inclusion in the validation set but did have associated MSLT scores. These data are used to test whether MAD-predicted arousals are related to objective sleepiness measured clinically by the MSLT.

Finally, a separate test set of 30 PSGs from each of WSC and SSC was used to compare model performance to the performance of human scorers. The demographics as well as data usage are described in Table 1. Subjects are aged 6 to 90 and are generally overweight with an average BMI of 30.4. Research performed in this manuscript has been reviewed and approved by Stanford Institutional Review Boards.

2.2 Preprocessing of Biomedical Signals

Physiological signals were acquired using varying signal montages and can be contaminated with artifacts that decrease the signal quality. To enable consistent quality, preprocessing steps described below and summarized in Fig. 1 were performed.

This study used convolutional neural networks that employs a series of filters to compute descriptive signal features i.e. a low dimensional representation of the signals. Filter kernels designed by a convolutional neural network are static in size, hence networks expect an input with a consistent sampling frequency. For this reason, we resampled all signals to 128 Hz. Aliasing effects were avoided by applying an anti-aliasing low pass least-squares Finite Impulse Response (FIR) filter. The FIR filter was designed to minimize the weighted integrated squared error between the filter's magnitude response and an ideal piecewise function over a set of desired frequency bands that allow resampling without aliasing. Finally, a Kaiser window with a shape factor $\beta = 5$ was used to normalize the filter gain.

Infinite Impulse Response (IIR) band pass filters were used to remove frequency content that represent power line interference at 60 Hz and low frequency artifacts, while preserving most physiological meaningful data. The IIR filters designed for EEG and EOG had passband ranges from 0.5 to 35 Hz with an allowed passband ripple of 1 dB, and stopband frequencies of 0.1 and 50 Hz with a stopband attenuation of 20 dB using the Butterworth method. The IIR filter used for EMG was similar but the first stopband ends at 5 Hz and the passband starts at 10 Hz. Filters were implemented with zero-phase filtering, that is the filter is applied forward and backward, to avoid the issue of non-linear phase response and frequency-dependent group delay.

Recursive Least Square (RLS) adaptive filters as previously implemented (He et al. 2004; Moore et al. 2014) were used to respectively remove ECG and ocular movement artifacts in the EEG. Both studies reported that using a filter order of $p = 4$ and a forgetting factor $\lambda = 0.995$ resulted in satisfactory results. RLS adaptive filters with the same settings were similarly used in this study. The study by (Moore et al. 2014) was also based on the WSC and SSC, which supports the use of same filter settings.

Prior to feeding signals to the deep neural network, signal distributions were standardized by subtracting the mean and dividing by the standard deviation. This is a commonly used simple standardization (LeCun et al. 2012, 2015), which is necessary to avoid a vanishing gradient during training of the neural network.

2.3 Classification with Convolutional and LSTM Neural Networks

Uses of convolutional and recurrent neural networks have been shown to achieve state-of-the-art performance in various fields (LeCun et al. 2015; Lipton et al. 2015), including sleep analysis (Supratak et al. 2017; Biswal et al. 2018; Stephansen et al. 2018). A convolutional neural network (CNN) works by taking a static input such as a signal or image, and the CNN processes it with a network of filters. In 1-dimensional CNNs, non-linear features are extracted in each layer by transforming a set of inputs $a^l(t)$ into new feature maps as

$$a_j^{l+1}(t) = f\left(\sum_{n=1}^{N^l}\left[\sum_{k=1}^{K^l} w_{jn}^l(k)a_n^l(t-k)\right] + b_j^l\right) \quad \#(1)$$

where $a_j^{l+1}(t)$ is a new feature map j in layer $l+1$, f is a non-linear activation function, b_j^l is the bias term, and w_{jn}^l is the convolution kernel of size K^l that is convolved with the n^{th} feature map in layer l. Long short-term memory (LSTM) networks were used in the network, which is a type of recurrent neural network capable of modelling long-term dependencies without the problem of exploding or vanishing gradients (Hochreiter and Schmidhuber 1997). The LSTM network contains a cell state C_t, which saves information about long and short-term changes of CNN input features. The information saved is regulated by an input gate i_t, a forget gate f_t, and a new candidate cell state $\widetilde{C}_t$, which depending on the input features and previous hidden state h_{t-1}, decides what information is saved and what is discarded. The operation performed by these gated can be defined as

$$i_t = \sigma(W_i[x_t, h_{t-1}] + b_i) \quad \#(2)$$

$$f_t = \sigma\left(W_f[x_t, h_{t-1}] + b_f\right) \quad \#(3)$$

$$\widetilde{C}_t = \tanh(W_C[x_t, h_{t-1}] + b_C) \quad \#(4)$$

$$C_t = f_t \bullet C_{t-1} + i_t \bullet \widetilde{C}_t \quad \#(5)$$

Where σ is the softmax function, $[x_t, h_{t-1}]$ is the concatenated input features and previous hidden state, and (W, b) are the learned weight and bias terms. The output of the LSTM network uses the saved information in the cell state C_t to output a new hidden state h_t using an output gate o_t as

$$o_t = \sigma(W_o[x_t, h_{t-1}] + b_o) \quad \#(6)$$

$$h_t = o_t \bullet \tanh C_t \quad \#(7)$$

This memory allows features to incorporate temporal context, such those indicating prior sleep stage or frequency shift, which is ideally suited to detect arousals.

2.4 Network Architecture

The architecture of the proposed network is based on similar studies of sleep staging (Supratak et al. 2017; Biswal et al. 2018; Olesen et al. 2018; Stephansen et al. 2018). The general idea is to use CNNs to automatically design a set of features describing the preprocessed EEG, EOG, and EMG signals in 1 second bins. These features are then fed to a bi-directional LSTM network followed by fully connected neural networks to predict labels (*arousal*, *non-arousal*) and (*wake*, *sleep*) as probabilities for each successive second of data.

The fully connected layers enable correct classification of arousal and wake by making a highly complex non-linear mapping of features computed at the level of the LSTM layers. Labels are associated with a 1 second signal bin, consisting of 4 signals with 128 samples. The proposed CNN is based heavily on the network structure used in the work of He et al. (2016). This network structure performs very well in the field of image recognition (He et al. 2016) and is easily restructured for 1-D inputs. The network is comprised of residual building blocks that use convolutional layers, batch normalization (Ioffe and Szegedy 2015), ReLU activation functions (Dahl et al. 2013), and residual learning. He et al. (2016) proposed a deep CNN with residual learning that improved performance without adding further network complexity. Residual learning is implemented as shortcut connections as identity mappings between every second convolutional layer. Identity mappings are straight forward with identical dimensions, for increasing dimensions zero-padding is used, and decreasing dimensions are handled with max pooling. A simple illustration of the network architecture and how signals are processed is shown in Fig. 2. The full network is displayed in Fig. 3. In the network, all convolutional and fully connected layers are followed by batch normalization and ReLU activation, while output probabilities are computed using the softmax activation.

2.5 Network training

The proposed deep neural network was optimized iteratively using an average of the cross entropy cost for arousal and wake as

$$C(\theta) = -\frac{1}{N}\sum_{1}^{N} y_n \log(\hat{y}_n) + (1 - y_n)\log(1 - \hat{y}_n) \quad \#(8)$$

$$C_{total}(\theta) = \frac{1}{2}C_{arousal}(\theta) + \frac{1}{2}C_{wake}(\theta) \quad \#(9)$$

where $C(\theta)$ is the cost function with variables θ, N is the mini-batch size, y_n is the n^{th} true label, and $\hat{y}_n$ is the n^{th} predicted class probability. The dependency of the cost function C on variables θ is given by the implicit dependency of $\hat{y}_n$ on θ. The cross entropy cost $C_{total}(\theta)$ was minimized using the Adam optimization algorithm (Kingma and Ba 2014) with $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Data was processed using a mini-batch size of $N = 300$ of 1 second bins, corresponding to 5 minutes of PSG. This ensures enough temporal context for prediction of arousals and wake. A learning rate $\eta = 10^{-3}$ was used and weights were initialized according to Glorot and Bengio (2010).

The network was validated on a subset of 20 PSGs every 5,000 iterations, which was used to visually determine when network performance saturated. The network was trained for 350,000 iterations over roughly 30 hours, corresponding to a total of 3,240 full night's PSGs. The architecture of the model was selected based on a hyper-parameter grid search, in which different number of LSTM cells and convolutional layers were used. In this grid search, the forward and backward LSTM layer had either 64, 128 or 256 cells, and a total of 13 or 19 convolutional layers. The optimal configuration of hyper-parameters was selected based on the minimal cross entropy cost obtained after 50,000 training iterations.

2.6 Probability Postprocessing

The cost function C provides a measure of prediction error, however it does not directly provide information about how many arousal events are detected correctly. In fact, the model does not even predict events, but rather a probability of arousal or wake for each second. Events can be detected from the output probability by the proposed postprocessing, which binarizes the probability.

The postprocessing for the arousal probability P(Arousal) was implemented as follows

1. Threshold arousal probability, P(Arousal) > T_{ar}.
2. Connect arousal events closer than 10 seconds.
3. Discard detected events shorter than 3 seconds.

where the threshold T_{ar} was optimized using the validation set. Rule 2 and 3 are based on the AASM manual (Berry et al. 2018) as scored arousals require a preceding 10 seconds of sleep and a duration of 3 seconds or more.

The postprocessing for the wake probability was implemented as

1. Threshold wake probability, P(Wake) > T_w.
2. Connect wake periods closer than 15 seconds.
3. Remove wake periods with duration less than 15 seconds.

where the wake threshold T_w was also optimized using the validation set. Periods of wake are connected and discarded based on a 15 seconds criterion due to the gold standard, which states that 30-second epochs should be labelled as wake if more than half has the characteristics of wake (see introduction).

2.7 Model Validation and Testing of the Multimodal Arousal Detector

The mini-batch window processes 300 seconds of data at a time. During model validation and testing, the mini-batch window was moved 150 seconds at a time while using the central 150 seconds of each mini-batch, thereby excluding the ends of each window that do not have sufficient temporal context for optimal classification. The following set of performance metrics were used to evaluate the model. Predictions of arousals and wake in 1 second bins were compared to target labels. For sleep/wake, target labels of 30 seconds were split into 1 second epochs to allow for direct comparison of model decisions. Detected arousal events do not have to match the labels to indicate the same event, therefore an arousal event true positive (TP) is defined as a predicted and target event having any overlap. In the context of measuring arousal scoring performance, true negatives (TNs) are trivial as predicting *non-arousal* is a very easy task. Performance of arousal scoring was therefore measured by precision and recall, defined as

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \qquad \#(10)$$

where TPs are either matching true 1 second bins or overlapping arousal events, FP are false positives, and FN false negatives. As both metrics are essential to measure performance, the F1 score was also used to summarize performance:

$$\mathrm{F1} = 2 \bullet \frac{\mathrm{Precision} \bullet \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \quad \#(11)$$

In addition to F1 scores, the false positive rate (FPR), implemented as described in Eq. 5, was used to compare arousal scoring performance in different sleep stages.

$$FPR = \frac{FP}{FP + TN} \quad \#(12)$$

The FPR measure uses TN instead of TP to enable measuring performance in wake, where there are no TPs as arousals can only have onset during sleep by definition. The FPR metric for arousal events is defined as

$$FPR_{event} = \frac{FP_{event}}{FP + TN} \quad \#(13)$$

where the *event* subscript refers to number of events rather than 1 second bins.

To measure accuracy of classifying wake, TNs are important as they measure correctly identified sleep. Wake scoring performance is measured by recall, specificity and accuracy. The performance metrics of specificity and accuracy are defined by Eq. 7 and 8.

$$Specificity = \frac{TN}{TN + FP} \quad \#(14)$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad \#(15)$$

The probability postprocessing employ a set of thresholds (T_{ar}, T_w) to binarize these probabilities. A threshold of 0.5 intuitively makes sense, as it distinguishes between which class is more likely based on a cross entropy cost. However, this may not be optimal since performance is measured mostly by F1 score and events are connected and discarded during the postprocessing step. For this reason, the two thresholds, T_{ar} and T_w, were calculated using the validation set by maximizing arousal event F1 score and wake accuracy, respectively.

As described above, a test set of 996 PSGs derived of multiple cohorts was used to calculate an unbiased estimate of MAD performance. We also used these data to evaluate performance in the cohorts and across sleep stages. This analysis gives insight as to if the model has any problems processing specific parts of the dataset. A second test set of 30 PSGs from each of the WSC and SSC datasets was used to compare performance of the model against that of multiple individual human scorers. PSG studies in this dataset had been scored five times using a pool of nine human scorers. In this analysis, an individual scorer and the model's performance were evaluated based on a pseudo-consensus scoring as gold standard, which

was established based on a majority vote of the four remaining scorers in a leave-one-out scheme. The majority vote was defined as at least two scorers agreeing on an event i.e. an agreement of 50 % or more. The comparison was carried out by iterating through the five scorers for each PSG, creating a pseudo-consensus from the remaining four and calculating F1 score for both the scorer and model based on that consensus. Thereby, as each of the nine technicians had scored either 33 or 34 of these PSGs, the average F1 score was compared between individual scorers and the model. This test evaluates whether the model performs at least as well or better than a human expert when scoring new unseen data. Thereby, we can evaluate whether the model can be used instead of human scoring in clinical and research settings, which is the main motivation of the study.

Bias of annotations by human scorers within the 30-second epoch was investigated as it is suspected visualization tools, used to score PSGs in 30-second epochs, can introduce bias in human annotations. Automatic systems can avoid this type of bias as the 30-second epoch is not the basis of the model.

2.8 Relationship of Daytime Sleepiness with PSG arousals and wake events as detected using MAD

The MSL on an MSLT is considered a gold standard to measure daytime sleepiness objectively. We used a total of 1,447 PSGs with associated MSLT from the WSC to analyze the effect of sleep disruption as measured with the MAD on daytime sleepiness. The 1,447 PSG-MSLTs were performed in 873 subjects and Generalized Estimating Equations (GEE) statistics were used to account for repeated samples (see below).

To address properly the effects of arousals of all durations (i.e. detection of microarousals less than 15 sec and more than 15 sec leading a full-blown wake epoch), arousals and wake scored by our detector were combined to measure all sleep disruption and wakefulness. This global arousal measure of any transition to wake was computed as the union of predicted arousal and wake. As arousals are defined by the AASM to be preceded by at least 10 seconds of stable sleep, the new arousal measure was further postprocessed to combine wakefulness events preceded by fewer than 10 seconds of sleep.

We next examined the effect of arousals in the context of SDB and leg movement (LM) occurrences. The WSC is a thoroughly investigated cohort, and both manual scoring and various prior event detectors have been proposed to score SDB events, blood oxygen desaturations and LMs, with high correlations with manually scored events. For SDB events, we applied the algorithm described by Koch et al. (2017), which is a rule-based algorithm adapted to model the AASM 2012 manual (Berry et al. 2018) that detects 10 sec breathing disturbances with and without hypoxia events independently of any detected arousal (Breathing Disturbance Index, BDI, composed of BDI-Hypoxia and BDI non Hypoxia). LMs were scored using the Stanford PLM automatic detector (S-PLMAD) proposed by Moore et al. (2014), which is also a rule-based algorithm adapted to the AASM scoring guidelines (Berry et al. 2018). The S-PLMAD uses adaptive filtering of ECG artifacts and additional rules to account for specific noise types., and scores PLM independently of the presence of arousals. For more information about these detectors see original publications in (Moore et al. 2014; Koch et al. 2017).

2.8.1 Coupling of automatically detected arousals to disturbed breathing or leg movement events—The distribution of new onset arousal/wake occurrences relative to SDB or LM events were investigated. These distributions were computed with histograms over the 1447 PSGs from the WSC by counting the occurrence of events relative to onset or offsets of other events. Coupling rules were initially based on the AASM manual but revised if data suggested improvements could be made.

As illustrated in Fig. 4, and reported in Moore et al. (2014) and others (Manconi et al. 2014; Aritake et al. 2015), we found that the original AASM scoring definition which suggested disregarding LM within 0.5 sec of SDB events was not adequate as movements typically peaks a few seconds after the end of an SDB event. LMs secondary to SBDs were discarded if the onset of a LM occurred 15 seconds preceding a BD or −5 to 10 seconds at the offset of a SBD event. PLMs were then scored in agreement with the AASM manual, as at least four LMs not secondary to BDs with an inter-movement-interval between 5 and 90 seconds [5].

Breathing disturbances were coupled with arousals and desaturations based on similar distributions of time-locked events. The AASM manual do not state any objective rules for coupling these events, therefore the rules are based on the distributions visualized in Fig. 5 (a) and (b). Arousals were coupled with BDs if the arousal onset in the interval of −5 to 10 seconds at the BD offset. Desaturations were coupled with BDS if the peak desaturation was between 5 and 35 seconds after the BD offset. PLMs were also coupled with arousals if their onsets were within a range of 5 seconds. The AASM manual uses a rule of minimum 0.5 seconds overlap of these events, but this rule is based on the distribution seen in Fig. 5 (c).

Based on these coupling rules, a set of PSG biomarkers was computed, which are described in Table 2. These were then used as covariates to explore how they best predict daytime sleepiness using the MSLT. In these analyses all variables were $\log_2$-transformed to ensure normal distribution.

2.8.2 Statistical Analysis—We first explored collinearity of the $\log_2$transformed variables described in Table 2 using Pearson correlation coefficient statistics. Next, the effect of each variable on MSLT was examined using generalized estimating equations (GEE), a statistical technique that maximizes power for cross-sectional analyses that have repeated measures in some subjects (Zeger and Liang 1986) to estimate robust standard errors and allow for usage of repeated measurements. GEE was implemented in GEEQBOX (Ratcliffe and Shults 2008), a MATLAB toolbox for GEE, using the identity link function and the first order autoregressive correlation structure. This correlation structure is well suited for repeated measurements that are roughly evenly spaced in time (Ratcliffe and Shults 2008), which is the case for the WSC as repeated measurements for subjects are conducted as 4-year follow ups. The *sandwich estimator* (Zeger and Liang 1986) was used to provide a robust estimate of the standard errors.

Finally, stepwise linear regression was implemented to find an optimal set of variables that provided most information about the MSL. Stepwise linear regression initially fits the set of variables, which we wanted to adjust for, namely age, BMI, sex, habitual sleep duration,

minutes of sleep the two nights preceding the MSLT, and predicted WASO. Then, from a set of explanatory variables, it proceeds to iteratively add the variable with lowest p-value larger than 0.01 and remove the largest p-value larger than 0.05 in the linear model. The stepwise linear regression algorithm terminates when no parameters can be added or removed. For stepwise linear regression a subset of 873 PSGs was used, as repeated measurements could not be included. The best regression model was determined based on the adjusted coefficient of determination R^2, which is given by

$$\bar{R}^2 = 1 - \frac{SS_{res}/(n-p-1)}{SS_{tot}/(n-1)} \quad \#(16)$$

where SS_{res} is the sum of squares of the residuals, SS_{tot} is the total sum of squares, n is the number of observations, and p is the number of explanatory variables. When additional explanatory variables are added to a model, the adjusted R^2 will only increase if R^2 increases by more than expected by chance. Of note, because many variables are colinear, the set of variables selected by the model may not necessarily represent the ideal best variable if a larger dataset was analyzed. It nonetheless helped us validate key features of the detector.

Code availability: The Matlab and Python code for the MAD detector is available on GitHub at: https://github.com/abrinkk/multimodal-arousal-detector. On average, the run time for a single PSG is 92.8 seconds on an Intel® Core™ i7–8750H CPU @ 2.2GHz (12 CPUs) for preprocessing and NVIDIA GeForce GTX 1080 for model predictions.

3 Results

3.1 Network Performance

3.1.1 Probability Threshold—The effect of changing probability thresholds for arousal and wake detection was examined by varying threshold in steps of 0.025. For each threshold, precision, recall, and F1 score were calculated for arousal events while sensitivity, recall, and accuracy were calculated for the wake predictions. Arousal predictions in 1 second bins will from this point referred to as *arousal samples* for convenience. The thresholds that maximize arousal event F1 score and wake accuracy were selected, which are $T_{ar} = 0.225$ and $T_w = 0.45$. The precision-recall (PR) curves for arousal predictions is shown in Fig. 6 and the receiver operating characteristic (ROC) curve for wake are shown in Fig. 7.

3.1.2 Test Performance—The model was evaluated on the test set of 996 PSGs from MrOS, CFS, and WSC. Predictions were postprocessed using the optimal thresholds $T_{ar} = 0.225$ and $T_w = 0.45$. Fig. 8 show an example of arousal and wake predictions over a full night's PSG, and Fig. 9 shows predictions of a segment of the same PSG in a 60 seconds window at an arousal prediction.

Table 3 shows the classification performance on the different cohorts as well as all test data. Arousal events were detected with an F1 score of 0.76 over all test data. On the WSC data, arousal events were detected with a lower F1 score of 0.70, which is caused by lower precision compared to the MrOS and CFS data. The slightly lower performance in the WSC is likely caused by two factors: the fact that this data was unseen during training

and the fact arousals in this dataset were only annotated as onsets, hence all annotated arousals were assumed to have a duration of 3 seconds, as this is the minimum duration of AASM arousals. Wake was predicted with an accuracy of 0.95 across all test data. The wake prediction accuracy was 0.93 on the WSC data, which suggests that the model works almost as well on data from sources unseen during training.

Variation in model's arousal event scoring performance is displayed in Fig. 10 as a scatter plot, which shows a good performance for the clear majority of PSG recordings, but also a set of PSGs with poor performance In Fig. 10, the size of each dots is proportional to the number of arousals detected in a single PSG, so that small dots that have poor performance could reflect limited sample size for performance evaluation. Through visual inspection it was observed that there were PSGs with mostly missing arousal target labels or target labels that appears random. This suggests that a substantial part of the error on PSG recordings with very poor performance is caused by human error.

Bias within 30-second epochs in human scoring was further investigated by computing the average arousal annotation frequency for each 30-second epoch. Arousals and their definition according to the AASM guidelines (Berry et al. 2018) is completely unrelated to the 30-second epoch. However, as shown in Fig. 11, arousal annotations in the MrOS and CFS cohort show a clear bias toward the central and late part of the 30-second epochs. This bias is likely to have been introduced as a result of software that visualize data in 30-second epochs. As a result of this edge bias, the arousal sample performance was lower in the beginning and end of each 30-second epoch. The arousal sample F1 score for the CFS and MrOS test data in the [10 −25] second interval of each 30-second epoch 0.71.

The dependency of the model's classification performance on sleep stage has also been examined. This analysis was performed by comparing the performance to manually scored sleep stages. Table 4 and 5 shows the performance in the different sleep stages for arousals and wake/sleep, respectively.

Table 4 shows that arousal events are detected well in sleep, although with a relatively poorer performance in N1. The FPR for arousal events is the lowest in wake and highest in N1. The performance metrics for the arousal samples also show a good performance in all stages, but with a slight decrease in performance in N1.

The wake/sleep accuracy displayed in Table 5, showing that sleep is detected well in N2, N3, and REM, while accuracy is lower in N1. As N2, N3, and REM sleep is highly distinctive from wake, these are scored as sleep confidently by the model.

3.1.3 Comparison to Multiple Scorers—The performance of the model was compared to multiple scorers on a dataset of 60 PSGs. The model and human scorer predictions were evaluated with respect to a pseudo-consensus of multiple scorers. The comparison was based on the arousal event F1 score. The results of this test are presented in Table 6.

As seen in Table 6, the model predicts arousals with a significantly higher average F1 score in comparison to 5 scorers, while there is no significant difference to the remaining

4 scorers. Further, the model outperforms the average scorer with a difference in F1 score of 0.09. Scorer I performed particularly bad in comparison, however the model also outperforms scorer A – H on average. However, it should be noted that each human scorer to model comparison is based on the subset of data that the human scored. This indicates that our automatic arousal scoring system performs substantially better than human scorers.

3.2 Statistical Analysis of Daytime Sleepiness

The statistical analysis was performed using a combined measure of arousal and wake, which does not discriminate based on the 15 seconds threshold used in current scoring rules (Berry et al. 2018). The duration of predicted arousals has the distribution shown in Fig. 12. The distribution peaks with an arousal duration of 9 seconds and decreases exponentially onward. The data does not provide any justification as to split the arousal measure at 15 seconds. Therefore, in the analysis of daytime sleepiness, we analyze PSGs with the combined measure of arousal and wake as a single type of event.

Fig. 13 shows a correlation matrix of $\log_2$-transformed sleep variables, computed using Pearson correlation coefficient statistics.

Results show that the BDI is highly correlated to all variables related to breathing disturbances, while the other breathing disturbance variables are correlated in a more complex pattern. As reported in Kock et al. (Koch et al. 2017), breathing disturbances associated with hypoxia are only weakly correlated with those not associated with hypoxia, thus two clusters are apparent. Arousal-associated SDB (Ar-BDI) encompass almost all SDB events, which indicate that our previously reported non-hypoxia-BDI (NonHyp-BDI) really corresponds to SDB events associated with arousals but no desaturation (Ar-NonHyp-BDI). The correlation between Ar-Hyp-BDI and NonAr-NonHyp-BDI was shown to be small ($r = 0.052$, $p = 0.13$), suggesting that subjects can be affected by either subtype of sleep disordered breathing.

The sleep variables relating to PLMs were also all highly correlated. The Me-Spon-Ar-Dur variable has no correlation to either breathing disturbance variables, PLMI or the ArI.

A GEE was used to estimate model parameters for each sleep variables at a time as seen in Table 7. These estimates were adjusted for age, BMI, sex, habitual sleep duration, minutes of sleep the two nights preceding the MSLT, and predicted WASO. The predicted WASO was included as it and arousal metrics are positively correlated (see Fig. 13), while WASO and arousal metrics have an inverse relation to MSLT as displayed in Table 7. Due to this interaction it was necessary to add it to the linear models to show the effect of arousals.

The results in Table 7 show that the ArI, Me-Spon-Ar-Dur, and most breathing disturbance variables have a significant ($p < 0.05$) negative effect on the MSL.

A series of MSL models was fitted with stepwise linear regression to provide information as to which sleep variables provides most independent information about the MSL. Thereby, these linear models investigate if the predicted arousal measure add additional information beyond the LM and BD measures. These models were fitted using the first available visit of

each subject, which reduces the number of observations to 873. The results of these models are summarized in Table 9.

Model 1 in Table 9 shows the explanatory power of the variables that each model is adjusted for. Breathing disturbances associated with arousals seems to have a stronger effect on the MSL as indicated by model 2 to 6. The Ar-Hyp-BDI and NonAr-NonHyp-BDI were shown in model 7 to be independent measures of breathing disturbances that both affects the MSL. Model 8 and 9 shows that the ArI and NonAr-BDI have an independent effect on MSL. The results of model 10 and 11 show that the inclusion of Me-Spon-Ar-Dur further strengthens the MSL models. The model with the most explanatory power as measured by the adjusted R^2 included the sleep variables NonAr-NonHyp-BDI, ArI, and Me-Spon-Ar-Dur. Surprisingly, Ar-PLMI was not associated with MSL in any model.

4 Discussion

4.1 Performance of MAD: Comparison to Multiple Scorers

The arousal event classification performance of our model was compared to that of nine individual scorers by evaluating predictions with respect to a pseudo-consensus. The pseudo-consensus was based on majority voting from the remaining four human scorers is not expected to be near perfect as the discrepancies between scorers is so large, but it is assumed to be good enough to justify the comparison to individual scorers. The comparison showed that the model, in terms of F1 score, significantly outperformed five of nine individual human scorers, while there was no significant difference to the remaining four. Further, the model outperformed the average human scorer with a difference in F1 score of 0.09.

The pseudo-consensus is an average estimate arousal scoring. The model's decisions are optimized using a cross entropy loss function, which attempts to converge to the average decision of expert technicians. Thereby, the model has learned to model the average technician's arousal scoring better or comparable to that of an individual expert technician.

The significance of this comparison is further emphasized by the fact that the performance of the model is unbiased, as the data from the WSC and SSC were unseen during training. The best performing human scorer achieved a F1 score of 0.71, which also suggests that the F1 score of 0.76 achieved on the larger dataset of 996 PSGs is satisfactory. In brief, our detector was able to score arousals better than most scorers.

The distribution of arousals annotated by humans was also found to depend on the 30-second epoch for the MrOS and CFS data, as arousals were more frequently annotated in the central to the late proportion of the 30-second epoch. This bias was not shown by the proposed model, which is a result of the model not viewing and processing data in 30-second epochs.

4.2 Comparison to Previous Methods

In Table 10 various methods for automatic detection of arousals are summarized. However, comparing the proposed arousal detecting system to existing published methods is difficult due to differences in used data, number of human scorers, scoring unit, performance metric

etc. The performance of the proposed method has achieved a higher F1 score than the methods validated in comparable datasets (Alvarez-Estevez and Fernández-Varela 2019; Olesen et al. 2019). The method that reported the highest performance was published by Sorensen et al. (2012) with an F1 score of 0.87 using cross-validation on 24 subjects. However, we believe the model would be unlikely to generalize on unseen data as it is trained on annotations from a single human scorer. Furthermore, this high performance suggest overfitting to this single scorer, as the best human scorer in this study only achieved a F1 score of 0.71. In general, the proposed method stands out from existing methods on the points of it being fully automatic, showing a robust performance on a very large dataset, and scoring both arousals and wake in 1-second intervals.

Wake was predicted with an overall accuracy of 0.95 in 1 second bin (epochs), which is different than the standardized 30-second epoch. The sleep/wake accuracy of the model is high, but it is inflated to some degree due to long periods of wake between recording start time and sleep onset in MrOS and CFS. Accuracy could have been measured between annotations of lights off/on, however wakefulness during time of lights on was also considered as wake. This also has the effect of simplifying the preprocessing and does not require manual annotations. The sleep/wake prediction performance of the model can be compared to the performance of published methods for sleep staging. The studies by Biswal et al. (2018) and Sun et al. (2017) both used a test set of 1000 PSGs similar to this project and achieved a wake/sleep accuracy of 0.937 and 0.929, respectively. It should be noted that the wake accuracy from these studies was estimated from normalized confusion matrices using the distribution of sleep stages of data from the WSC. Stephansen et al. (2018) and Supratak et al. (2017) achieved a higher wake/sleep accuracy of 0.967 and 0.961, respectively. The wake accuracies reported in these studies were measured using smaller test sets of 70 and 82 PSG. The wake accuracy of the proposed system is highly competitive and is at the level of state-of-the-art sleep staging methods, but it is difficult to infer which method clearly perform best.

4.3 Statistical Analysis of Daytime Sleepiness in relation to arousal

Regression analysis revealed a significant association between the breathing disturbance sleep variables and the MSL on the MSLT, except for NonAr-Hyp-BDI. The NonAr-Hyp-BDI is expected to be highly noisy as most breathing disturbances with hypoxia are suspected to provoke an arousal. The effect of breathing disturbances on MSL has previously been demonstrated by multiple studies (Roehrs et al. 1989; Martin et al. 1997; Koch et al. 2017). The stepwise linear regression model showed that breathing disturbances can be described with two independent variables Ar-Hyp-BDI and NonAr-NonHyp-BDI, which agrees with previous findings by Koch et al. (2017), who showed that Hyp-BDI and NonHyp-BDI are independent measures of sleep disordered breathing in the same dataset. The significance of the breathing disturbances without hypoxia or arousal is slightly counter-intuitive as the breathing disturbances do not provoke any measured physiological changes. A possible explanation is that these types of breathing disturbances may result or be a result of subcortical disturbances, which impair the restorative effect of sleep.

More surprisingly, sleep variables related to PLMs had no effect on MSL, even in the presence of an associated arousal. Chervin et al. (2001) found a weak, but significant association between PLMs and MSL. In this sample, PLMs with and without arousals seem to either have a negligible or a small effect on MSL. It should be noted that a p-value above 0.05 simply implies a lack of evidence to reject the null hypothesis, rather than showing that the sleep variable has no effect on daytime sleepiness. The regression coefficient as well and standard error is also highly dependent on the variables that the model is adjusted for, the test therefore only provides information about the relation to MSL based on the knowledge of the variables adjusted for. Together with the finding described above suggesting that breathing disturbances without EEG arousal is associated with sleepiness, this observation may reflect the fact cortical activations are not 100% associated with disturbances in sleep homeostasis.

The ArI predicted by the proposed method exhibited a strong association to the MSL, with an average decrease in the MSL of 40 seconds for each doubling in the ArI ($\beta = -0.67$, $p = 0.0075$). This result is in concordance with similar regression models of MSL (Roehrs et al. 1989; Martin et al. 1997; Leng et al. 2003). Additionally, the best stepwise regression model shown in Table 9 (model 10, R^2 (adj) = 0.151) retained both arousal and breathing disturbance measures, which shows that the arousal measure adds value to the prediction of daytime sleepiness beyond the remaining sleep biomarkers. The regression analysis also included median arousal duration with associations of BDs and PLMs. The Me-Spon-Ar-Dur variable was the only significant arousal duration measure ($\beta = -1.5$, $p = 0.0028$), while Spon-ArI was non-significant. This suggests again that all arousals (spontaneous versus associated with SDB or PLMs) are not created equal with respect to their effects on daytime sleepiness. Bigger samples would be needed to examine this question more thoroughly. These results suggest that the proposed scoring system have clinical applications, as arousal variables show a significant link to MSL. The significance of the Me-Spon-Ar-Dur variable further suggests that combining brief and long arousals constitutes a meaningful measure in describing sleep structure.

The regression analysis presented in this chapter incorporated scoring of arousals, wake, breathing disturbances, blood oxygen desaturations, and leg movements, but could be expanded by including sleep stages and lights on/off annotations. Furthermore, label probabilities as scored by the model (without postprocessing) could be used directly as they convey more information. However, it is already difficult to compare significant explanatory variables due to the small effect size of each variable. This could partly be sorted by simply using much more data, although this is difficult due to the lack of available databases that contains nocturnal PSGs with associated MSLT data. Alternatively, if the sole purpose is to predict MSL, then a deep learning framework could be implemented to directly model MSL from an entire PSG. Unfortunately, however, large datasets with PSGs and MSLTs are nonexistent, and to our knowledge the WSC is the only available large sample with such data.

5 Conclusion

We describe a fully automatic method that concurrently detect arousals and wake using convolutional and LSTM neural networks. Our model was trained on a dataset of 2,889 PSGs and performance evaluated on 1,026 PSGs. These PSGs came from four distinct cohorts that used different hardware for PSG recordings, ensuring robustness of the MAD detector. The model predicted arousal events with a precision of 0.72, recall of 0.81, and a F1 score of 0.76. Wake was predicted on the test set in 1-second intervals with an accuracy of 0.95. The arousal event scoring performance of the model was compared to that of 9 individual human scorers with respect to a pseudo-consensus scoring in 30 PSGs from both the Wisconsin Sleep Cohort and Stanford Sleep Cohort. The comparison showed that the model significantly outperformed 5 of 9 human scorers, while no significant differences were found to the remaining 4 scorers. The proposed method further achieved a higher F1 score than previously published arousal detectors that have been validated in comparable (and typically smaller) datasets.

The arousal index predicted by MAD on an additional 1,447 PSGs from the Wisconsin Sleep Cohort was compared to an associated MSLT through statistical analysis. The arousal index showed a significant association to the mean sleep latency with an average decrease in mean sleep latency of 40 seconds for each doubling in the arousal index ($\beta = -0.67$, $p = 0.0075$). An increase in the median duration of spontaneous arousal was also associated with a decrease in the mean sleep latency ($\beta = -1.5$, $p = 0.0028$). The model predictions were correlated with the mean sleep latency, showing that the model has clinical applications as a fully automated scoring tool that predicts next day sleepiness.

Acknowledgment:

Andreas Brink-Kjaer, Alexander Neergaard Olesen and Emmanuel Mignot were partially funded by the Klarman Family Foundation. Poul Jennum is supported by internal funding from Rigshospitalet. Additional support was provided to Andreas Brink-Kjaer by the Marie and M.B. Richters, Vera and Carl Johan Michaelsens, Froeken Marie Maanssons, Oticon, Dansk Tennis, Julie Damms, and IDA foundations.

Wisconsin Sleep Cohort polysomnography data collection was supported by the US National Institutes of Health (NIH) grants 1R01AG036838, R01HL62252, and 1UL1RR02501.

The MrOS Study was funded by: U01s AG027810, AG042124, U01 AG042139,40, 43 and 45, AG042168, U01 AR066160, UL1 TR000128 and R01s HL070837-39, 40-42, 48.

Furthermore, we would like to thank The National Sleep Research Resource for offering free access to large collections of data. The NSRR is supported by Grant Number HL114473 from the National Heart, Lung, and Blood Institute, NIH.

Declaration of interest:

Dr. Mignot has received funding from Jazz pharmaceutical and has shares in Rythm, a company doing a consumer portable EEG device, and Inoxia/Orexia, a company developing orexin agonists, but these involvements are unrelated to this project. Katie L. Stone has received grant funding from Merck, but this is unrelated to this project and they are not involved.

References

Alvarez-Estevez D, Fernández-Varela I. Large-scale validation of an automatic EEG arousal detection algorithm using different heterogeneous databases. Sleep Med. 2019;57:6–14. [PubMed: 30878899]

Andlauer O, Moore H, Jouhier L, Drake C, Peppard PE, Han F, et al.Nocturnal Rapid Eye Movement Sleep Latency for Identifying Patients With Narcolepsy/Hypocretin Deficiency. JAMA Neurol. 2013;70:891. [PubMed: 23649748]

Aritake S, Blackwell T, Peters KW, Rueschman M, Mobley D, Morrical MG, et al.Prevalence and associations of respiratory-related leg movements: the MrOS sleep study. Sleep Med. 2015;16:1236–44. [PubMed: 26429752]

Berry R, Albertario CL, Harding SM, Lloyd RM, Plante DT, Qyan SF, et al.The AASM Manual for the scoring of sleep and associated events : rules, terminology and technical specifications. Version 2.Darien, IL: American Academy of Sleep Medicine: American Academy of Sleep Medicine; 2018.

Biswal S, Sun H, Goparaju B, Westover MB, Sun J, Bianchi MT. Expert-level sleep scoring with deep neural networks. J Am Med Informatics Assoc. 2018;25:1643–50.

Blackwell T, Yaffe K, Ancoli-Israel S, Redline S, Ensrud KE, Stefanick ML, et al.Associations Between Sleep Architecture and Sleep-Disordered Breathing and Cognition in Older Community-Dwelling Men: The Osteoporotic Fractures in Men Sleep Study. J Am Geriatr Soc. 2011;59:2217–25. [PubMed: 22188071]

Blank JB, Cawthon PM, Carrion-Petersen M Lou, Harper L, Johnson JP, Mitson E, et al.Overview of recruitment for the osteoporotic fractures in men study (MrOS). Contemp Clin Trials. 2005;26:557–68. [PubMed: 16085466]

Bonnet M, Carley D, Carskadon M, Easton P, Guilleminault C, Harper R, et al.EEG Arousals: Scoring Rules and Examples. Sleep. 1992;15:173–173. [PubMed: 11032543]

Bonnet MH, Doghramji K, Roehrs T, Stepanski EJ, Sheldon SH, Walters AS, et al.The scoring of arousal in sleep: reliability, validity, and alternatives. J Clin Sleep Med. 2007;3:133–45. [PubMed: 17557423]

De Carli F, Nobili L, Gelcich P, Ferrillo F. A Method for the Automatic Detection of Arousals During Sleep. Sleep. 1999;22:561–72. [PubMed: 10450591]

Chervin RD. Periodic Leg Movements and Sleepiness in Patients Evaluated for Sleep-disordered Breathing. Am J Respir Crit Care Med. 2001;164:1454–8. [PubMed: 11704595]

Cho SP, Lee J, Park HD, Lee KJ. Detection of Arousals in Patients with Respiratory Sleep Disorders Using a Single Channel EEG. In: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference. IEEE; 2005. p. 2733–5.

Dahl GE, Sainath TN, Hinton GE. Improving deep neural networks for LVCSR using rectified linear units and dropout. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE; 2013. p. 8609–13.

Dean DA, Goldberger AL, Mueller R, Kim M, Rueschman M, Mobley D, et al.Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource. Sleep. 2016;39:1151–64. [PubMed: 27070134]

Fernández-Varela I, Alvarez-Estevez D, Hernández-Pereira E, Moret-Bonillo V. A simple and robust method for the automatic scoring of EEG arousals in polysomnographic recordings. Comput Biol Med. 2017a;87:77–86. [PubMed: 28554078]

Fernández-Varela I, Hernández-Pereira E, Álvarez-Estévez D, Moret-Bonillo V. Combining machine learning models for the automatic detection of EEG arousals. Neurocomputing. 2017b;268:100–8.

Findley LJ, Unverzagt ME, Suratt PM. Automobile Accidents Involving Patients with Obstructive Sleep Apnea. Am Rev Respir Dis. 1988;138:337–40. [PubMed: 3195832]

Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Int Conf Artif Intell Stat. 2010;9:249–56.

Halasz P, Terzano M, Parrino L, Bodizs R. The nature of arousal in sleep. J Sleep Res. 2004;13:1–23.

He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition. 2016. p. 770–8.

He P, Wilson G, Russell C. Removal of ocular artifacts from electro-encephalogram by adaptive filtering. Med Biol Eng Comput. 2004;42:407–12. [PubMed: 15191087]

Hobson JA. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Electroencephalogr Clin Neurophysiol. 1969;26:644.

Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997;9:1735–80. [PubMed: 9377276]

Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arxiv:150203167.2015;

Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arxiv:14126980.2014;

Koch H, Schneider LD, Finn LA, Leary EB, Peppard PE, Hagen E, et al.Breathing Disturbances Without Hypoxia Are Associated With Objective Sleepiness in Sleep Apnea. Sleep. 2017;40:zsx152.

Kryger MH, Roth T, Dement WC. Principles and Practice of Sleep Medicine (Fifth Edition). 5. Philadelphia: W.B. Saunders; 2011.

LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44. [PubMed: 26017442]

LeCun YA, Bottou L, Orr GB, Müller K-R. Efficient BackProp. In: Montavon G, Orr GB, Müller K-R, editors. Springer, Berlin, Heidelberg; 2012. p. 9–48.

Leng PH, Low SY, Hsu A, Chong SF. The Clinical Predictors of Sleepiness Correlated with the Multiple Sleep Latency Test in an Asian Singapore Population. Sleep. 2003;26:878–81. [PubMed: 14655923]

Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to Diagnose with LSTM Recurrent Neural Networks. arxiv:151103677v7.2015;

Magalang UJ, Chen N-H, Cistulli PA, Fedson AC, Gíslason T, Hillman D, et al.Agreement in the Scoring of Respiratory Events and Sleep Among International Sleep Centers. Sleep. 2013;36:591–6. [PubMed: 23565005]

Manconi M, Zavalko I, Bassetti CL, Colamartino E, Pons M, Ferri R. Respiratory-Related Leg Movements and Their Relationship with Periodic Leg Movements During Sleep. Sleep. 2014;37:497–504. [PubMed: 24587572]

Martin SE, Engleman HM, Kingshott RN, Douglas NJ. Microarousals in patients with sleep apnoea/hypopnoea syndrome. J Sleep Res. 1997;6:276–80. [PubMed: 9493529]

Moore H, Leary E, Lee S-Y, Carrillo O, Stubbs R, Peppard P, et al.Design and Validation of a Periodic Leg Movement Detector. Penzel T, editor. PLoS One. 2014;9:e114565. [PubMed: 25489744]

Olesen AN, Chambon S, Thorey V, Jennum P, Mignot E, Sorensen HBD. Towards a Flexible Deep Learning Method for Automatic Detection of Clinically Relevant Multi-Modal Events in the Polysomnogram. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2019. p. 556–61.

Olesen AN, Jennum P, Peppard P, Mignot E, Sorensen HBD. Deep residual networks for automatic sleep stage classification of raw polysomnographic waveforms. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2018. p. 1–4.

Orwoll E, Blank JB, Barrett-Connor E, Cauley J, Cummings S, Ensrud K, et al.Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study — A large observational study of the determinants of fracture in older men. Contemp Clin Trials. 2005;26:569–85. [PubMed: 16084776]

Popovic D, Khoo M, Westbrook P. Automatic scoring of sleep stages and cortical arousals using two electrodes on the forehead: validation in healthy adults. J Sleep Res. 2013;23:211–21. [PubMed: 24313630]

Ratcliffe SJ, Shults J. GEEQBOX : A MATLAB Toolbox for Generalized Estimating Equations and Quasi-Least Squares. J Stat Softw. 2008;25:1–14.

Redline S, Tishler PV., Schluchter M, Aylor J, Clark K, Graham G. Risk Factors for Sleep-disordered Breathing in Children. Am J Respir Crit Care Med. 1999;159:1527–32. [PubMed: 10228121]

Redline S, Tishler PV, Tosteson TD, Williamson J, Kump K, Browner I, et al.The familial aggregation of obstructive sleep apnea. Am J Respir Crit Care Med. 1995;151:682–7. [PubMed: 7881656]

Roehrs T, Zorick F, Wittig R, Conway W, Roth T. Predictors of Objective Level of Daytime Sleepiness in Patients with Sleep-Related Breathing Disorders. Chest. 1989;95:1202–6. [PubMed: 2721252]

Saper CB, Scammell TE, Lu J. Hypothalamic regulation of sleep and circadian rhythms. Nature. 2005;437:1257–63. [PubMed: 16251950]

Shahrbabaki SS, Dissanayaka C, Patti CR, Cvetkovic D. Automatic detection of sleep arousal events from polysomnographic biosignals. In: 2015 IEEE Biomedical Circuits and Systems Conference (BioCAS). IEEE; 2015. p. 1–4.

Shmiel O, Shmiel T, Dagan Y, Teicher M. Data mining techniques for detection of sleep arousals. J Neurosci Methods. 2009;179:331–7. [PubMed: 19428545]

Sorensen GL, Jennum P, Kempfner J, Zoetmulder M, Sorensen HBD. A Computerized Algorithm for Arousal Detection in Healthy Adults and Patients With Parkinson Disease. J Clin Neurophysiol. 2012;29:58–64. [PubMed: 22353987]

Stephansen JB, Olesen AN, Olsen M, Ambati A, Leary EB, Moore HE, et al.Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. Nat Commun. 2018126;9:5229. [PubMed: 30523329]

Sugi T, Kawana F, Nakamura M. Automatic EEG arousal detection for sleep apnea syndrome. Biomed Signal Process Control. 2009;4:329–37.

Sun H, Jia J, Goparaju B, Huang G-B, Sourina O, Bianchi MT, et al.Large-Scale Automated Sleep Staging. Sleep. 2017;40:zsx139.

Supratak A, Dong H, Wu C, Guo Y. DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. IEEE Trans Neural Syst Rehabil Eng. 2017;25:1998–2008. [PubMed: 28678710]

Wallant DC, Muto V, Gaggioni G, Jaspar M, Chellappa SL, Meyer C, et al.Automatic artifacts and arousals detection in whole-night sleep EEG recordings. J Neurosci Methods. 2016;258:124–33. [PubMed: 26589687]

Young T, Finn L, Peppard PE, Szklo-Coxe M, Austin D, Nieto FJ, et al.Sleep Disordered Breathing and Mortality: Eighteen-Year Follow-up of the Wisconsin Sleep Cohort. Sleep. 2008;31:1071–1078. [PubMed: 18714778]

Zeger SL, Liang K-Y. Longitudinal Data Analysis for Discrete and Continuous Outcomes. Biometrics. 1986;42:121. [PubMed: 3719049]

Highlights

- New method that detects cortical arousals and wake in polysomnographic recordings.
- The method performs comparable or better than human experts for arousal detection.
- An increase in arousal frequency is associated with decrease in mean sleep latency.

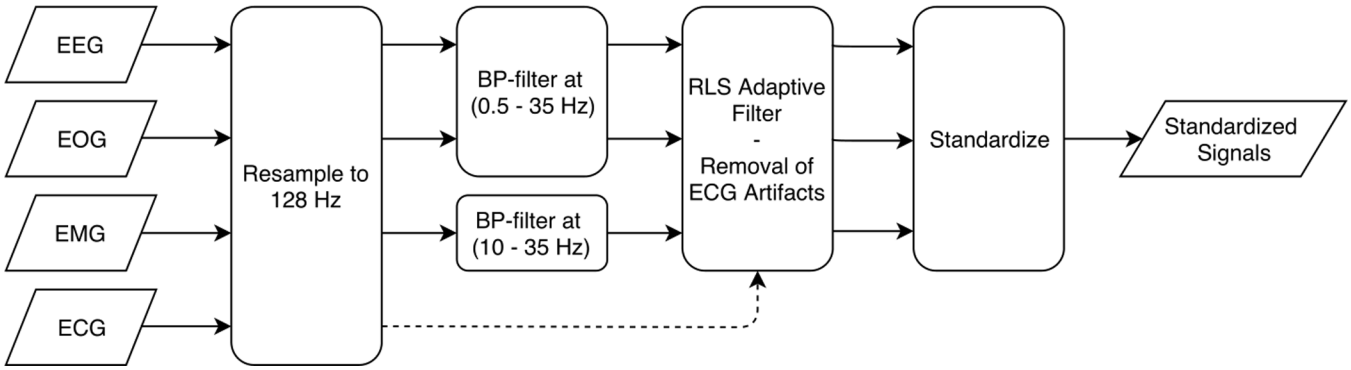

Figure 1:

Schematic illustration of the proposed preprocessing method. Input signals are resampled, band pass filtered, the RLS adaptive filter is applied, and finally signals are standardized.

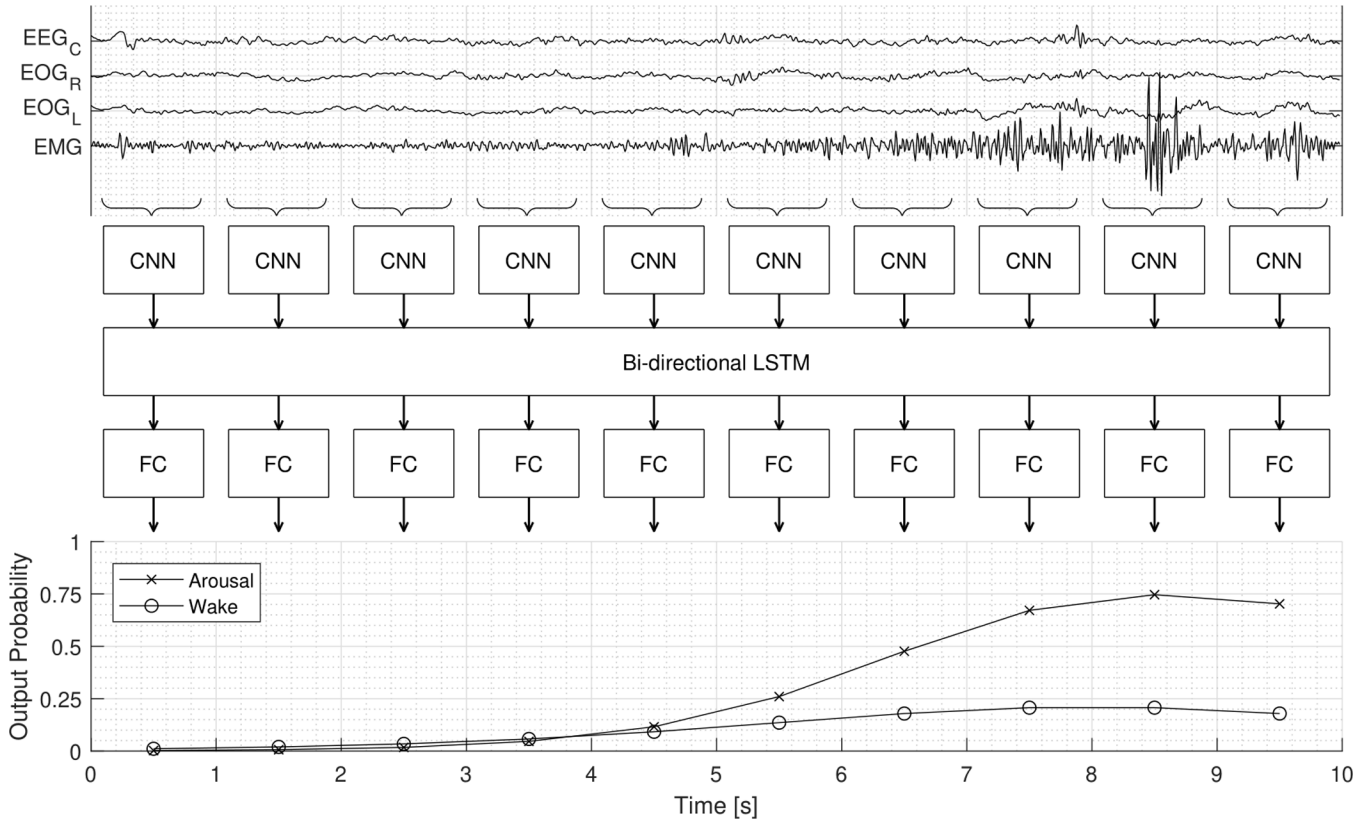

Figure 2:

Illustration of the process to which the preprocessed signals are reshaped and fed through the proposed MAD network. The top plot shows a 10 second window of signals with an arousal annotated at 5 seconds. Arousal and wake probabilities are computed from the reshaped input being fed through the network structure consisting of CNN (convolutional neural network), LSTM (Long Short-Term Memory), and FC (fully connected) layers. The three network types (CNN, LSTM, FC) each have a main function, (1) extraction of features describing the 1 second segment of the signals, (2) computing features with temporal information (e.g. frequency change), (3) mapping temporal features to predict wake and arousal labels.

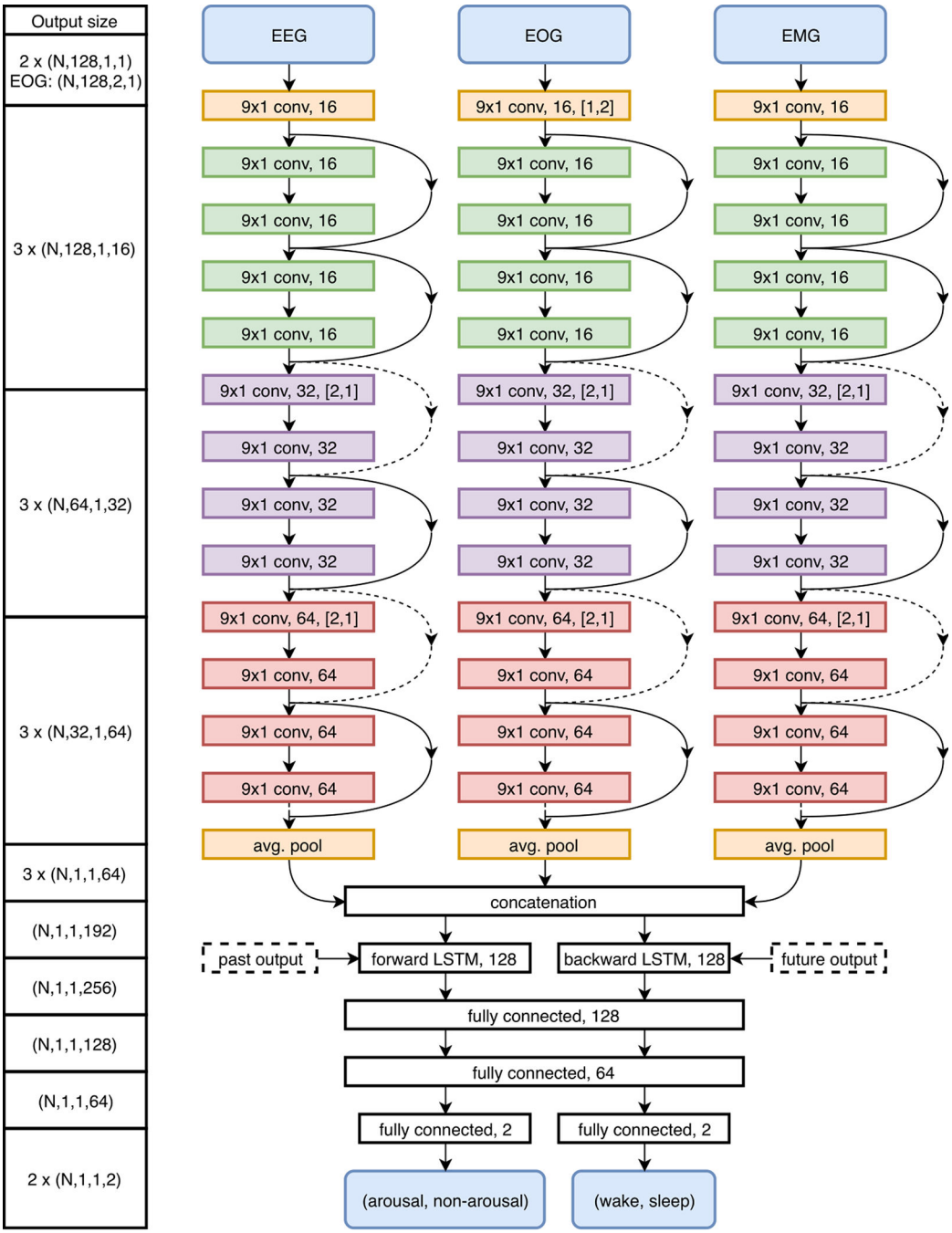

Figure 3:

Deep neural network architecture visualization for MAD. The left column displays the output size throughout the network with dimensions defined as (N: mini-batch size, height, width, channels). Convolutional layers are described by filter kernel size, feature map size, and stride. LSTM layers are described by the number of cells. Fully connected layers are described by the number of hidden units. Shortcut connections are displayed as arrows that are either filled or dashed, dashed arrows indicate the use of zero-padding and max pooling to match dimensions.

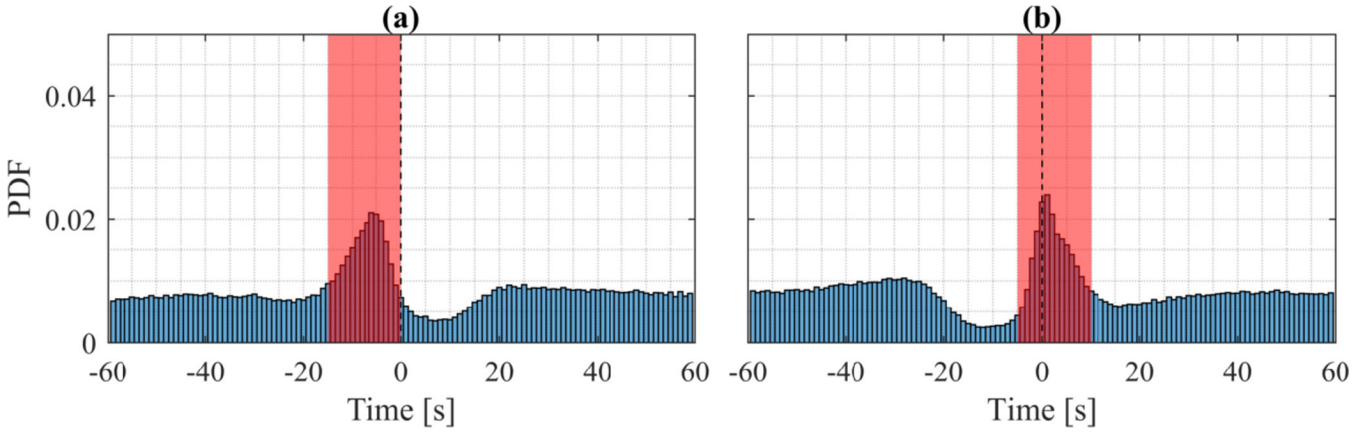

Figure 4:

Distribution of LMs nearby BDs in 1447 PSGs from the WSC. LMs secondary to BDs are discarded based on the windows shown in red. **(a)** Distribution of LMs time-locked to the onset of BDs with no preceding BDs for 60 seconds. **(b)** Distribution of LMs time-locked to the offset of BDs.

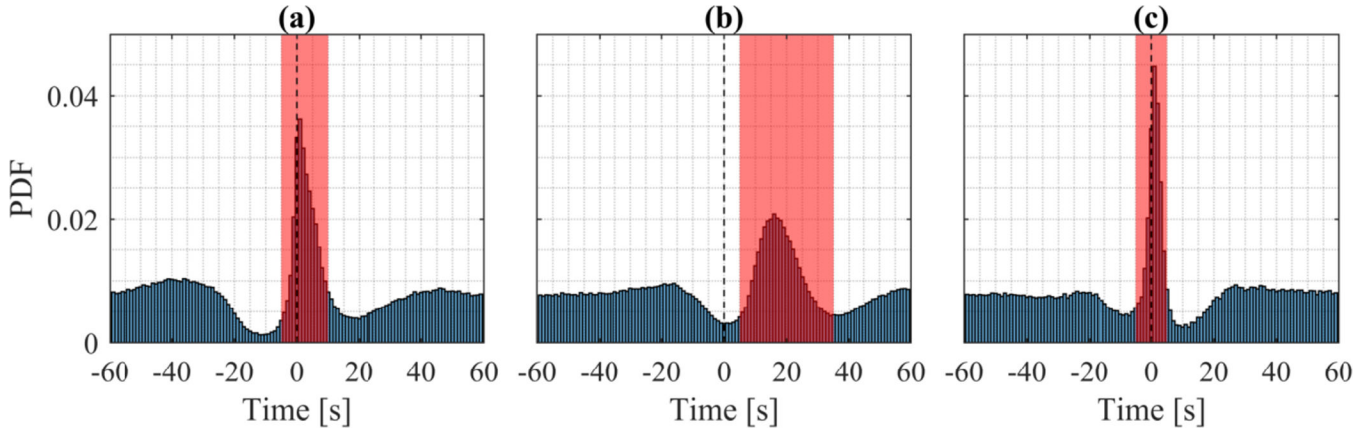

Figure 5:

Distribution of events relative to other event positions in 1447 PSGs from the WSC. Event coupling based on relative distribution of time-locked events. **(a)** Distribution of peak desaturation time-locked to BD offsets. **(b)** Distribution of arousal onset time-locked to BD offsets. **(c)** Distribution of arousal onset time-locked to PLMs onset.

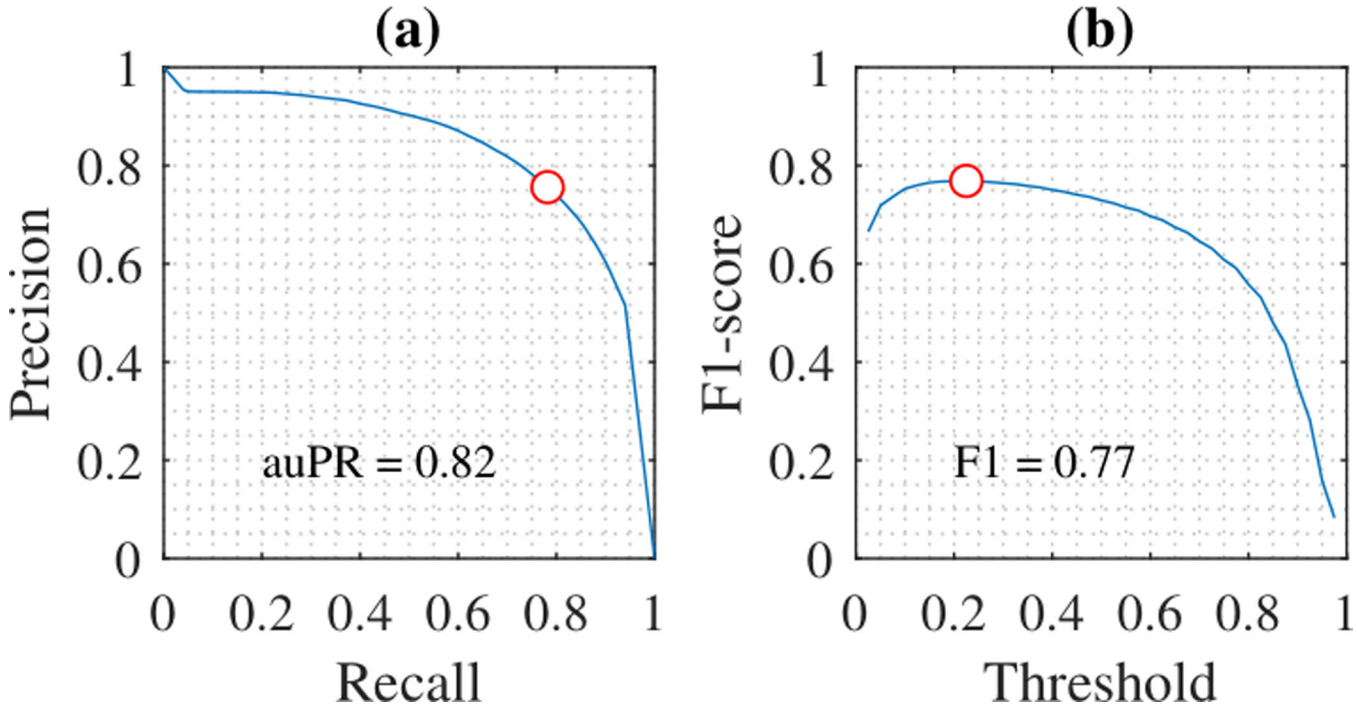

Figure 6:

PR-curves and ROC-curves for predictions. The red circle indicates the optimal threshold at $T_{ar} = 0.255$. **(a)** PR-curve for arousal events. **(b)** F1 score for arousal events.

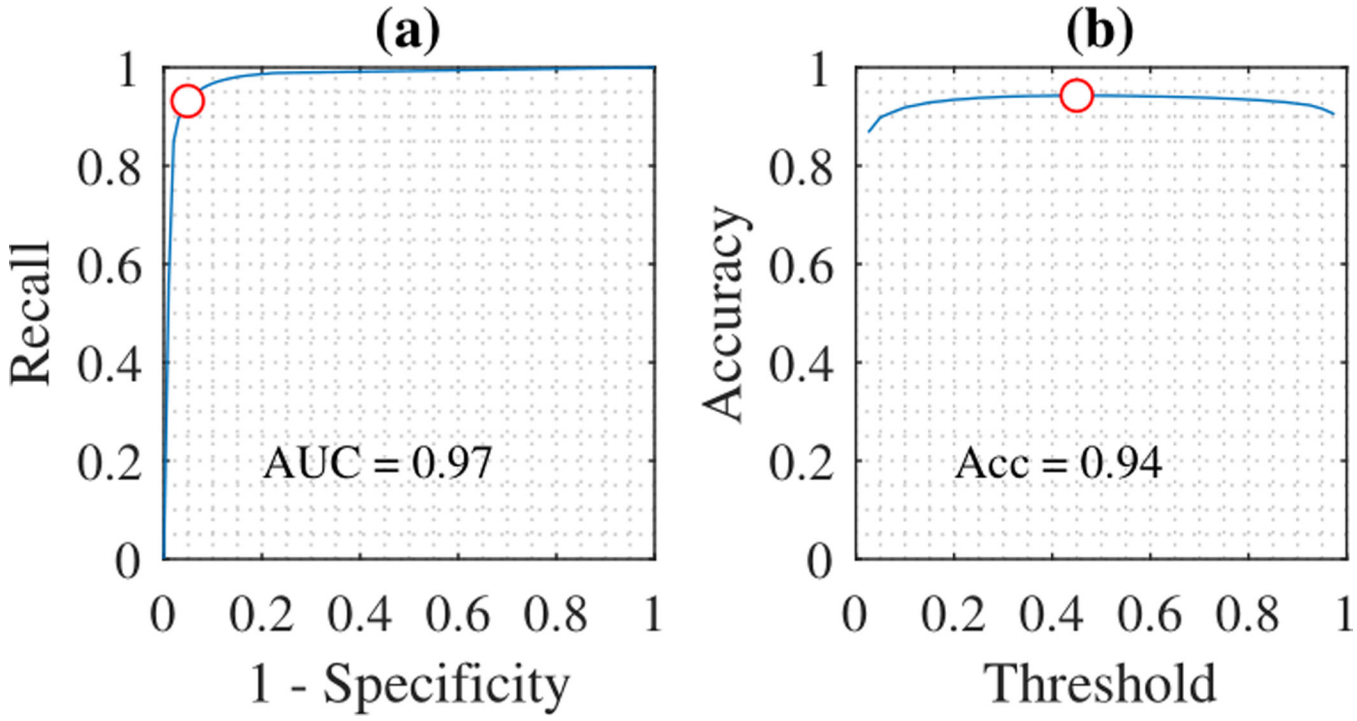

Figure 7:

ROC-curves for predictions. The red circle indicates the optimal threshold at $T_W = 0.45$. **(a)** ROC-curve for wake. **(b)** Accuracy for wake.

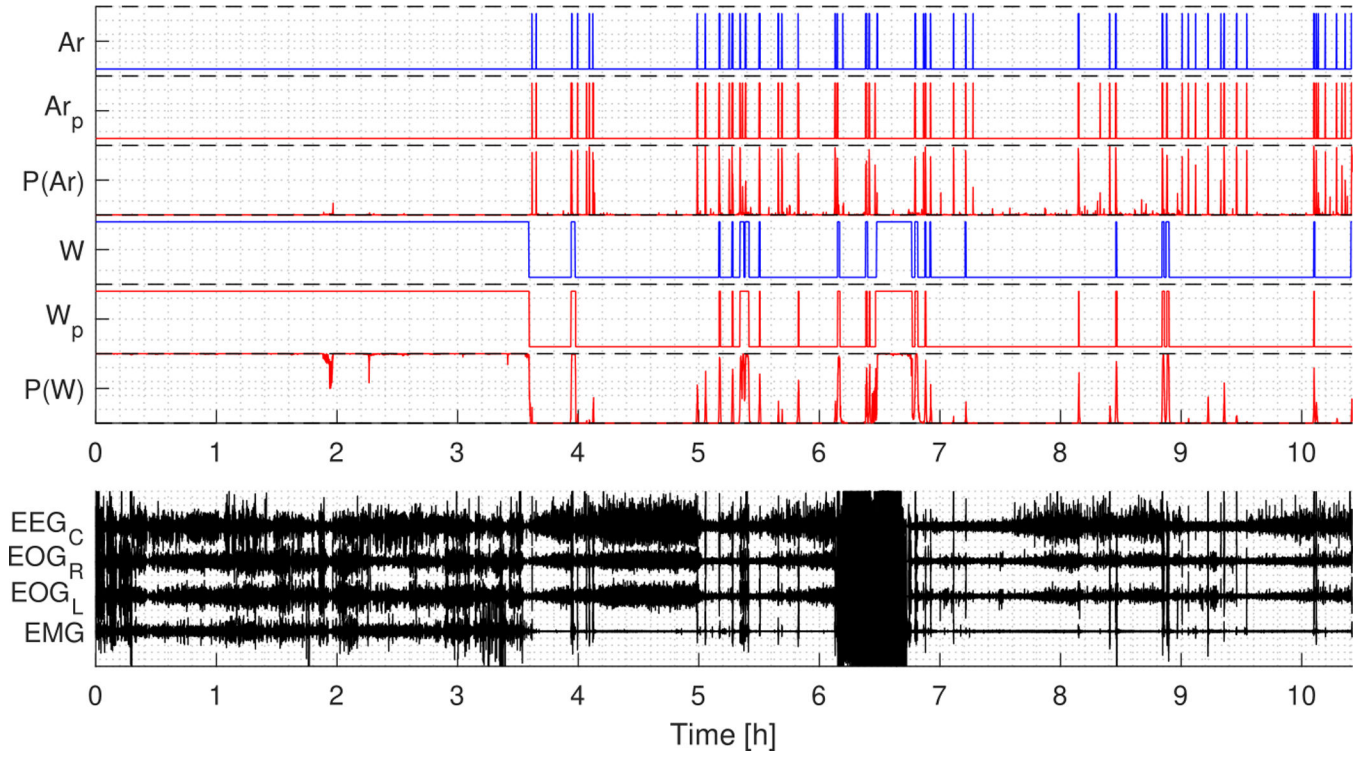

Figure 8:

Example of arousal and wake predictions for a full night's PSG. P(Ar) and P(W) are the probability output, Ar_p and W_p are the predicted labels, while Ar and W are the target labels. The predictions for both arousals and wake are in agreement with the target labels.

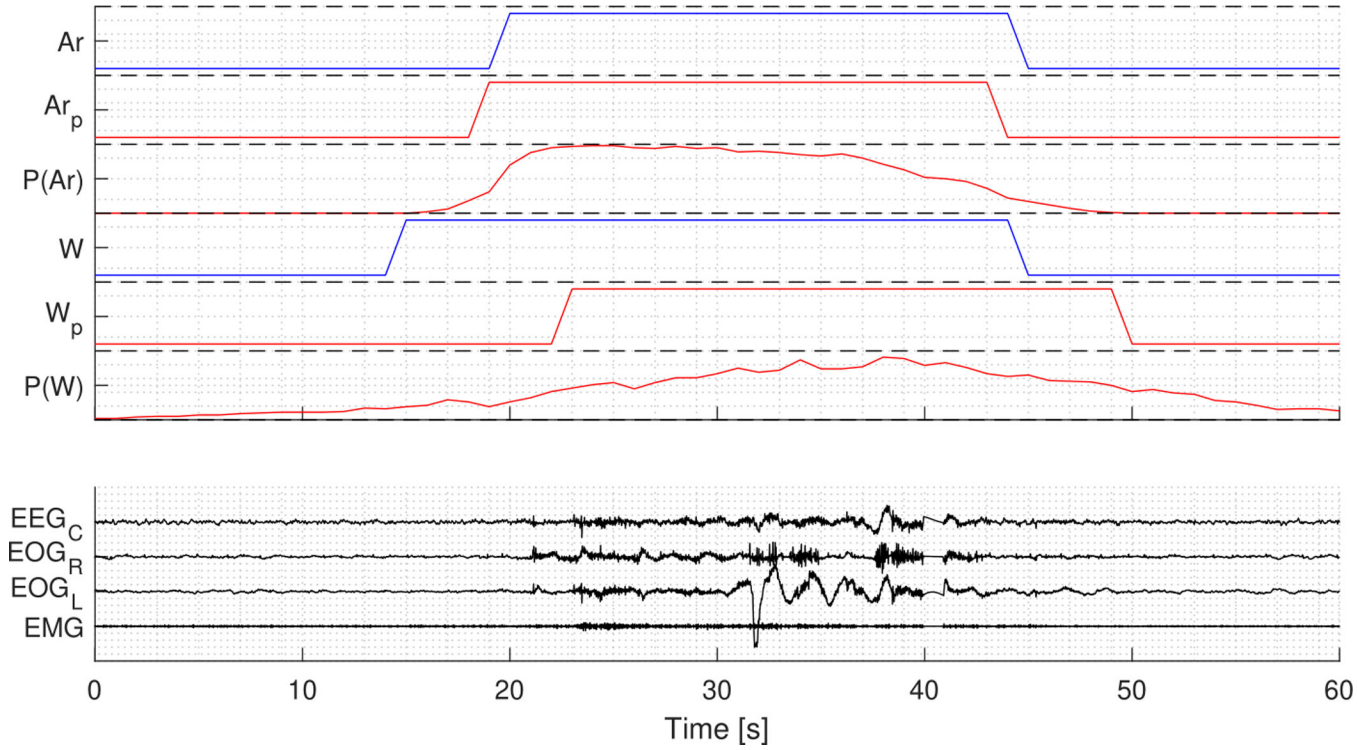

Figure 9:

Example of arousal and wake predictions in a 60 seconds segment from the same PSG displayed in Fig. 8. P(Ar) and P(W) are the probability output, Ar_p and W_p are the predicted labels, while Ar and W are the target labels.

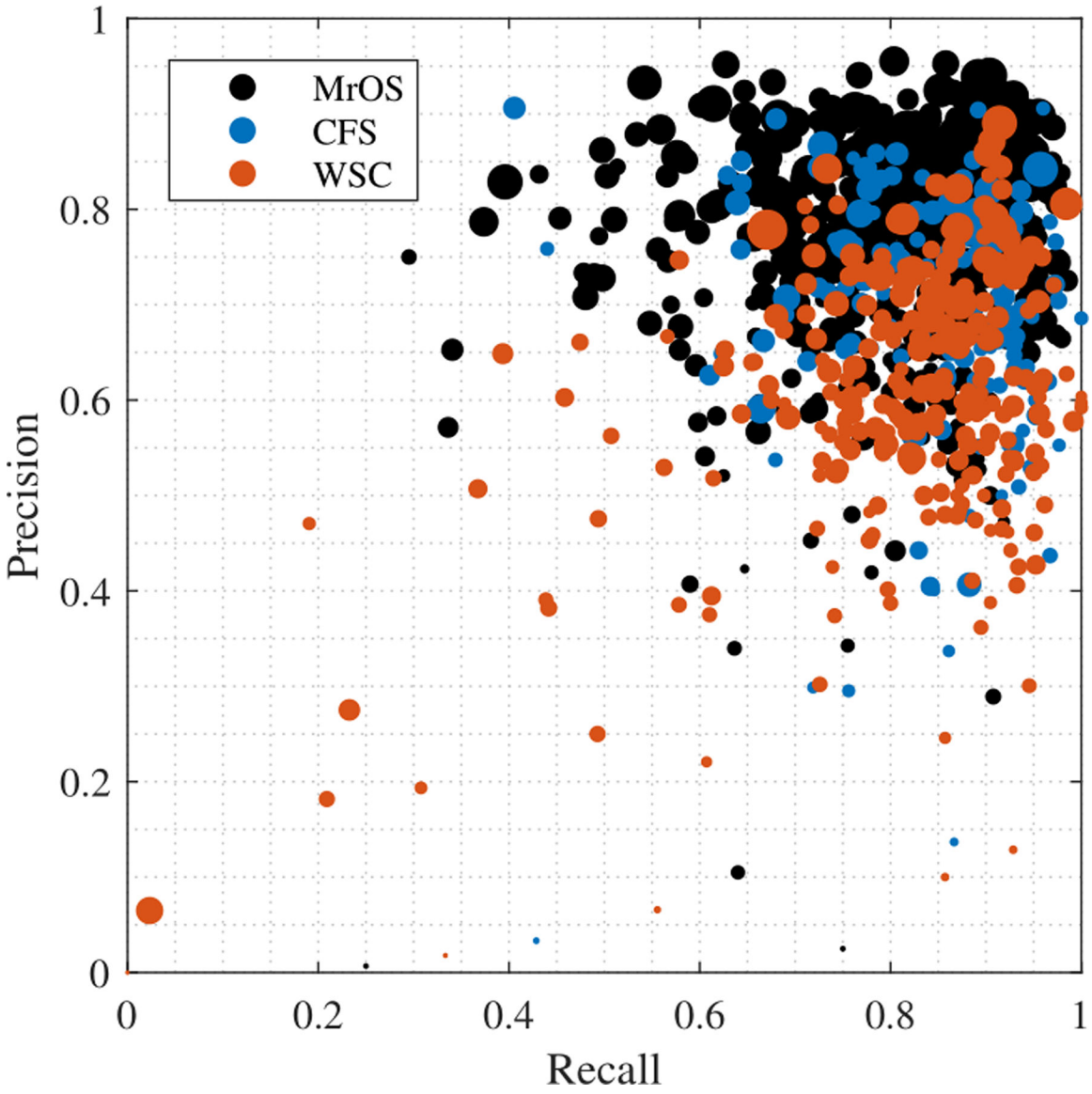

Figure 10:
Precision-recall scatter plot of all test data. Each dot represents a full PSG, and the size is proportional to the number of annotated

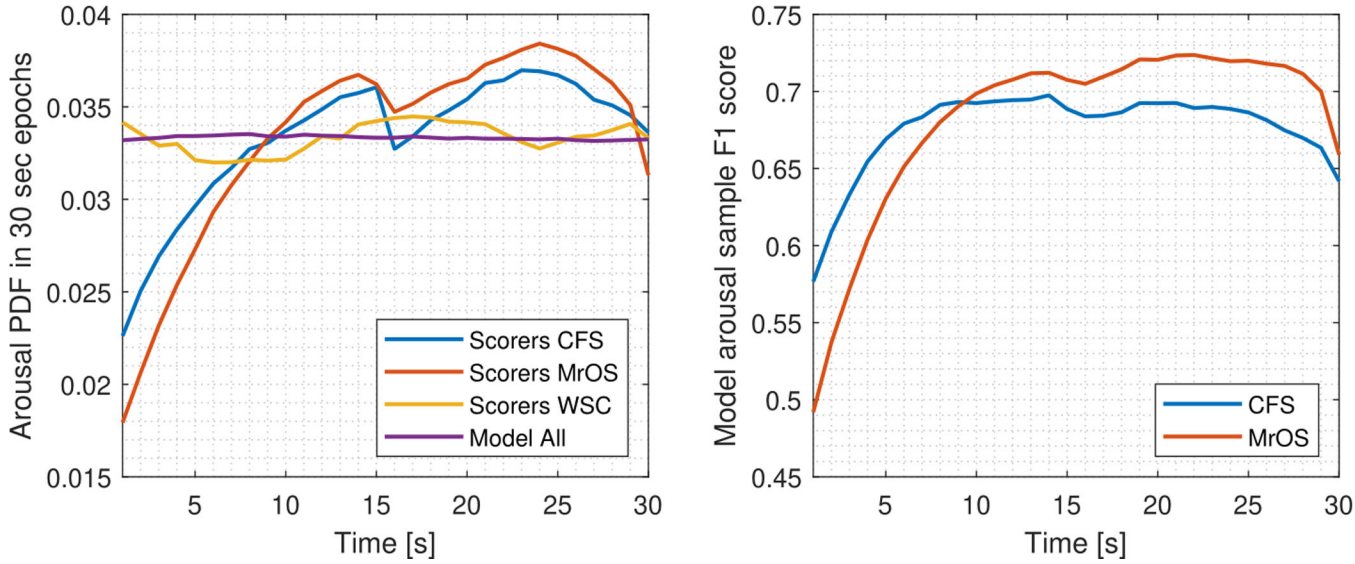

Figure 11:

Left: Arousal PDF over 30-second epochs for manual annotations in CFS (n = 145), MrOS (n = 580), WSC (n = 271) and model predictions over all test data (n = 996). The distribution is expected to be uniform as scoring of arousals is unrelated to this epoch, however annotations in CFS and MrOS exhibit a strong bias toward the central and later part of the 30-second epoch. Right: The arousal F1 score evaluated per second is lower in the first 10 seconds of each 30-second epoch due to this edge bias.

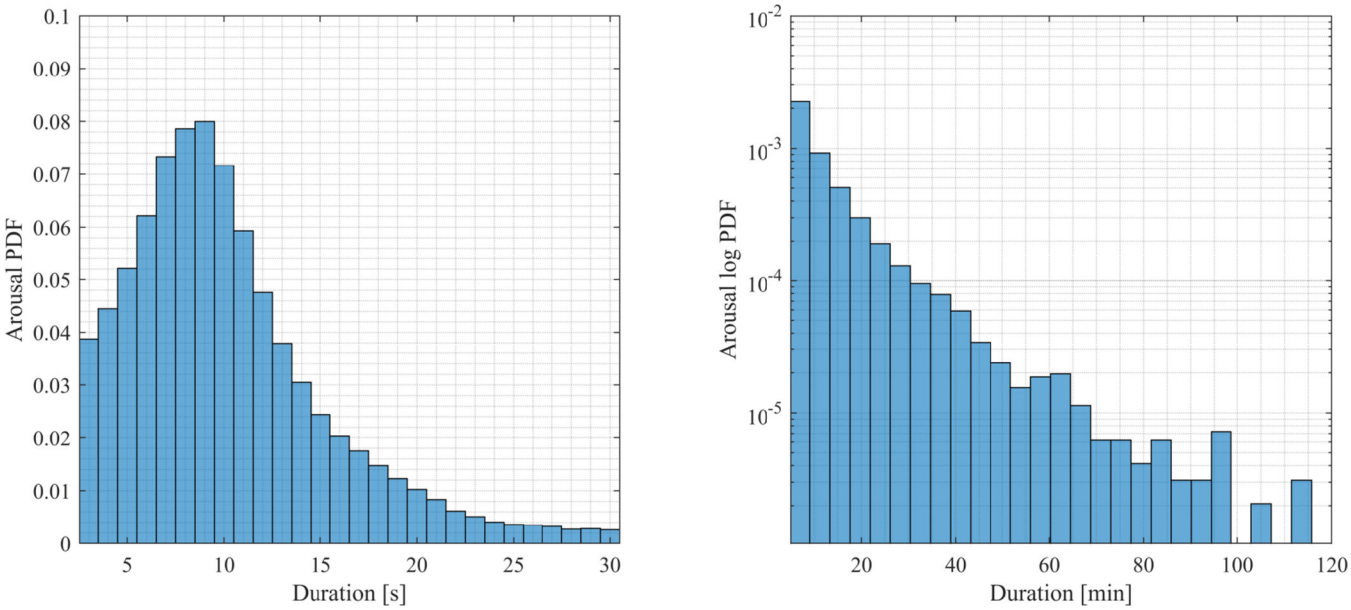

Figure 12:

Distribution of duration of predicted arousals in WSC PSG data (n = 1447). The left plot shows the distribution peaking at 9 seconds, while the right figure shows that the probability decreases exponentially as a function of duration in minutes (shown as a linear decrease in a logarithmic plot).

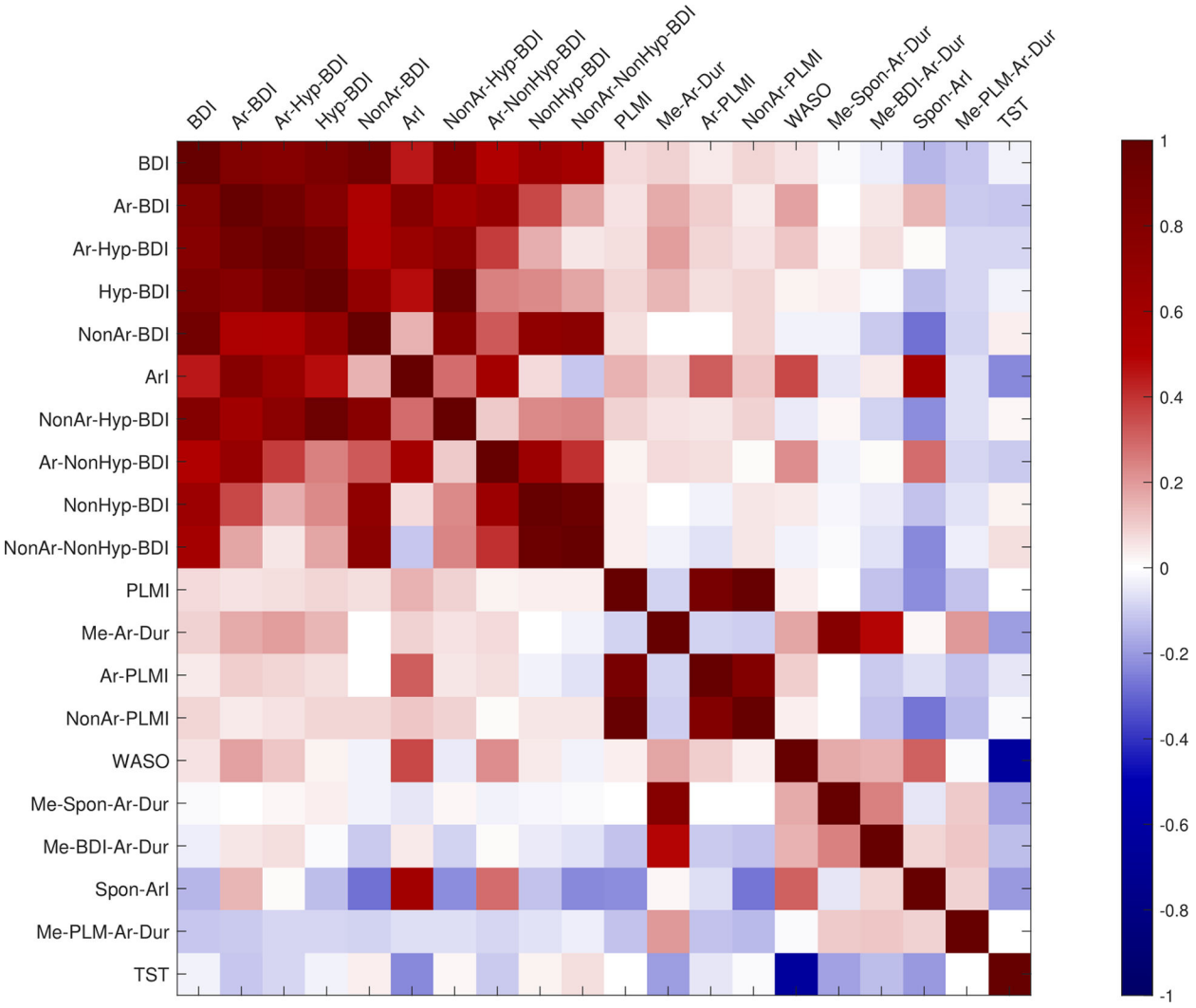

Figure 13:

Correlation of biomarker variables after $\log_2$-transform. Correlation are sorted in a descending order from left to right. The description of these biomarkers can be found in Table 2.

Table 1:

Summary of data demographics and data split. MSL: Mean sleep latency, MrOS: MrOS Sleep Study, CFS: Cleveland Family Study, WSC: Wisconsin Sleep Cohort, and SSC: Stanford Sleep Cohort.

Name	Age (μ ± σ) [min-max]	BMI (μ ± σ)	Sex (% Male)	PSGs (subjects)		
				Arousal Scoring		MSL Statistics
				Training	Testing	
MrOS	76.4 ± 5.5 [67–90]	27.2 ± 3.8	100	2308	580	-
CFS	41.4 ± 19.3 [6–88]	32.4 ± 9.5	44.8	581	145	-
WSC	60.0 ± 8.5 [37–85]	31.7 ± 7.2	53.8	-	271 (269)	1447 (873)
SSC	53.5 ± 15.8 [20–90]	29.1 ± 8.8	66.7	-	30	-
Total	65.4 ± 15 [6–90]	30.4 ± 7.1	74.3	2889	1026 (1024)	1447 (873)

Table 2:

Description of PSG biomarkers. BD: breathing disturbance, PLM: periodic leg movement.

PSG Biomarkers		
Name	**Description**	**Mean ± Std**
BDI	BDs/h	26.1 ± 14.1
Ar-BDI	(BDs w/ arousal)/h	7.6 ± 7.6
NonAr-BDI	(BDs w/o arousal)/h	18.1 ± 9.1
Hyp-BDI	(BDs w/ hypoxia)/h	11.3 ± 12.1
NonHyp-BDI	(BDs w/o hypoxia)/h	14.8 ± 6.8
Ar-Hyp-BDI	(BDs w/ arousal and hypoxia)/h	4.9 ± 6.8
Ar-NonHyp-BDI	(BDs w/ arousal and w/o hypoxia)/h	2.7 ± 2.3
NonAr-Hyp-BDI	(BDs w/o arousal and w/ hypoxia)/h	6.2 ± 6.5
NonAr-NonHyp-BDI	(BDs w/o arousal and hypoxia)/h	11.9 ± 5.4
PLMI	PLMs/h	11.4 ± 18.5
Ar-PLMI	(PLMs w/ arousal)/h	2 ± 3.3
NonAr-PLMI	(PLMs w/o arousal)/h	9.4 ± 16.5
ArI	Arousals/h	21.9 ± 10
Spon-ArI	(Spontaneous arousal)/h	12.4 ± 4.9
Me-Ar-Dur	Median arousal duration (s)	9.5 ± 1.5
Me-BD-Ar-Dur	Median BD arousal duration (s)	10.4 ± 5
Me-PLM-Ar-Dur	Median PLM arousal duration (s)	12.3 ± 33.9
Me-Spon-Ar-Dur	Median spontaneous arousal duration (s)	9.3 ± 2.3
TST	Total Sleep Time (h)	6.7 ± 1
WASO	Wake after sleep onset (h)	1 ± 0.7

Table 3:

Classification performance on test set.

	Arousal Event			Arousal Samples			Wake		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Specificity**	**Recall**	**Accuracy**
MrOS	0.77	0.81	0.79	0.65	0.73	0.69	0.96	0.93	0.95
CFS	0.69	0.83	0.75	0.61	0.76	0.68	0.97	0.92	0.95
WSC	0.62	0.82	0.7	-	-	-	0.97	0.84	0.93
All	0.72	0.81	0.76	0.64	0.73	0.68	0.96	0.92	0.95

Table 4:

Arousal Scoring performance in manually scored sleep stages.

Arousal						
		Wake	**N1**	**N2**	**N3**	**REM**
Events	FPR	6.1e-4	2.5e-3	1.8e-3	6.2e-4	1.6e-3
	Precision	-	0.79	0.77	0.69	0.74
	Recall	-	0.64	0.83	0.86	0.86
	F1	-	0.71	0.8	0.77	0.8
Samples	FPR	0.016	0.026	0.026	0.007	0.022
	Precision	0.64	0.7	0.66	0.56	0.6
	Recall	0.75	0.5	0.75	0.76	0.76
	F1	0.69	0.58	0.7	0.65	0.67

Table 5:

Performance of wake and sleep classification compared to manual scoring of sleep stages.

Wake/Sleep Accuracy					
	Wake	**Sleep**			
	W	**N1**	**N2**	**N3**	**REM**
MrOS	0.935	0.746	0.977	0.996	0.965
CFS	0.916	0.805	0.976	0.996	0.989
WSC	0.842	0.792	0.981	0.997	0.99
All	0.919	0.771	0.978	0.996	0.975

Table 6:

Arousal event F1 score of human scorers and model on pseudo-consensus. *p*-values below 0.05 are highlighted as significant and are calculated using a two-sample t-test with the null hypothesis being equal means while assuming equal but unknown variances.

F1 score		A	B	C	D	E	F	G	H	I	Mean	Mean (A – H)
Human Scorer	**μ**	0.62	0.57	0.65	0.68	0.62	0.65	**0.71**	0.61	0.32	0.60	0.64
	(σ)	(0.17)	(0.16)	(0.14)	(0.16)	(0.19)	(0.14)	(0.11)	(0.17)	(0.2)	(0.19)	(0.16)
Model	**μ**	**0.7**	**0.67**	**0.68**	**0.69**	**0.70**	**0.72**	0.7	**0.66**	**0.71**	**0.69**	**0.69**
	(σ)	(0.14)	(0.14)	(0.11)	(0.1)	(0.13)	(0.08)	(0.08)	(0.12)	(0.12)	(0.12)	(0.11)
***p*-val**		**0.033**	**0.006**	0.318	0.755	**0.038**	**0.016**	0.67	0.19	**1.6e-14**	**6.9e-12**	**1.7e-5**

Table 7:

MSL parameters describing the relationship between the $\log_2$transformed sleep variables and MSL. The table is sorted with respect to *p*-value in a descending order. The parameter estimates are adjusted for age, BMI, sex, habitual sleep duration, minutes of sleep the two nights preceding the MSLT, and predicted WASO.

MSL Model Parameters						
Predictive Variables	**β Coefficient**	**95 % CI**	***p*-value**	**β Coefficient***	**95 % CI***	***p*-value***
WASO	-	-	-	1.45	(0.94, 1.96)	3.1e-8
BDI	−0.65	(−0.97, −0.34)	4.5e-5	−0.66	(−0.97, −0.34)	4.6e-5
Ar-NonHyp-BDI	−0.63	(−0.96, −0.31)	0.00014	−0.44	(−0.77, −0.11)	0.008
NonHyp-BDI	−0.61	(−0.94, −0.28)	0.00028	−0.57	(−0.90, −0.24)	0.0008
Ar-BDI	−0.46	(−0.71, −0.21)	0.00032	−0.38	(−0.64, −0.13)	0.003
NonAr-BDI	−0.51	(−0.84, −0.18)	0.0022	−0.57	(−0.90, −0.24)	0.0007
Ar-Hyp-BDI	−0.34	(−0.56, −0.12)	0.0027	−0.32	(−0.54, −0.1)	0.0046
Me-Spon-Ar-Dur	−1.5	(−2.4, −0.51)	0.0028	−0.95	(−1.9, 0.004)	0.051
NonAr-NonHyp-BDI	−0.49	(−0.81, −0.16)	0.0036	−0.50	(−0.83, −0.16)	0.0034
ArI	−0.67	(−1.2, −0.18)	0.0075	−0.22	(−0.7, 0.26)	0.37
Hyp-BDI	−0.26	(−0.46, −0.05)	0.014	−0.29	(−0.5, −0.08)	0.006
TST	−1.2	(−2.8, 0.32)	0.12	−2.94	(−2.8, 0.32)	5.3e-6
NonAr-Hyp-BDI	−0.18	(−0.41, 0.05)	0.13	−0.26	(−0.49, −0.02)	0.03
Me-PLM-Ar-Dur	0.2	(−0.13, 0.53)	0.24	0.2	(−0.13, 0.53)	0.24
NonAr-PLMI	−0.071	(−0.2, 0.06)	0.29	−0.073	(−0.21, 0.06)	0.29
Me-BDI-Ar-Dur	0.31	(−0.37, 1)	0.37	0.53	(−0.16, 1.22)	0.13
PLMI	−0.058	(−0.19, 0.07)	0.38	−0.058	(−0.19, 0.07)	0.38
Me-Ar-Dur	−0.44	(−1.7, 0.79)	0.49	0.13	(−1.08, 1.34)	0.84
Ar-PLMI	−0.074	(−0.31, 0.16)	0.53	−0.03	(−0.27, 0.21)	0.81
Spon-ArI	0.011	(−0.51, 0.53)	0.97	0.44	(−0.06, 0.95)	0.084

*The effect of the predictive variables is shown when WASO is not included. The effect of the adjusted parameters is shown in Table 8.

Table 8:

Effect of adjusted parameters on MSL. HSD: habitual sleep duration. The table is sorted with respect to *p*-value in a descending order.

MSL Model Adjusted Parameters			
Predictive Variables	**β Coefficient**	**95 % CI**	***p*-value**
Age	0.09	(0.57,0.12)	1e-7
Sex = female	−1.35	(−1.89, −0.82)	7.6e-7
Min night 0	0.0076	(0.004, 0.011)	1.5e-5
HSD	0.49	(0.22, 0.76)	3.4e-4
Min night 1	0.0049	(0.002, 0.008)	0.001
BMI	−0.06	(−0.10, −0.02)	0.0027
Intercept	−1.37	(−4.58, 1.85)	0.40

Table 9:

Variables included (shown in bold font) with stepwise linear regression in a set of fitted models. The models are adjusted for age, BMI, sex, habitual sleep duration, minutes of sleep the two nights preceding the MSLT, and predicted WASO. WASO had a significant effect in all models.

#	Model	*p*-value	R^2 (adj)
1	-	2.85e-25	0.134
2	**BDI**	2.43e-27	0.146
3	**Ar-BDI**, NonAr-BDI	1.21e-26	0.143
4	**Ar-Hyp-BDI**, NonAr-Hyp-BDI	3.62e-26	0.14
5	**Ar-NonHyp-BDI**, NonAr-NonHyp-BDI	1.82e-26	0.142
6	Ar-Hyp-BDI, **Ar-NonHyp-BDI**	1.82e-26	0.142
7	**Ar-Hyp-BDI**, Ar-NonHyp-BDI, NonAr-Hyp-BDI, **NonAr-NonHyp-BDI**	6.58e-27	0.146
8	**NonAr-BDI**, **ArI**	1.75e-26	0.144
9	NonAr-Hyp-BDI, **NonAr-NonHyp-BDI**, **ArI**	7.66e-27	0.146
10	**NonAr-NonHyp-BDI**, **ArI**, **Me-Spon-Ar-Dur**	1.34e-27	0.151
11	**BDI**, **Me-Spon-Ar-Dur**	8.84e-28	0.15

Table 10:

Comparison of methods for automatic arousal detection. In these methods, true positive events are defined as either overlapping (Overlap) or agreement in a set timeframe (1, 1.28 or 30 seconds). CV: Cross-validation.

Method	PSGs		Scoring unit	Performance		
	Train	Test		Recall	Precision	F1
Proposed method, MAD	2889	996	Overlap	**0.81**	0.72	**0.76**
Alvarez-Estevez and Fernández-Varela (2019)	-	2768	30 second	0.58	0.71	0.64
Olesen et al. (2019)	1650	1000	Overlap	0.75	**0.77**	0.75
Coppieterst Wallant et al. (2016)	60	CV	1 second	0.61	-	-
Popovic et al. (2013)	10	29	Overlap	0.72	0.67	0.69
Fernández-Varela et al. (2017b)	20	26	30 seconds	0.78	-	-
Sorensen et al. (2012)	24	CV	Overlap	0.89	0.86	0.87
Fernández-Varela et al. (2017a)	6	22	30 seconds	0.75	0.86	0.79
Shmiel et al. (2009)	6	20	30 second	0.77	0.75	0.76
Shahrbabaki et al. (2015)	9	CV	30 seconds	0.79	-	-
De Carli et al. (1999)	3	8	Overlap	0.88	-	-
Cho et al. (2005)	3	6	1 second	0.75	-	-
Sugi et al. (2009)	8	0	1.28 seconds	0.82	-	-