KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Huysmans D., Borzée P., Buyse B., Testelmans D., Van Huffel S, Varon C. (2021), **Sleep Diagnostics for Home Monitoring of Sleep Apnea Patients** Frontiers in Digital Health, vol. 3, June 2021, 1-13
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	https://www.frontiersin.org/article/10.3389/fdgth.2021.685766
Journal homepage	https://www.frontiersin.org/
Author contact	Dorien.Huysmans@esat.kuleuven.be + 32 (0)16 37 92 69
Abstract	Objectives: Sleep time information is essential for monitoring of obstructive sleep apnea (OSA), as the severity assessment depends on the number of breathing disturbances per hour of sleep. However, clinical procedures for sleep monitoring rely on numerous uncomfortable sensors, which could affect sleeping patterns. Therefore, an automated method to identify sleep intervals from unobtrusive data is required. However, most unobtrusive sensors suffer from data loss and sensitivity to movement artifacts. Thus, current sleep detection methods are inadequate, as these require long intervals of good quality. Moreover, sleep monitoring of OSA patients is often less reliable due to heart rate disturbances, movement and sleep fragmentation. The primary aim was to develop a sleep-wake classifier for sleep time estimation of suspected OSA patients, based on single short-term segments of their cardiac and respiratory signals. The secondary aim was to define metrics to detect OSA patients directly from their predicted sleep-wake pattern and prioritize them for clinical diagnosis. Methods: This study used a dataset of 183 suspected OSA patients, of which 36 test subjects. First, a convolutional neural network was designed for sleep-wake classification based on healthier patients (AHI < 10). It employed single 30 s epochs of electrocardiograms and respiratory inductance plethysmograms. Sleep information

	and Total Sleep Time (TST) was derived for all patients using the short-term segments. Next, OSA patients were detected based on the average confidence of sleep predictions and the percentage of sleep-wake transitions in the predicted sleep architecture. Results: Sleep-wake classification on healthy, mild and moderate patients resulted in moderate κ scores of 0.51, 0.49, and 0.48, respectively. However, TST estimates decreased in accuracy with increasing AHI. Nevertheless, severe patients were detected with a sensitivity of 78% and specificity of 89%, and prioritized for clinical diagnosis. As such, their inaccurate TST estimate becomes irrelevant. Excluding detected OSA patients resulted in an overall estimated TST with a mean bias error of 21.9 (± 55.7) min and Pearson correlation of 0.74 to the reference. Conclusion: The presented framework offered a realistic tool for unobtrusive sleep monitoring of suspected OSA patients. Moreover, it enabled fast prioritization of severe patients for clinical diagnosis.
IR	NA

(article begins on next page)

frontiers

Sleep Diagnostics for Home Monitoring of Sleep Apnea Patients

Dorien Huysmans [1,*], **Pascal Borzée** [2], **Bertien Buyse** [2], **Dries Testelmans** [2], **Sabine Van Huffel** [1] **and Carolina Varon** [1,3]

[1] *STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium*
[2] *Department of Pneumology, UZ Leuven, Leuven, Belgium*
[3] *e-Media Research Lab, Department of Electrical Engineering, KU Leuven, Leuven, Belgium*

Correspondence*:
Corresponding Author
dorien.huysmans@esat.kuleuven.be

ABSTRACT

Objectives. Sleep time information is essential for monitoring of obstructive sleep apnea (OSA), as the severity assessment depends on the number of breathing disturbances per hour of sleep. However, clinical procedures for sleep monitoring rely on numerous uncomfortable sensors, which could affect sleeping patterns. Therefore, an automated method to identify sleep intervals from unobtrusive data is required. However, most unobtrusive sensors suffer from data loss and sensitivity to movement artefacts. Thus, current sleep detection methods are inadequate, as these require long intervals of good quality. Moreover, sleep monitoring of OSA patients is often less reliable due to heart rate disturbances, movement and sleep fragmentation. The primary aim was to develop a sleep-wake classifier for sleep time estimation of suspected OSA patients, based on single short-term segments of their cardiac and respiratory signals. The secondary aim was to define metrics to detect OSA patients directly from their predicted sleep-wake pattern and prioritize them for clinical diagnosis. **Methods.** This study used a dataset of 183 suspected OSA patients, of which 36 test subjects. First, a convolutional neural network for sleep-wake classification was designed based on healthier patients (AHI $<$ 10). It employed single 30s epochs of electrocardiograms and respiratory inductance plethysmograms. Sleep information and Total Sleep Time (TST) was derived for all patients using the short-term segments. Next, OSA patients were detected based on the average confidence of sleep predictions and the percentage of sleep-wake transitions in the predicted sleep architecture. **Results.** Sleep-wake classification on healthy, mild and moderate patients resulted in moderate κ scores of 0.51, 0.49 and 0.48, respectively. However, TST estimates decreased in accuracy with increasing AHI. Nevertheless, severe patients were detected with a sensitivity of 78% and specificity of 89%, and prioritized for clinical diagnosis. As such, their inaccurate TST estimate becomes irrelevant. Excluding detected OSA patients resulted in an overall estimated TST with a mean bias error of 21.9 ($\pm$ 55.7) minutes and Pearson correlation of 0.74 to the reference. **Conclusion.** The presented framework offered a realistic tool for unobtrusive sleep monitoring of suspected OSA patients. Moreover, it enabled fast prioritization of severe patients for clinical diagnosis.

Keywords: sleep, sleep apnea, unobtrusive sensor, wearable sensor, ECG, respiration, convolutional neural network

1 ACRONYMS

- **AHI** Apnea-Hypopnea Index
- **CNN** Convolutional Neural Network
- **OSA** Obstructive Sleep Apnea
- **ECG** Electrocardiography
- **IHR** Instantaneous Heart Rate
- **N1, N2, N3** Non-Rapid Eye Movement sleep 1, 2, 3
- **PSG** Polysomnography
- **REM** Rapid Eye Movement
- **RIP** Respiratory Inductance Plethysmography
- **SD** Standard Deviation
- **TST** Total Sleep Time

2 INTRODUCTION

Obstructive Sleep Apnea (OSA) is the most common sleep related breathing disorder. It is characterized by events of breathing disturbances causing hypoxemia, intrathoracic pressure changes and arousals from sleep. Consequently, OSA is an acknowledged risk factor for excessive daytime sleepiness, hypertension and cardiovascular diseases (Young et al., 2002). As OSA is closely associated with obesity and advancing age, the prevalence is expected to further increase (Senaratna et al., 2017). Nevertheless, many patients remain undiagnosed. One of the reasons is the limited hospital capacity for performing polysomnography (PSG) (Flemons et al., 2004). Furthermore, the clinical diagnostic procedure poses a high level of discomfort for the patient. Therefore, it is desired to identify OSA patients at risk with unobtrusive sensors at home, allowing a comfortable sleeping environment and follow up over multiple nights. Clinically, the severity of sleep apnea is assessed by the Apnea-Hypopnea Index (AHI), which is the number of respiratory events (apneas, hypopneas and respiratory effort-related arousals) per hour of sleep. The events are annotated based on the patient's airflow and oxygen saturation (Berry et al., 2012). A patient is then categorized as not suffering from OSA if $0 \leqslant$AHI< 5, mild OSA if $5 \leqslant$AHI< 15 with presence of symptoms, moderate OSA if $15 \leqslant$AHI< 30 or severe OSA if AHI $\geqslant 30$ (Sateia, 2014). The calculation of this AHI requires the quantification of the hours of sleep, i.e. Total Sleep Time (TST). In fact, there are five sleep stages defined by the American Academy of Sleep Medicine, which are Wakefulness, Rapid Eye Movement sleep (REM sleep) and non-REM (NREM) sleep 1, 2 and 3 (respectively N1, N2 and N3) (Berry et al., 2012). Usually, stages N1 and N2 are referred to as *light* sleep and N3 as *deep* sleep. The rules for annotating sleep stages (i.e. performing sleep staging) are based on patterns and wave characteristics found in the electroencephalogram (EEG), the electrooculogram, and the submental electromyogram. The PSG records these signals, among others such as the respiratory airflow, oxygen saturation and electrocardiogram (ECG). To facilitate the sleep staging, these signals are scored in consecutive windows of 30s, which are referred to as epochs (Rechtschaffen and Kales, 1968). Hence, in this paper, monitoring of sleep apnea patients refers to the whole process of sleep staging, sleep time estimation and severity assessment.

Although clinical sleep staging mainly relies on EEG analysis, many emerging unobtrusive sensor technologies for sleep monitoring are based on cardiac and respiratory signals. Consequently, the development of novel algorithms for automated sleep staging based on these unobtrusive signals is an active topic of research. The following studies developed specific sleep staging algorithms for OSA patients based on cardiac and respiratory information. Often, feature-based approaches were implemented to differentiate between

sleep stages when expert knowledge was available (Willemen et al., 2015; Radha et al., 2019; Dietz-Terjung et al., 2021; Bakker et al., 2021). This implied a disadvantage of the method as prior knowledge was required to find appropriate features. Another disadvantage was the extensive data processing needed to perform accurate feature extraction. To alleviate the manual feature extraction, a deep learning network can be developed, as done by Korkalainen et al. (2020). The network required an input sequence of 100 $\times$ 30s epochs and obtained good performance results for classifying the five sleep stages. These algorithms by previously mentioned authors required long signal segments surrounding a 30s (or 60s) epoch as an input for the epoch's sleep stage classification. As such, these longer segments provided contextual information to improve classification performance. However, long intervals of good quality are in reality not available as unobtrusive sensors are very sensitive to movement artefacts. In addition, OSA patients often show more movements during their sleep compared to healthy subjects. Therefore, the required algorithm input should consist of single and independent signal epochs, to alleviate the requirement of successive good quality segments. However, state-of-the-art sleep staging algorithms rarely take into account the potential data loss and distortion of unobtrusive sensors. Malik et al. (2018) did perform a two-class sleep-wake classification with an input consisting of single 30s epochs, or longer sequences. They solely used the instantaneous heart rate (IHR) (i.e. tachograms) and a one-dimensional convolutional neural network (1D CNN). However, the method was only applied on healthy subjects and the performance on 30s epochs was insufficient. Also in the study of Huysmans et al. (2020), a sleep-wake classifier was developed with 30s epochs, for healthy to mild OSA patients and based on the 1D CNN of Malik et al. (2018). A difference with the classifier of Malik et al. (2018) was that respiratory inductance plethysmography (RIP) signals were added to improve performance. Moreover, the use of tachograms allowed a straight-forward application of other sensors capturing the beat-to-beat variability. As such, the CNN was preliminarily tested with recordings from unobtrusive capacitively coupled ECG. However, the study was based on a limited dataset.

Additionally, in OSA patients, heart rate disturbances and sleep fragmentation complicates algorithm design and validation (Norman et al., 2000; Varon and Van Huffel, 2017). The complexity and validation issue are related to the increase of the uncertainty in clinical sleep staging with the AHI of a patient. It is partially a consequence of the restrictions posed by the scoring rules, as defined in (Berry et al., 2012). For example, patients can pass through two or even three different sleep stages during a 30s interval, although sleep stages are annotated per epoch of 30s. Also micro-sleeps or micro-awakenings of a few seconds will not be annotated. Additionally, apneic events can only be scored if they last at least 10s. State-of-the-art non-EEG sleep staging algorithms are aware of the decrease in prediction performance for a patient population, however the problem is not mitigated (Radha et al., 2019; Fonseca et al., 2020). Therefore, it is desired to detect OSA patients with complex sleep architectures, as they would receive less reliable sleep-wake predictions and can be prioritized for a clinical PSG.

The primary aim of this actual work is to reliably estimate TST for healthy subjects as well as the whole range of OSA patients, based on PSG signals which could be acquired unobtrusively. This means the TST is estimated based on single short-term segments, as unobtrusive data likely includes artefacts and data loss. Therefore, this study proposes a sleep-wake classifier based on Huysmans et al. (2020), which can handle data acquired by unobtrusive sensors. For this, the approach proposed here has a preprocessing phase based on single 30s segments, as opposed to the previous algorithm, which makes it more usable for future application on unobtrusive data. Furthermore, the robustness of the network is verified by training the CNN model multiple times using a variation in training and validation set and by comparing the performance of each model on a test set. This is in contrast with the application of a fixed training and validation set. In addition, the network is tested on the whole range of OSA patients, instead of only healthy and mild OSA patients. The secondary aim is to assess the applicability of the classifier's outcome for detection

of OSA patients, who would receive less reliable sleep-wake predictions. The TST estimates of these patients would be less accurate, but they can be directly prioritized for a clinical diagnostic test. Thus, the relationships between a patient's classification outcome and its OSA severity is analyzed. As the sleep-wake network is trained on healthy subjects and mild OSA patients, a relatively small amount of apneic events is included in the training set. Thus, a first hypothesis is that the CNN classifier will exhibit uncertain sleep-wake predictions in the presence of apneic events. The second hypothesis is that more transitions from sleep to wake and vice versa occur in the predicted sleep pattern of OSA patients, also caused by apneic events. Hence, this study addresses the need for a sleep monitoring framework that accommodates signals acquired by unobtrusive sensors, as it takes into account data losses through the analysis of single short-term segments. Furthermore, the framework investigates how the predicted sleep architecture of OSA patients and the decrease in reliability can be applied to detect these patients, and increase overall sleep monitoring performance.

3 MATERIALS AND METHODS

This study is organized as illustrated in figure 1. First, the different datasets and their demographics and sleep information are described in section 3.1. Section 3.2 presents the preprocessing methodology of ECG and RIP data. The classifier's architecture, its training procedure and the derivation of the TST are described in section 3.3. Furthermore, section 3.4 studies the link between a patient's sleep-wake prediction and its OSA severity in order to detect OSA patients.

3.1 Datasets

The dataset comprised 183 patients who were referred to the sleep laboratory of the University Hospitals Leuven (UZ Leuven, Belgium) for a diagnostic PSG. The B3IP device from Medatec (Haillot, Belgium) served as polysomnograph and provided data from the built-in ECG (SPES electrodes) and built-in thoracic RIP (SleepSense belts) (Medatec, 2021). Medatec Brainnet Winacq 5.0 was the acquisition software and Medatec Brainnet Winrel 5.0 the analyzing software. A clinical sleep expert annotated the sleep stages and apneic events according to the AASM 2012 scoring rules (Berry et al., 2012). The collection of data was approved by the ethical committee of UZ Leuven (S60319) and all patients signed an informed consent. From the full dataset, 36 patients were left out as an independent, unseen dataset for validation of sleep-wake classification, TST estimation and detection of OSA patients. These patients were part of an additional data collection later in time, complying with the same ethical standards. The remaining patients were split into subsets for different purposes, as described in section 3.3.2. The overview of the different subdatasets can be found in table 1. Figure 1 indicates which datasets were applied for parameter optimization or model selection.

3.2 Data Preprocessing

The sleep-wake classification network was developed based on full-night recordings of ECG and RIP, extracted from the PSG. The preprocessing steps took into account the application on unobtrusive, movement-sensitive sensor recordings, with frequent episodes of insufficient quality. As such, the full signal was first segmented into non-overlapping windows of 30s and preprocessing was performed on these individual segments.

ECG: First, R-peak detection was performed on 30s segments, with the method proposed by Moeyersons et al. (2019). Segments with less than 15 detected R-peaks were discarded. From the remaining segments, the IHR was derived and expressed in beats per minute. The unevenly sampled IHR data points were

interpolated at 4 Hz by a piecewise cubic hermite interpolating polynomial, resulting in segments of 120 samples. To avoid border problems during interpolation, the first and last beat of the segment were shifted in time. The first beat time was calculated by subtracting the mean value of the second and third interbeat interval from the second beat time. Similarly, the last beat time was calculated by adding the mean value of the second and third last interbeat interval to the second last beat time. Next, outliers were identified whenever the IHR value was outside the range of 40 to 180 beats per minute, or outside the segment's median value $\pm$ 20 beats per minute, or outside the segment's median value $\pm$ (3 $\times$ the segment's standard deviation (SD)). The first condition were physiological boundaries. The second and third were defined empirically using visual inspection and logical values. Next, the outliers were indicated with NaN. The NaN interval was corrected as long as the duration of subsequent NaNs was smaller or equal to 10 samples (i.e. 2.5s). This NaN gap was filled by mirroring the values preceding the gap (Pichot et al., 2016). Outlier correction was important to not discard epochs with minor artefacts and preserve a maximal number of epochs. Finally, the interpolated values of remaining segments were concatenated and the overall median for each subject was subtracted from every segment. In this way, inter-subject variability was removed but the inter-sleep stage variability retained. As a neural network cannot process NaN values, every segment with remaining NaN values was discarded.

RIP: The segments of the RIP signal were bandpass filtered at [0.04, 2] Hz and downsampled to 4 Hz by spline interpolation, resulting in segments of 120 samples. Then, the median and SD value of every segment was considered. As such, every patient recording had a distribution of median values and one of SD values. Next, every segment was normalized by subtraction with the 50th percentile of the median values and dividing by the 50th percentile of the SD values, to reduce the influence of respiratory artefacts. This was followed by the subtraction of the individual median per segment. Segments discarded after ECG preprocessing as they contained remaining NaN values, were also discarded from the RIP data. Remaining epochs, i.e. without NaNs, were fed to the neural network.

3.3 Sleep-Wake Classification

3.3.1 Neural Network Architecture

The neural network consisted of a convolutional part for feature representation and a dense part for classification (see figure 2). Two separate unimodal networks were first optimized using the cardiac or respiratory signal, based on Malik et al. (2018). After training, the convolutional layers of these networks were combined into a multimodal network, retaining the weights of these layers. Only the dense layers of the multimodal network were optimized using training. All networks consisted of four types of layers. The convolutional layers were defined as $(\mathrm{f}, \mathrm{k}, \mathrm{s}) - \mathrm{Conv}$, with a depth f, a kernel size k, a stride s and an activation of type ReLu. After the convolutional block, dense layers, $(\mathrm{n}) - \mathrm{Dense}$, with n neurons were included. A third type were dropout layers, $(\mathrm{d}\%) - \mathrm{Dropout}$, where $\mathrm{d}\% = 50\%$ of the nodes were set to zero in every training step to avoid overfitting (Srivastava et al., 2014). The output layer is a softmax layer, $\mathrm{Softmax}(1, \mathrm{c})$, delivering posterior class probabilities for every one of the $\mathrm{c} = 2$ classes, where class 0 represented Sleep and class 1 Wake. As an optimization scheme, Adam was chosen, which uses an adaptive learning rate for weight updates instead of a fixed rate (Kingma and Ba, 2014). The network trained with balanced and shuffled batches of sixteen non-sequential epochs. Balancing was achieved by over-sampling classes, such that every batch contained on average an equal number of samples of every class. The threshold of posterior class probability for classification was set at 0.5, thus assigning a segment to class *Wake* if $\mathrm{p}_{\mathrm{class}} \geqslant 0.5$.

3.3.2 Neural Network Training and Selection

Training of the network was performed on 56 patients from UZ Leuven with a low AHI (i.e. AHI $<$ 10), so that the network purely learned patterns of sleep or wake and not to recognize apneic events for classification. Moreover, patients with higher OSA severity have stronger physiological dynamics, which may block the learning process of typical sleep patterns. The training dataset was randomly split into a subset using 70% (N = 39) of the patients for weight training of the neural network ($\mathrm{CNN_Train}$) and 30% (N = 17) for validation during training ($\mathrm{CNN_Val}$), with N the number of subjects. The subdivision changed ten times, using a different seed for randomization, to train and validate ten models. The same ten seeds were used for both the unimodal ECG and RIP networks as well as for the multimodal network. The final multimodal model was selected based on the highest Cohen's Kappa score (κ) obtained using the fixed (i.e. non-randomized) set $\mathrm{CNN_Test}$. The κ score is a measure of inter-rater agreement, while compensating for the degree of agreement expected by chance. It ranges from –1 (total disagreement) through 0 (random classification) to 1 (total agreement). The interpretation of κ, however, varies among different studies (McHugh, 2012).
In addition, the patients of dataset $\mathrm{CNN_Test}$ were merged with patients with higher AHI and split again according to clinical OSA categories in the subsets No, Mild, Mod and Sev. As such, the selected sleep-wake classifier tested these populations with varying AHI. Finally, a Wilcoxon signed rank test verified the performance differences between the unimodal networks, and between the unimodal versus multimodal network on the patients in No, Mild, Mod and Sev.

3.3.3 Assessment of Total Sleep Time (TST)

The TST was estimated as the total time spent asleep in minutes, for datasets No, Mild, Mod, Sev and Test. The comparison was performed by subtracting the reference TST from the estimated TST and calculating the mean and SD of this difference. In addition, the Pearson's correlation coefficient ρ between the reference TST and estimated TST was calculated.

3.4 Detection of OSA Patients based on Sleep-Wake Classifier Outcome

The secondary aim of this study was to assess the applicability of the classifier's outcome for detection of OSA patients. Therefore, the relationships between a patient's outcome of the sleep-wake classifier and its OSA severity was analyzed in section 3.4.1. These relations were used as metrics for which appropriate thresholds were required to detect OSA patients. Threshold selection was performed in section 3.4.2.

3.4.1 Relations between Sleep-Wake Classifier Outcome and OSA Severity

The sleep-wake classifier network was trained on a rather healthy population ($\mathrm{CNN_Train}$ with AHI $<$ 10), in which a relatively small amount of apneic events was present. It was hypothesized that the network output would exhibit uncertain sleep-wake predictions in the presence of apneic events, as mentioned in the introduction. Therefore, the probabilistic outcome of $\mathrm{CNN_Test}$ was further inspected to increase insight into the predictions, as explained further on and illustrated in figure 3. The top row represented the outcome of the CNN, which was the wake probability of each epoch, i.e. $\mathrm{p(Wake)}$. The second row shows the predicted sleep-wake classification with the threshold for posterior class probability at 50% (see section 3.3.1). The last row showed the ground truth sleep stages, which clinicians annotated. However, as can be seen from the top row, some epochs had a $\mathrm{p(Wake)}$ just above 50%. Thus, the prediction of these epochs was rather uncertain. On the other hand, an epoch with a very low $\mathrm{p(Wake)}$, e.g. 10%, indicated an epoch which was predicted as *Sleep* with a high confidence. Based on these observations, a distinction was made between *confident* and *uncertain* predicted epochs by defining confidence thresholds (table 3). The wake

confidence threshold $\mathrm{T_w}$ served as the threshold for epochs predicted as *Wake*. It was the median $\mathrm{p(Wake)}$ of epochs predicted as *Wake* minus its SD, calculated over all subjects of $\mathrm{CNN_Test}$. For epochs predicted as *Sleep*, the $\mathrm{p(Sleep)} = 1 - \mathrm{p(Wake)}$ was considered. Thus, the sleep confidence threshold $\mathrm{T_s}$ was the median $\mathrm{p(Sleep)}$ of epochs predicted as *Sleep* minus its SD, calculated over all subjects of $\mathrm{CNN_Test}$. Epochs with a $\mathrm{p(Wake)}$ between these margins had an uncertain prediction. These margins were applied on sets No, Mild, Mod, Sev and Test. Thus, the amount of uncertain sleep or wake predictions over the total number of predicted epochs was investigated as an indicator of apneic severity, referred to as $\%\mathrm{Uncertain\ Sleep\ Epochs}$ and $\%\mathrm{Uncertain\ Wake\ Epochs}$.

In addition, the predicted sleep architecture was expected to exhibit more frequent sleep-wake transitions with increasing AHI. Reasons for this included the expected increase of sleep fragmentation with the amount of apneic events (Kimoff, 1996), the presence of micro-awakenings due to apneas and the sympathetic activation related to apneas that resemble cardiorespiratory behaviour during wakefulness (Guilleminault et al., 1984; Varon and Van Huffel, 2017). Due to the latter, the network might predict a wake epoch shortly after the occurrence of an apneic event although the patient continued sleeping. Therefore, the percentage of wake-sleep *plus* sleep-wake transitions in the prediction was examined as a second identification metric for high risk OSA patients, referred to as $\%\mathrm{Sleep\ Wake\ Transitions}$. More precisely, every change in the prediction from wake to sleep or vice versa was counted and divided over the total number of predicted epochs. Only remaining (i.e. without NaNs) epochs were counted.

3.4.2 Detection of OSA Patients

The goal was to apply the sleep-wake classifier outcome, namely the metrics $\%\mathrm{Uncertain\ Sleep\ Epochs}$ and $\%\mathrm{Sleep\ Wake\ Transitions}$, for detection of OSA patients. Firstly, to gain insight into the suitability of these metrics for patient detection, the distributions of both metrics were visualised with boxplots per OSA severity class. This was performed using the four datasets No, Mild, Mod and Sev. An upward trend of each metric with OSA severity was expected. Thus, a Kruskal–Wallis test with Bonferroni correction tested significant differences ($p<0.05$) between OSA classes. As a patient is regarded as suffering from OSA if the AHI $\geqslant 15$, regardless of having symptoms, the presented method should be able to select moderate ($15 \leqslant$ AHI < 30) and severe patients (AHI $\geqslant 30$). For simplicity, it was chosen that if at least one of both metrics exceeded a selected threshold, the patient was identified as being at high risk of OSA, i.e. detected positive. Therefore, ROC analysis was carried out to select a suitable OSA detection threshold for each metric. A large specificity was preferred when setting the thresholds, as this meant the identified OSA group would contain few false positives, i.e. few non-OSA patients falsely detected to have OSA. Hence, this implied the detection of patients with rather high AHI values, as opposed to AHI values close to 15 events/h. Hence, when detecting OSA patients at home using only unobtrusive cardiac and respiratory sensors, moderate and severe OSA patients could be detected with a high confidence and given prioritization for a diagnostic PSG. This procedure for detecting OSA patients was assessed on the Test data set .

4 RESULTS

4.1 Sleep-Wake Classifier Selection and Performance

The multimodal network was trained ten times on different distributions of $\mathrm{CNN_Train}$ and $\mathrm{CNN_Val}$. Application of these ten networks onto $\mathrm{CNN_Test}$ resulted in moderate κ scores ranging between 0.46 and 0.51. The multimodal model with the highest κ was chosen[1]. The weights of the convolutional layers of this chosen multimodal network were the same as the final ECG and RIP unimodal networks. Application

[1] The network will be made publicly available after publication.

of CNN_Test on the selected ECG, RIP and multimodal networks resulted in κ = 0.31, 0.46 and 0.51, respectively. In addition, the multimodal CNN tested all other datasets. Table 2 shows the resulting κ scores. Using all patients with varying AHI, the Wilcoxon signed rank test indicated significant different κ scores ($p<0.05$) for the RIP and ECG+RIP networks compared to the ECG network, and the RIP compared to the ECG+RIP network. Next, the TST estimates were compared with the reference value for all datasets (table 2).

4.2 Uncertainty in Sleep-Wake Classifier Outcome

The probability of an epoch predicted as sleep, p(Sleep), or wake, p(Wake), are shown in table 3. The median and SD values of p(Sleep) and p(Wake) of dataset CNN_Test defined the confidence thresholds for the multimodal network (see section 3.4.1). This resulted in $T_s = 0.87 - 0.06 = 0.81$ and $T_w = 0.69 - 0.06 = 0.63$. Taking these thresholds into account, the percentages of uncertain predicted sleep and wake epochs were derived and displayed in table 3. ECG based predictions appear more difficult as the %Uncertain Sleep Epochs was highest compared to RIP and ECG+RIP. Instead, for RIP and ECG+RIP outcomes, this number increased with AHI.

To further investigate the origin of uncertain epochs, a distinction was made between uncertain epochs with and without apneas. For No, Mild, Mod, and Sev, the %Uncertain Sleep Epochs *with* the presence of an apneic event were respectively 1%, 4%, 13% and 39%. Thus, it was found that apneic events caused the increase in %Uncertain Sleep Epochs with AHI. The %Uncertain Sleep Epochs *without* the presence of an apneic event were respectively 32%, 34%, 31% and 21%. These values stayed rather stable over the datasets with increasing AHI, however, a clear decrease was seen for Sev. To investigate the cause of uncertainty for non-apneic epochs, the ground truth sleep stages of these uncertain epochs were extracted for CNN_Test. The largest portion of uncertain sleep predicted, non-apneic epochs were present during N2 and REM sleep. On the other hand, N2 was also the most frequent sleep stage, as seen in table 1. Therefore, the portion of uncertain non-apneic epochs per sleep stage was investigated. For this, the classes N1 and REM had the largest ratio, being 55.1% and 53.2%, respectively. However, uncertain predictions did not necessarily imply incorrect predictions. Nevertheless, classes N1 and REM also had the largest ratio of uncertain non-apneic epochs which were wrongly predicted, respectively 9.1% and 4.3%. These results can be found in more detail in the Supplementary Material.

4.3 Detection of OSA Patients

The values of No, Mild, Mod, and Sev for %Uncertain Sleep Epochs and %Uncertain Wake Epochs increased with OSA severity class, as shown in table 3. However, the trend was more pronounced for %Uncertain Sleep Epochs and was therefore chosen as the preferred metric. The distributions for No, Mild, Mod, and Sev with corresponding ROC curve for detection of AHI $\geqslant$ 15 are displayed in figure 4. The significance tests confirmed the increasing trend of %Uncertain Sleep Epochs with OSA severity. The area under the ROC curve was 0.77. Furthermore, an operating point on the ROC curve was chosen where the specificity was $\geqslant$ 95%, since a larger specificity for detection of OSA patients was preferred. As such, a threshold of 64% was selected, at which specificity reached 97% and sensitivity 37%. A similar study was carried out for %Sleep Wake Transitions, for which the area under the ROC curve was 0.75. Also the upward trend with OSA severity was confirmed by a Kruskal–Wallis test (figure 4). A threshold of 24% was selected, at which a specificity of 95% and sensitivity of 33% was obtained. The detection capabilities of these metrics and corresponding thresholds on Test are shown in figure 5. Detection of OSA patients in set Test resulted in a κ of 32%, accuracy of 64%, sensitivity of 56% and specificity of 89%.

The specificity was relatively high, as expected, as there was only one false positive out of 36 patients.

5 DISCUSSION

5.1 Sleep-Wake Classification

For sleep diagnostics of OSA patients in a home setting, sleep staging algorithms based on cardiac and respiratory signals are required, as these signals can be acquired by unobtrusive sensor technologies. However, many state-of-the-art sleep staging algorithms require long temporal dependencies in the data, which cannot be garantueed in data acquired by unobtrusive sensors. Therefore, this study explicitly focussed on single short-term signal inputs for sleep staging. More specifically, this study proposed a deep learning network for sleep-wake classification based on single 30s epochs from cardiac and respiratory signals in suspected OSA patients. Furthermore, the network was validated on an unseen test set.

The Wilcoxon signed rank tests showed that the RIP based network was more informative compared to the ECG based equivalent, as higher κ values were reached (see section 4.1). Nevertheless, application of the cardiac tachogram did have a benefit as combining the ECG and RIP signals into the multimodal network outperformed the RIP unimodal network. An additional advantage of including the cardiac tachogram is the usage of beat-to-beat variability, allowing the use of other cardiac sensors. Examples are pulse photoplethysmography and ballistocardiography, which enable heart beat extraction.

Furthermore, a distinction was made between epochs that reached prediction confidence tresholds and those which were uncertain. As reported in section 4.2, the $\%\mathrm{Uncertain\ Sleep\ Epochs}$ *without* the presence of an apneic event was on average 30% of sleep predicted epochs. For these type of epochs, the prediction of N1 and REM epochs showed the lowest confidence. Both N1 and REM are more *active* stages of sleep, where the heart rate is elevated and the respiration more irregular (Douglas et al., 1982; Bassetti et al., 2014). This ressembles the cardiorespiratory behaviour during wake and partially explains the larger confusion in prediction of these epochs. Furthermore, the ratio of N1 and REM epochs in the training data was low, as seen in table 1. Hence, the network had less diverse examples to learn from, adding to the lower testing performances for N1 and REM epochs.

Comparison of the sleep-wake classification to literature was difficult as studies generally do not focus on using single short term epochs. Most studies include contextual information, by applying epoch sequences, which improves performance, at the cost of requiring long segments of good quality. This is extremely difficult to guarantee when using real and unobtrusive techology. Only the study of Malik et al. (2018) fed single 30s epochs from ECG to a CNN, but achieved a low κ of 0.25 for sleep-wake classification on healthy subjects. In contrast, the current study achieved a superior κ of 0.49 and 0.48 for mild and moderate OSA patients, respectively, which is in addition more challenging than classification in healthy subjects. On the other hand, Korkalainen et al. (2020) obtained a κ of 0.65 for wake-NREM-REM classification with pulse photoplethysmography in OSA patients with a median AHI of 16.8. Their performance was superior, but the used CNN was fed with a sequence of 100 epochs of 30s. Similarly, Dietz-Terjung et al. (2021) reached a κ of 0.62 for wake-NREM-REM using actigraphy and RIP in patients with an average AHI of 19.0. Their algorithm required a manual feature extraction on 25 epochs of 30s. Although the current network reached lower κ scores compared to the latter studies, it offers a realistic approach for sleep-wake classification with unobtrusive sensors, as it is based on single 30s epochs.

5.2 Total Sleep Time Estimation

The comparison of TST estimates with the reference value in table 2 shows an increase in SD with an increase in AHI. It demonstrates a decrease in reliability of the outcome. Next, the estimation of TST

on dataset Test was performed twice, once including all subjects (pre-detection) and once on subjects detection as non-OSA (post-detection). The reason for this was twofold. First, estimation of the TST becomes irrelevant when severe OSA patients can be detected, as they are directly prioritized for a clinical diagnostic test. Thus, an AHI estimation at home becomes redundant, as well as the corresponding TST estimation. Second, TST estimates becomes more reliable for milder OSA patients, due to more stable physiological dynamics, as further discussed in 5.3. For dataset Test, the ρ increased for post-detection ($\rho = 0.74$) compared to pre-detection ($\rho = 0.46$). Although the mean difference between the estimated TST and reference TST increased from -9.7 min to -21.9 min, the SD decreased from 101.0 min to 55.7 min. Korkalainen et al. (2020) reported a mean difference of -12.2 min ($\pm$52.9 min) and Dietz-Terjung et al. (2021) an overestimation of 14 min and $\rho = 0.81$. These studies performed slightly better on a population with a similar AHI range as expected, since their sleep staging performances were higher as well. Nevertheless, these studies required longer input intervals for the algorithm, making them less suited for usage on unobtrusive technologies. Moreover, this study slightly underestimated the TST, which would result in an overestimated AHI. In general, slight overestimation has minor consequences compared to underestimation, as these patients would receive a diagnostic PSG as a follow-up procedure.

5.3 Detection of OSA Patients

Despite the fact that the CNN was trained for sleep-wake classification, its outcome contained information relevant for detection of OSA patients. As discussed in 3.4, more uncertain sleep-wake predictions were expected in the presence of apneic events, similar to the fact that the uncertainty of clinical sleep staging labels increases as well with the AHI of a patient. Additionally, there was an expected increase of sleep fragmentation, sympathetic activation and micro-awakenings related to apneas. As such, two metrics for detection of OSA patients were derived from the CNN outcome, namely the %Uncertain Sleep Epochs and %Sleep Wake Transitions. This improved interpretability of the network is beneficial when proposing the framework as a sleep diagnostics tool for OSA patients to clinicians. Another advantage was that OSA patient detection only relied on ECG and RIP signals, instead of including oxygen saturation sensors. A specificity of 89% was reached on the dataset Test, for detection of patients with AHI $\geqslant$ 15. However, the corresponding sensitivity was only 56% and $\kappa = 0.32$. In addition, mainly severe OSA patients were detected, as illustrated in figure 5). Indeed, when identifying an AHI $\geqslant$ 30, the specificity remained stable at 89%, but the sensitivity increased to 78% and κ to 0.67. This result is beneficial, as severe OSA patient indeed require prioritization for diagnostic PSG at the hospital. Additionally, detection of patients with many events as a first step is advantageous for future refined OSA severity categorisation. The reason is that severe OSA patients can have much stronger physiological dynamics compared to milder patients. This enables an OSA patient detection algorithm to focus training on patients with lower AHIs. It should be noted that one patient from Test with an AHI < 5 was falsely detected as being an OSA patient. For this, a follow-up over multiple nights could increase the OSA detection capabilities, as a single night recording might not be fully representative, due to accidental decreased data quality or the first night effect (Agnew Jr et al., 1966). If patients would consistently have values around the decision boundaries, it could indicate a pathological risk factor.

5.4 Future Work

To complete the proposed framework for OSA patient detection, apneic event detection from a minimal set of sensors is desired. This could be achieved by analyzing the SpO2 signal (Deviaene et al., 2018; Mendonça et al., 2020) or the cardiac and respiratory signals, which are already included in the current

sensor set (Feng et al., 2020; Deviaene et al., 2021). Wearable trackers from several commercial companies already provide these signals, such as Fitbit (2021), Garmin (2021) and Apple (2021) . The number of apneic events could then be combined with the sleep-wake staging to calculate the patient's AHI and provide feedback on the OSA severity.

In addition, the algorithmic pipeline for sleep-wake classification and OSA patient detection requires further validation on unobtrusive data, as the presented study used PSG signals. This was partially performed by Huysmans et al. (2020). However, this unobtrusive dataset was limited in number of subjects. Additionally, it only applied unobtrusively acquired ECG in combination with RIP from PSG. Thus, when accomodating the CNN to recordings from a different respiratory sensor, transfer learning of the new CNN is proposed. For this, the unimodal RIP network with the pretrained weights (see 3.3.1) is updated with the new data using a very low learning rate. A small learning rate allows the model to learn an optimal set of weights. This retrained RIP network is then recombined into the multimodal network, after which the dense layers are retrained. A smaller number of subjects is required as the model was pretrained.

The presented framework could also benefit from training with a larger dataset to improve sleep-wake classification performance. Moreover, extending the problem to classes wake-NREM-REM could increase the relevance of the network and deliver insight into REM-related apneic events. These events are still being researched for their adverse effects on cardiac comorbidities (Aurora et al., 2018; Varga and Mokhlesi, 2019).

Furthermore, the application domain of confident epochs could be further extended. For example, the percentage of confident predicted epochs in a patient's recording could serve as a data quality indicator. In a sleep study recording patients over multiple nights, it is expected that this percentage would remain relatively stable for a subject. An outlier value could indicate a recording from different quality and instability of the percentage or a constant low percentage could even indicate sleep problems.

6 CONCLUSION

Standard clinical procedures for sleep monitoring rely on uncomfortable and burdensome electroencephalography analysis. On the other hand, cardiac and respiratory signals have a great potential for comfortable sleep monitoring at home as unobtrusive sensors can record these. However, most unobtrusive sensors suffer from data loss and sensitivity to movement artefacts, especially in OSA patients. In addition, state-of-the-art sleep staging algorithms require long temporal dependencies, which cannot be garantueed in unobtrusive data. Therefore, this study developed a sleep-wake classifier to estimate the TST of suspected OSA patients based on single short-term (30s) segments of their cardiac and respiratory signals. Application of the network on healthy, mild and moderate sleep apnea patients resulted in moderate κ scores of 0.51, 0.49 and 0.48. Furthermore, two metrics derived from the sleep-wake classifier's outcome were applied for detecting OSA patients in an unseen test set with patients of varying AHI. As such, severe OSA patients (AHI $\geqslant$ 30) were detected in the unseen dataset with a sensitivity of 78% and specificity of 89%. Additional TST estimation was irrelevant for these detected patients, as they are directly prioritized for a clinical diagnostic test. Thus, their AHI estimation at home becomes redundant. Moreover, after excluding these severe patients, the overall accuracy of TST estimates increased to a mean bias error of 21.9 ($\pm$ 55.7) minutes and Pearson correlation of 0.74 to the reference. As this patient detection was only based on cardiac and respiratory inputs, it might enable comfortable and fast prioritization of OSA patients for a diagnostic PSG. Overall, the presented framework offered a realistic tool for unobtrusive monitoring of sleep apnea patients.

AUTHOR CONTRIBUTIONS

Conceptualization, D.H., C.V.; Methodology, D.H., C.V. ; Software, D.H. ; Validation, D.H.; Formal Analysis, D.H.; Investigation, D.H.; Resources, P.B., D.T., B.B.; Data Curation, P.B., D.T., B.B.; Writing – Original Draft Preparation, D.H.; Writing – Review & Editing, D.H., P.B., D.T., B.B., S.V.H., C.V.; Visualization, D.H.; Supervision, S.V.H., C.V.; Project Administration, S.V.H., C.V.; Funding Acquisition, S.V.H., C.V.

FUNDING AND ACKNOWLEDGMENTS

Bijzonder Onderzoeksfonds KU Leuven (BOF) Prevalentie van epilepsie en slaapstoornissen in de ziekte van Alzheimer: C24/18/097 ; Fonds voor Wetenschappelijk Onderzoek-Vlaanderen (FWO) PhD/Postdoc grants ; Agentschap Innoveren en Ondernemen (VLAIO) 150466: OSA+ ; KU Leuven Stadius acknowledges the financial support of imec ; EU: EU H2020 FETOPEN 'AMPHORA' #766456, EU H2020 MSCA-ITN-2018: 'INtegrating Magnetic Resonance SPectroscopy and Multimodal Imaging for Research and Education in MEDicine (INSPiRE-MED)', funded by the European Commission under Grant Agreement #813120, EU H2020 MSCA-ITN-2018: 'INtegrating Functional Assessment measures for Neonatal Safeguard (INFANS)', funded by the European Commission under Grant Agreement #813483 ; EIT 19263 – SeizeIT2: Discreet Personalized Epileptic Seizure Detection Device ; Flemish Government: This research received funding from the Flemish Government (AI Research Program). Sabine Van Huffel, Carolina Varon and Dorien Huysmans are affiliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium.

REFERENCES

Agnew Jr, H., Webb, W. B., and Williams, R. L. (1966). The first night effect: an eeg studyof sleep. *Psychophysiology* 2, 263–266

[Dataset] Apple (2021). Apple SpO2. https://www.apple.com/watch/

Aurora, R. N., Crainiceanu, C., Gottlieb, D. J., Kim, J. S., and Punjabi, N. M. (2018). Obstructive sleep apnea during rem sleep and cardiovascular disease. *American journal of respiratory and critical care medicine* 197, 653–660

Bakker, J. P., Ross, M., Vasko, R., Cerny, A., Fonseca, P., Jasko, J., et al. (2021). Estimating sleep stages using cardiorespiratory signals: validation of a novel algorithm across a wide range of sleep-disordered breathing severity. *Journal of Clinical Sleep Medicine* , jcsm–9192

Bassetti, C., Dogas, Z., and Peigneux, P. (2014). *Sleep medicine textbook* (European sleep research society)

Berry, R. B., Budhiraja, R., Gottlieb, D. J., Gozal, D., Iber, C., Kapur, V. K., et al. (2012). Rules for scoring respiratory events in sleep: update of the 2007 aasm manual for the scoring of sleep and associated events. *Journal of clinical sleep medicine* 8, 597–619

Deviaene, M., Castro, I., Borzée, P., Patel, A., Torfs, T., Buyse, B., et al. (2021). Capacitively-coupled ecg and respiration for the unobtrusive detection of sleep apnea. *Physiological Measurement*

Deviaene, M., Testelmans, D., Buyse, B., Borzée, P., Van Huffel, S., and Varon, C. (2018). Automatic screening of sleep apnea patients based on the spo 2 signal. *IEEE journal of biomedical and health informatics* 23, 607–617

Dietz-Terjung, S., Martin, A. R., Finnsson, E., Ágústsson, J. S., Helgason, S., Helgadóttir, H., et al. (2021). Proof of principle study: diagnostic accuracy of a novel algorithm for the estimation of sleep stages and disease severity in patients with sleep-disordered breathing based on actigraphy and respiratory inductance plethysmography. *Sleep and Breathing* , 1–8

Douglas, N. J., White, D. P., Pickett, C. K., Weil, J. V., and Zwillich, C. (1982). Respiration during sleep in normal man. *Thorax* 37, 840–844

Feng, K., Qin, H., Wu, S., Pan, W., and Liu, G. (2020). A sleep apnea detection method based on unsupervised feature learning and single-lead electrocardiogram. *IEEE Transactions on Instrumentation and Measurement* 70, 1–12

[Dataset] Fitbit (2021). Fitbit SpO2. https://www.fitbit.com/global/us/technology/health-metrics

Flemons, W. W., Douglas, N. J., Kuna, S. T., Rodenstein, D. O., and Wheatley, J. (2004). Access to diagnosis and treatment of patients with suspected sleep apnea. *American journal of respiratory and critical care medicine* 169, 668–672

Fonseca, P., van Gilst, M. M., Radha, M., Ross, M., Moreau, A., Cerny, A., et al. (2020). Automatic sleep staging using heart rate variability, body movements, and recurrent neural networks in a sleep disordered population. *Sleep* 43, zsaa048

[Dataset] Garmin (2021). Garmin SpO2. https://www.garmin.com/en-US/

Guilleminault, C., Winkle, R., Connolly, S., Melvin, K., and Tilkian, A. (1984). Cyclical variation of the heart rate in sleep apnoea syndrome: Mechanisms, and usefulness of 24 h electrocardiography as a screening technique. *The Lancet* 323, 126–131

Huysmans, D., Heffinck, E., Castro, I., Deviaene, M., Borzee, P., Buyse, B., et al. (2020). Sleep-wake classification for home monitoring of sleep apnea patients. In *Proc. of the 47th Annual Computing In Cardiology Conference* (IEEE), Page–1

Kimoff, R. J. (1996). Sleep fragmentation in obstructive sleep apnea. *Sleep* 19, S61–S66

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*

Korkalainen, H., Aakko, J., Duce, B., Kainulainen, S., Leino, A., Nikkonen, S., et al. (2020). Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea. *Sleep* 43, zsaa098

Malik, J. et al. (2018). Sleep-wake classification via quantifying heart rate variability by convolutional neural network. *Physiol. Meas* 39, 085004

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica* 22, 276–282

[Dataset] Medatec (2021). Medatec. https://www.medatec.eu/en/sleep

Mendonça, F., Mostafa, S. S., Morgado-Dias, F., and Ravelo-García, A. G. (2020). An oximetry based wireless device for sleep apnea detection. *Sensors* 20, 888

Moeyersons, J., Amoni, M., Van Huffel, S., Willems, R., and Varon, C. (2019). R-deco: An open-source matlab based graphical user interface for the detection and correction of r-peaks. *PeerJ Computer Science* 5, e226

Norman, R. G., Pal, I., Stewart, C., Walsleben, J. A., and Rapoport, D. M. (2000). Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep* 23, 901–908

Pichot, V., Roche, F., Celle, S., Barthélémy, J.-C., and Chouchou, F. (2016). Hrvanalysis: a free software for analyzing cardiac autonomic activity. *Frontiers in physiology* 7, 557

Radha, M. et al. (2019). Sleep stage classification from heart-rate variability using long short-term memory neural networks. *Scientific Reports* 9, 1–11

Rechtschaffen, A. and Kales, A. (1968). A manual of standardized terminology, techniques, and scoring system for sleep stages for human subjects. *National Institute of Health* 204

Sateia, M. J. (2014). International classification of sleep disorders. *Chest* 146, 1387–1394

Senaratna, C. V. et al. (2017). Prevalence of obstructive sleep apnea in the general population: A systematic review. *Sleep Medicine Reviews* 34, 70 – 81

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1929–1958

Varga, A. W. and Mokhlesi, B. (2019). Rem obstructive sleep apnea: risk for adverse health outcomes and novel treatments. *Sleep and Breathing* 23, 413–423

Varon, C. and Van Huffel, S. (2017). Complexity and nonlinearities in cardiorespiratory signals in sleep and sleep apnea. In *Complexity and Nonlinearity in Cardiovascular Signals* (Springer). 503–537

Willemen, T. et al. (2015). Probabilistic cardiac and respiratory based classification of sleep and apneic events in subjects with sleep apnea. *Physiol. Meas* 36, 2103

Young, T., Peppard, P. E., and Gottlieb, D. J. (2002). Epidemiology of obstructive sleep apnea: a population health perspective. *American journal of respiratory and critical care medicine* 165, 1217–1239

7 FIGURES AND TABLES

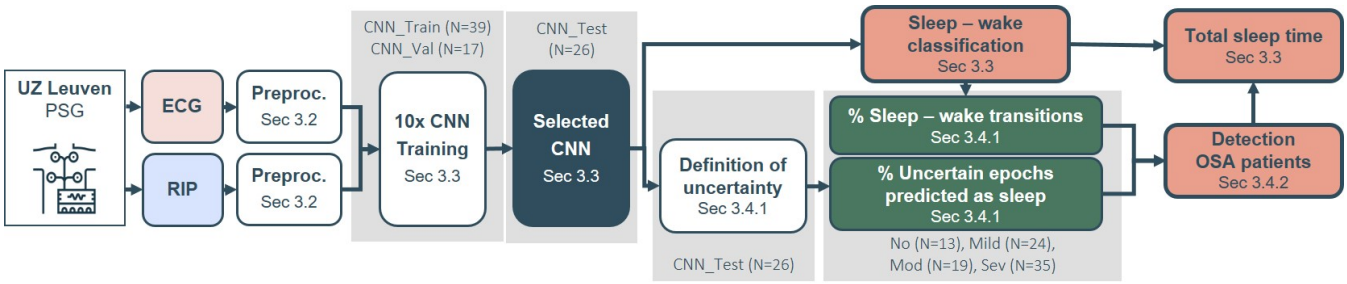

Figure 1. Framework pipeline from polysomnography (PSG) data to Total Sleep Time (TST) estimation and detection of OSA patients. First, the electrocardiogram (ECG) and respiratory inductance plethysmogram (RIP) data were preprocessed (section 3.2). The classifier's architecture was based on a convolutional neural network (CNN). Its training procedure and derived TST outcome is reported in section 3.3. The relation between uncertainties and sleep-wake transitions in a patient's prediction and the obstructive sleep apnea (OSA) severity were studied in section 3.4.1. These relations were subsequently applied for detection of OSA patients (section 3.4.2). The grey boxes indicate the datasets used for parameter optimization or selection.

....

Dataset	N	AHI (1/h)	Average Nr. of Epochs (% of Total Nr. Epochs)				
	(AHI<5)	Mean (SD)	N3	N2	N1	REM	Wake
CNN_Train	39 (13)	5.9 (2.2)	230 (22)	412 (40)	38 (4)	172 (17)	175 (17)
CNN_Val	17 (9)	5.1 (2.6)	253 (24)	393 (36)	39 (4)	159 (15)	236 (22)
CNN_Test	26 (13)	4.8 (2.6)	221 (22)	416 (40)	44 (4)	182 (17)	168 (17)
No	13	2.2 (1.3)	233 (22)	444 (42)	46 (4)	196 (18)	147 (14)
Mild	24	8.7 (2.9)	186 (19)	388 (39)	50 (5)	148 (15)	235 (23)
Mod	19	20.4 (4.3)	189 (19)	391 (38)	55 (5)	139 (14)	239 (24)
Sev	35	61.9 (20.1)	120 (12)	383 (39)	139 (14)	111 (12)	236 (24)
Test	36	38.6 (29.6)	112 (11)	406 (39)	104 (11)	128 (12)	281 (27)

N: Number of subjects in dataset, with number of No OSA indicated in brackets. AHI: Apnea-Hypopnea Index. No: AHI < 5, Mild: $5 \leqslant$ AHI < 15, Mod: $15 \leqslant$ AHI < 30, Sev: AHI $\geqslant 30$.

Table 1. Demographic and clinical information of the study datasets. The whole dataset comprised 183 patients. Training of the sleep-wake classifier was performed on 39 patients (CNN_Train) and validated on 17 patients (CNN_Val), all with AHI < 10. The trained classifier was tested with CNN_Test. In addition, dataset CNN_Test was merged with patients with higher AHI and split again according to OSA severity in the subsets No, Mild, Mod and Sev. Out of 183 patients, 36 patients were left out as an unseen dataset Test for independent validation of sleep-wake classification, total sleep time estimation and detection of OSA patients.

	Data	κ		TST		
		Median	SD	Mean(Δ)	SD(Δ)	ρ
ECG	CNN_Test	0.31	0.15	47 min	88 min	0.21
RIP	CNN_Test	0.46	0.17	52 min	60 min	0.72
	CNN_Train	0.53	0.16	2 min	48 min	0.83
	CNN_Val	0.48	0.21	10 min	71 min	0.44
ECG	CNN_Test	0.51	0.13	0 min	50 min	0.78
+	No	**0.51**	0.14	-9 min	**34** min	0.90
RIP	Mild	**0.49**	0.16	-9 min	**65** min	0.60
	Mod	**0.48**	0.17	-10 min	**75** min	0.42
	Sev	**0.36**	0.20	26 min	**86** min	0.56
	Test	0.32	0.18	-9.7 min	101 min	0.46

TST: Total Sleep Time. Δ: (Estimated TST - Reference TST).
ρ: Pearson's correlation coefficient.

Table 2. Sleep-wake classification of the unimodal and multimodal networks and TST estimates. **Left**: Performance score κ, calculated with decision threshold 50%. The indicated outcomes show a similar performance for No, Mild and Mod. **Right**: The estimated TST was compared to the TST derived from the expert annotations of the tested epochs by the mean and SD of the difference, and ρ. The indicated outcomes show an increase in SD(Δ) for increasing AHI. It demonstrates a decrease in reliability of the outcome.

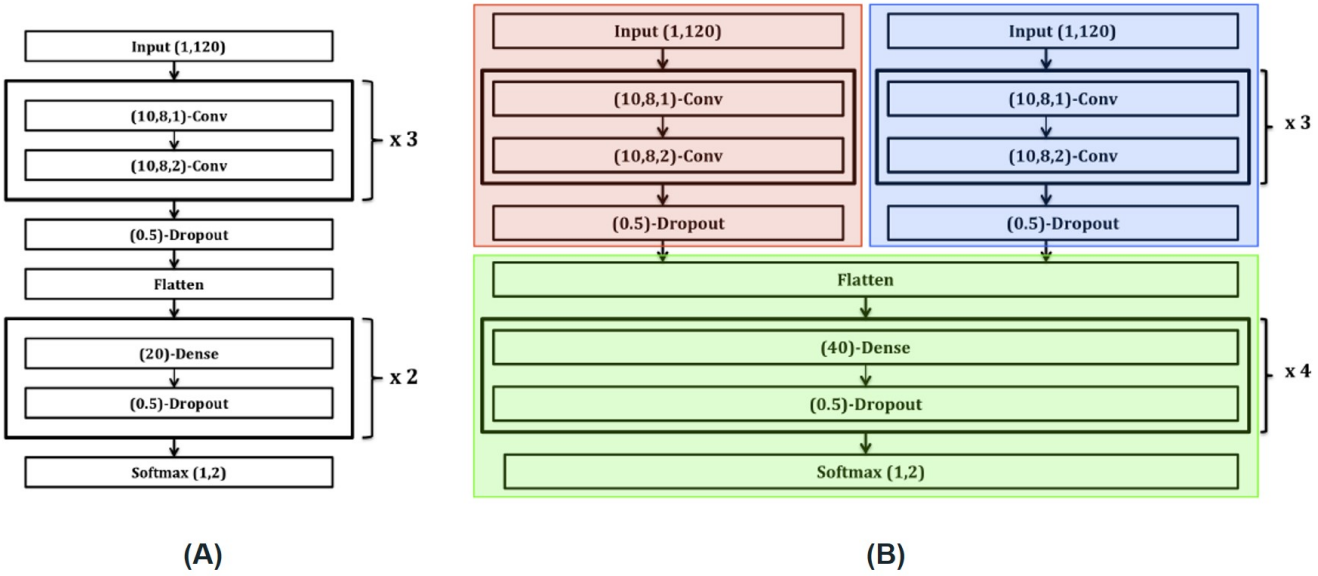

Figure 2. **(A)** Illustration of the network architecture used for the cardiac and the respiratory signals separately. This CNN extracted characteristics from the input epoch with length 120 samples. The layer $(\mathrm{f}, \mathrm{k}, \mathrm{s}) - \mathrm{Conv}$ indicates a layer with f filters, kernel size l and stride s, and an activation of type ReLu. After the convolutional block, dense layers, $(\mathrm{n}) - \mathrm{Dense}$, with n neurons were included. A third type were dropout layers, $(\mathrm{d}\%) - \mathrm{Dropout}$, where $\mathrm{d}\% = 50\%$ of the weights were equal to zero in every training step to avoid overfitting. The output layer is a softmax layer, $\mathrm{Softmax}(1, \mathrm{c})$, delivering posterior class probabilities for every one of the $\mathrm{c} = 2$ classes, where class 0 represented Sleep and class 1 Wake. **(B)** Architecture of the combined deep learning network, consisting of a cardiac CNN (red), a respiratory CNN (blue) and a combined dense network (green). During training, the weights of convolutional branches were frozen, while the dense layers were optimized.

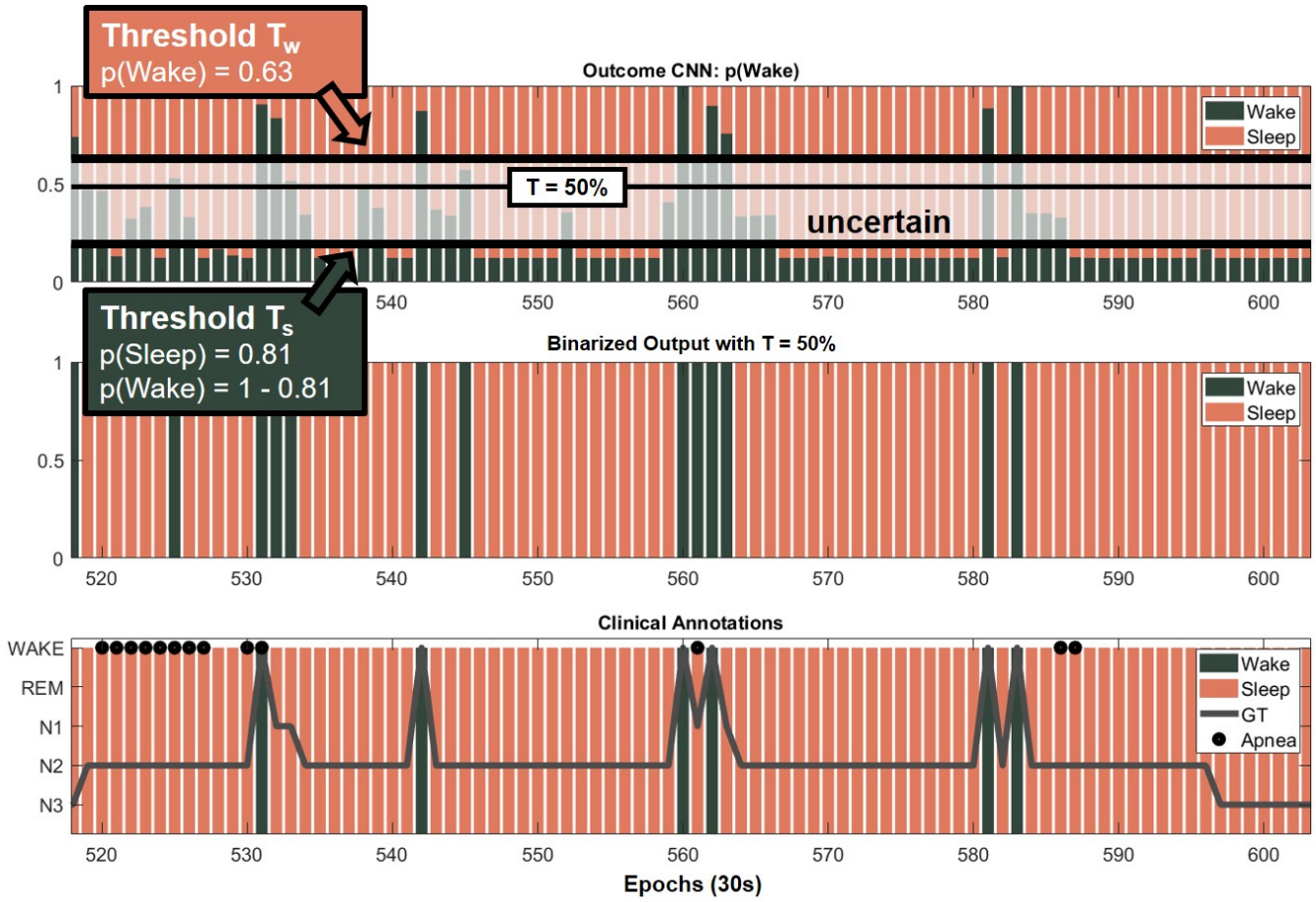

Figure 3. Definition of uncertain epochs. The top row represents the outcome of the CNN of each epoch as $\mathrm{p(Wake)}$ (green) and $\mathrm{p(Sleep)} = 1 - \mathrm{p(Wake)}$ (red) (see section 3.3.1). A distinction was made between *confident* and *uncertain* predicted epochs by defining confidence thresholds. The second row shows the predicted sleep-wake classification with the default threshold at 50%. The last row shows the ground truth (GT) sleep stages which were clinically annotated.

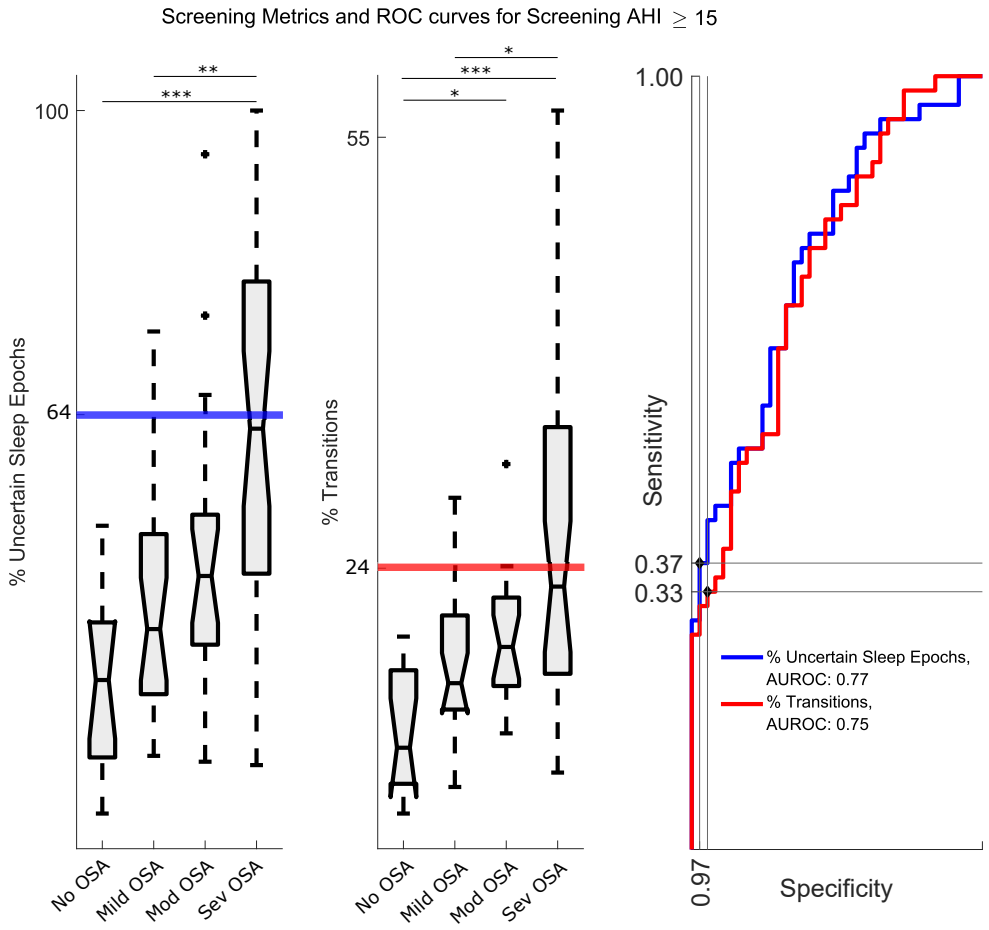

Figure 4. Left and Middle: The distribution of %Uncertain Sleep Epochs and %Sleep Wake Transitions for No, Mild, Mod and Sev. The increasing trends supported the application of the metrics for detection of OSA patient (i.e. with an AHI$\geqslant$ 15). The thick lines represent the chosen thresholds. The level of statistical significance is marked with one asterisk (*) if p$<$0.05, two (**) if p$<$0.01 and three (***) if p$<$0.001. **Right :** Separate ROC curves for both metrics. Indicated dots were the chosen operating points and the corresponding thresholds are indicated on the left plots. For %Uncertain Sleep Epochs, an operating point on the ROC curve was chosen where the specificity was $\geqslant$ 95%, since a larger specificity for OSA patient detection was preferred. As such, a threshold of 64% was selected, at which specificity reached 97% and sensitivity 37%. For %Sleep Wake Transitions a threshold of 24% was selected, at which a specificity of 95% and sensitivity of 33% was obtained. AUROC: Area Under the ROC curve.

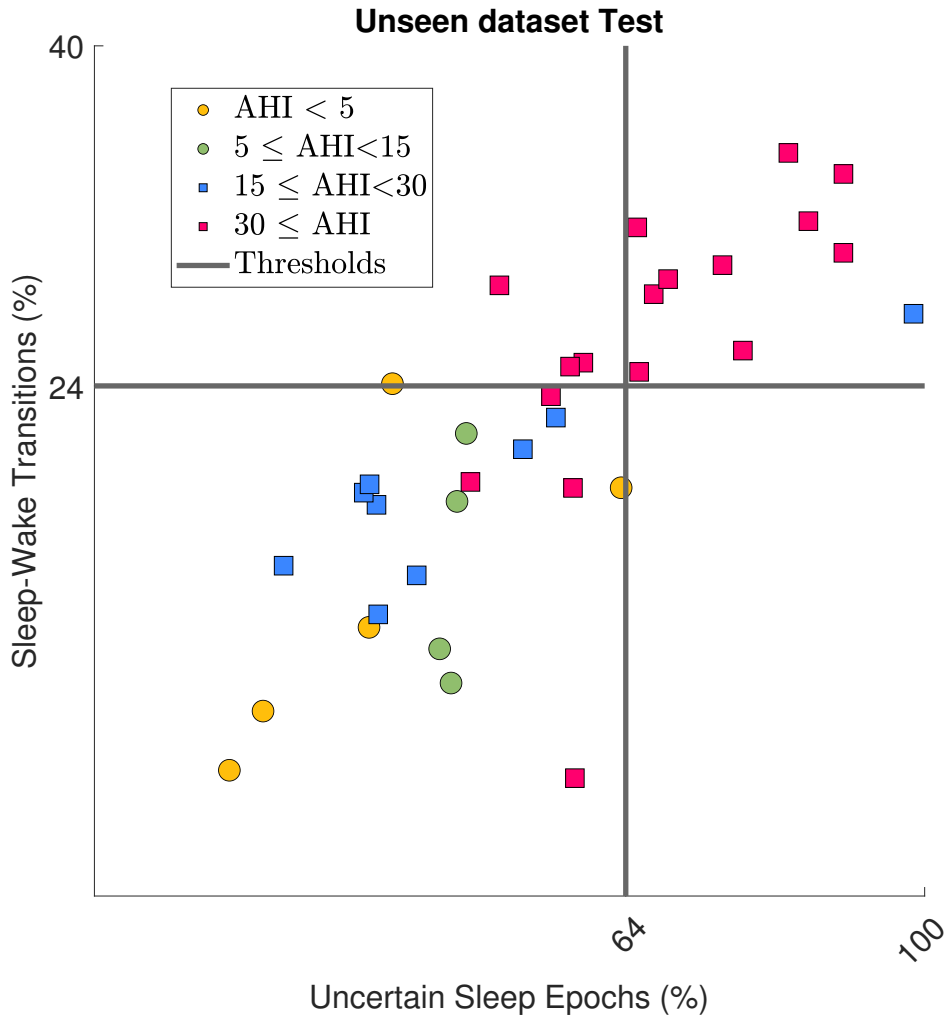

Figure 5. Application of two detection metrics on dataset Test where circles represent patients with AHI < 15 and squares AHI $\geqslant 15$. Grey lines represent the corresponding metric thresholds. Severe OSA patients (AHI $\geqslant 30$) were detected correctly by the metrics, as expected.

	Data	p(Sleep)		p(Wake)		% USE			% UWE		
		Med	SD	Med	SD	Med (SD)	NA	A	Med (SD)	NA	A
ECG	CNN_Test	0.65	0.01	0.59	0.03	42 (10)	40	2	41 (8)	39	2
RIP	CNN_Test	0.89	0.02	0.67	0.05	37 (13)	34	2	34 (12)	32	2
	CNN_Train	0.87	0.09	0.69	0.06	37 (19)	33	3	40 (11)	35	3
	CNN_Val	0.80	0.10	0.67	0.05	51 (21)	47	3	40 (11)	37	1
ECG	CNN_Test	0.87	0.06	0.69	0.06	33 (12)	31	3	40 (9)	37	2
+	No	0.87	0.02	0.69	0.07	**33 (11)**	32	1	40 (9)	38	2
RIP	Mild	0.87	0.08	0.69	0.04	**39 (13)**	34	4	40 (9)	35	5
	Mod	0.86	0.09	0.68	0.07	**45 (17)**	31	13	41 (8)	31	7
	Sev	0.66	0.11	0.65	0.04	**62 (21)**	21	39	47 (9)	20	27

M: Median value.
USE: Uncertain Sleep Epochs. UWE: Uncertain Wake Epochs. NA: No apneic events. A: With apneic events.

Table 3. Uncertainty in the sleep-wake classifier outcome, averaged over individual patient outcomes. **Left**: Prediction probability of sleep predicted epochs ($\mathrm{p(Sleep)} = 1 - \mathrm{p(Wake)}$ with $\mathrm{p(Wake)} < 50\%$), and wake predicted epochs ($\mathrm{p(Wake)} \geqslant 50\%$). The values indicated in green were used to calculate $\mathrm{T_s}$ and red for $\mathrm{T_w}$. **Right**: Percentage of uncertain epochs predicted as sleep or wake, derived using $\mathrm{T_s}$ and $\mathrm{T_w}$. From the indicated values it can be seen that $\%\mathrm{Uncertain\ Sleep\ Epochs}$ increased with OSA severity, supporting the usage of this metric as an indicator of apneic severity. Furthermore, a distinction was made between uncertain epochs during which an apneic event occured (A) and those without (NA), indicated with the mean value over all subjects.