Computers in Biology and Medicine 100 (2018) 123–131

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Real-time apnea-hypopnea event detection during sleep by convolutional neural networks

Sang Ho Choi[a], Heenam Yoon[a], Hyun Seok Kim[a], Han Byul Kim[a], Hyun Bin Kwon[a], Sung Min Oh[b], Yu Jin Lee[b], Kwang Suk Park[c,*]

[a] Interdisciplinary Program in Bioengineering, Seoul National University, Seoul, South Korea
[b] Department of Neuropsychiatry and Center for Sleep and Chronobiology, Seoul National University Hospital, Seoul, South Korea
[c] Department of Biomedical Engineering, College of Medicine, Seoul National University, South Korea

ARTICLE INFO

Keywords:
Apnea-hypopnea event detection
Convolutional neural networks
Real-time monitoring
Sleep apnea and hypopnea syndrome diagnosis
Nasal pressure signal

ABSTRACT

Sleep apnea-hypopnea event detection has been widely studied using various biosignals and algorithms. However, most minute-by-minute analysis techniques have difficulty detecting accurate event start/end positions. Furthermore, they require hand-engineered feature extraction and selection processes. In this paper, we propose a new approach for real-time apnea-hypopnea event detection using convolutional neural networks and a single-channel nasal pressure signal. From 179 polysomnographic recordings, 50 were used for training, 25 for validation, and 104 for testing. Nasal pressure signals were adaptively normalized, and then segmented by sliding a 10-s window at 1-s intervals. The convolutional neural networks were trained with the data, which consisted of class-balanced segments, and were then tested to evaluate their event detection performance. According to a segment-by-segment analysis, the proposed method exhibited performance results with a Cohen's kappa coefficient of 0.82, a sensitivity of 81.1%, a specificity of 98.5%, and an accuracy of 96.6%. In addition, the Pearson's correlation coefficient between estimated apnea-hypopnea index (AHI) and reference AHI was 0.99, and the average accuracy of sleep apnea and hypopnea syndrome (SAHS) diagnosis was 94.9% for AHI cutoff values of ≥5, 15, and 30 events/h. Our approach could potentially be used as a supportive method to reduce event detection time in sleep laboratories. In addition, it can be applied to screen SAHS severity before polysomnography.

1. Introduction

Sleep apnea and hypopnea syndrome (SAHS) is a common sleep breathing disorder characterized by repetitive events of complete (apnea) or partial (hypopnea) cessation of breathing during sleep [1]. SAHS disrupts sleep architecture and leads to several symptoms, such as daytime sleepiness, tiredness, somnolence, neurocognitive deficits, and inattention [2,3]. In addition, SAHS is a risk factor for several complications, such as cardiac arrhythmias [3], coronary artery disease [4], stroke [5], hypertension [6], depression [7], and diabetes mellitus [8]. Early detection and treatment of SAHS is needed to reduce the occurrence of these health problems.

According to an American Academy of Sleep Medicine (AASM) manual [9], sleep apnea is scored when there is a 90% or more reduction in the pre-event baseline of the oronasal thermal sensor signal amplitude, and the duration of events in the signal is longer than 10 s [9]. Hypopnea is scored when all of the following is observed: 1) a 30% or more reduction in the pre-event baseline of the nasal pressure signal amplitude that lasts longer than 10 s, accompanied by 2) 3% or greater oxygen desaturation from the pre-event baseline, or the event is associated with electroencephalography (EEG) arousal [9]. Polysomnography (PSG) is considered to be the gold standard method for diagnosing SAHS. However, PSG is expensive, requires well-controlled facilities and equipment, and is inconvenient, because of the large number of sensors attached to the subject's body. In addition, the manual scoring of apnea-hypopnea (AH) events is laborious, time-consuming, and requires trained sleep experts. Therefore, it is necessary to develop alternative methods that use fewer physiological signals and provide simplified automatic event detection.

Numerous alternative methods have been proposed for overcoming the drawbacks of PSG. For example methods have been proposed that detect events based on several biosignals, such as electrocardiogram (ECG) signals [10–15], airflow signals [16–19], pulse oximetry (SpO_2) signals [20,21], and combinations of signals [22]. In recent years,

* Corresponding author. Department of Biomedical Engineering, College of Medicine, Seoul National University, Yongon-dong Chongno-gu Seoul, 03080, South Korea.
E-mail address: pks@bmsil.snu.ac.kr (K.S. Park).

https://doi.org/10.1016/j.compbiomed.2018.06.028
Received 14 February 2018; Received in revised form 25 June 2018; Accepted 25 June 2018

several methods for detecting AH events in real time have been proposed. For example, Burgos et al. [21] implemented a real-time apnea monitoring system based on SpO_2, using a decision tree. Bsoul et al. [13] proposed a real-time sleep monitoring system based on ECG using support vector classifier. Xie et al. [22] proposed a real-time SAHS detection method based on ECG and SpO_2 by classifier combination. These proposed methods perform minute-by-minute analysis of apneic events, identifying whether an event is present in each minute. However, multiple AH events can occur in 1 min (epoch), and one event can be prolonged over multiple epochs. Therefore, a window overlapping method is required to detect AH events more accurately, including these cases, and give a precise apnea-hypopnea index (AHI) that represents SAHS severity.

Several methods have been used for AH event detection, such as thresholding [16,17,23], linear discriminant analysis [24], support vector machines [11,13,25], and neural networks [10,26]. Almost all of these methods utilize the following procedure: 1) preprocessing, 2) feature extraction, 3) feature selection, and 4) classification. In conventional methods, the design of a feature extractor requires considerable domain expertise. Such extractors are used to extract predictive parameters from a signal, and then construct an optimal combination of features that can be input into the classifier. Unlike conventional methods, deep learning [27] automatically learns and find features from the input signals without using any domain knowledge. Convolutional neural networks (CNNs) are the most widely used deep learning method and have achieved state-of-the-art performance in the image recognition field [28–30]. Recently, CNNs have been applied to biosignal classification problems, such as EEG [31], electrooculography (EOG) [32], and ECG [33,34]. Haidar et al. [44] proposed an apnea-hypopnea events detection method using CNNs. However, our proposed method differs from their proposed method. In their method, they assess the ability of the CNNs to detect sleep apneic events in non-overlapped 30 s intervals and do not evaluate the event detection performance for the entire sleep recording time. In addition, they provide no information on the CNN model optimization process.

In this study, we assessed the ability of CNNs to detect AH events in real time without using any hand-engineered features. The main contributions of this paper are as follows:

- An optimal CNN architecture that automatically extracts AH event features from nasal pressure signals is presented.
- Overlapping nasal pressure signal segments are used to detect AH events more precisely.
- Event detection performance is investigated via segment analysis and AHI estimation analysis.
- We compare the SAHS diagnostic performance with previous studies to prove CNN model effectiveness.

2. Material and methods

2.1. Subjects and polysomnography

Two different datasets were used in this study. The first dataset is an internal dataset, which consists of overnight PSG recordings from subjects at the Center for Sleep and Chronobiology, Seoul National University Hospital. We only included subjects whose ages were ≥20 years. The exclusion criteria were presence of 1) major behavioral or mental disorders, 2) unstable vital signs, 3) arrhythmia or heart failure, and 4) sleep disorders other than SAHS (e.g., periodic limb movements during sleep, sleepwalking, night terrors, and rapid eye movement sleep behavior disorders). One hundred and twenty nine PSG recordings that met the inclusion and exclusion criteria were used to study. Each PSG data recordings was acquired from different subjects. The Institutional Review Board of Seoul National University Hospital approved this retrospective study (IRB No. H-1607-146-778). PSG data were recorded using a standard PSG routine [9], and the following physiological parameters were collected: EEG at positions F4-M1, C4-M1, and O2-M1, electromyogram (EMG) at the chin and bilateral tibialis anterior muscles, a bilateral EOG, ECG at lead II, oronasal airflow using a thermistor, nasal pressure using a cannula/pressure transducer (PTAF 2, Pro-Tech, Woodinville, WA, USA), thoracic and abdominal volume changes using piezoelectric sensors, body position using a tri-axis accelerometer, snoring sounds using a microphone, and SpO_2 using a pulse oximeter. All channels were recorded with a NEUVO system (Compumedics Ltd., Victoria, Australia). After PSG recording, sleep stages and AH events were scored by sleep technologists and verified by a sleep clinician according to the 2012 AASM manual (version 2.0) [9]. Event start/end points were marked from the nadir preceding the first breath that is clearly reduced to the beginning of the first breath that approximates the baseline breathing amplitude [9]. For the results of AH events scoring, on average, inter-scorer agreement was 95.2%.

The second dataset consists of 50 PSG recordings in the Multi-Ethnic Study of Atherosclerosis (MESA) [47,48]. We selected 50 PSG recordings from the MESA dataset that have a high signal quality score and were not included in the exclusion criteria. Of the 179 PSG recordings consisting of internal and external datasets, 50 recordings were used for training, 25 recordings were used for model validation, and the remaining 104 recordings (54 recordings from the internal dataset, 50 recordings from the external dataset) were used for independent testing.

The number of apnea and hypopnea events per hour of sleep was defined as the AHI. Based on the AHI value, the apnea-hypopnea severity of each subject was classified as non-SAHS (AHI < 5 events/h), mild SAHS (5 ≤ AHI < 15 events/h), moderate SAHS (15 ≤ AHI < 30 events/h), or severe SAHS (AHI ≥ 30 events/h). Four groups were formed, consisting of 38 (non-SAHS), 41 (mild SAHS), 50 (moderate SAHS), and 50 (severe SAHS) subjects, respectively. The demographic and sleep-related parameters for the four groups are summarized in Table 1.

According to the AASM manual, an oronasal thermal airflow sensor is recommended for identifying apnea. Alternatively, a pressure transducer can be used. However, a nasal pressure transducer sensor is particularly recommended for identifying hypopnea. Based on these guidelines, and considering that a pressure sensor can reflect respiration change from hypopnea more rapidly than a thermal sensor, we used nasal pressure sensor signals to detect AH events.

2.2. Signal preprocessing and segmentation

The proposed CNN method, depicted in Fig. 1, consists of three phases: 1) An adaptively normalized preprocessed signal is segmented and fed to the CNNs. 2) After features are automatically extracted in the convolution and max-pooling layers, a segment is classified from the

Table 1
Summary of sleep-related variables and subject characteristics (Mean ± Standard deviation).

Variable \Group	Non-SAHS	Mild SAHS	Moderate SAHS	Severe SAHS
N (M/F)	38 (14/24)	41 (20/21)	50 (33/17)	50 (42/8)
Age (years)	40.2 ± 13.4	51.0 ± 15.0	50.5 ± 13.2	57.4 ± 14.7
BMI (kg/m^2)	22.8 ± 3.1	24.8 ± 2.9	25.5 ± 3.7	27.0 ± 5.3
TRT (min)	459.5 ± 63.4	496.8 ± 55.2	477.0 ± 27.6	487.2 ± 42.8
SE (%)	89.0 ± 6.5	84.0 ± 10.3	87.0 ± 8.4	80.2 ± 10.1
SOL (min)	11.8 ± 12.7	14.8 ± 20.7	9.0 ± 14.9	11.0 ± 19.0
AHI (events/h)	2.7 ± 1.3	9.8 ± 2.6	21.4 ± 4.4	54.5 ± 18.9

BMI, body mass index; TRT, total recording time; SE, sleep efficiency; SOL, sleep onset latency; AHI, apnea-hypopnea index.
Subjects were classified as non-SAHS (AHI < 5 events/h), mild SAHS (5 ≤ AHI < 15 events/h), moderate SAHS (15 ≤ AHI < 30 events/h), and severe SAHS (AHI ≥ 30 events/h).

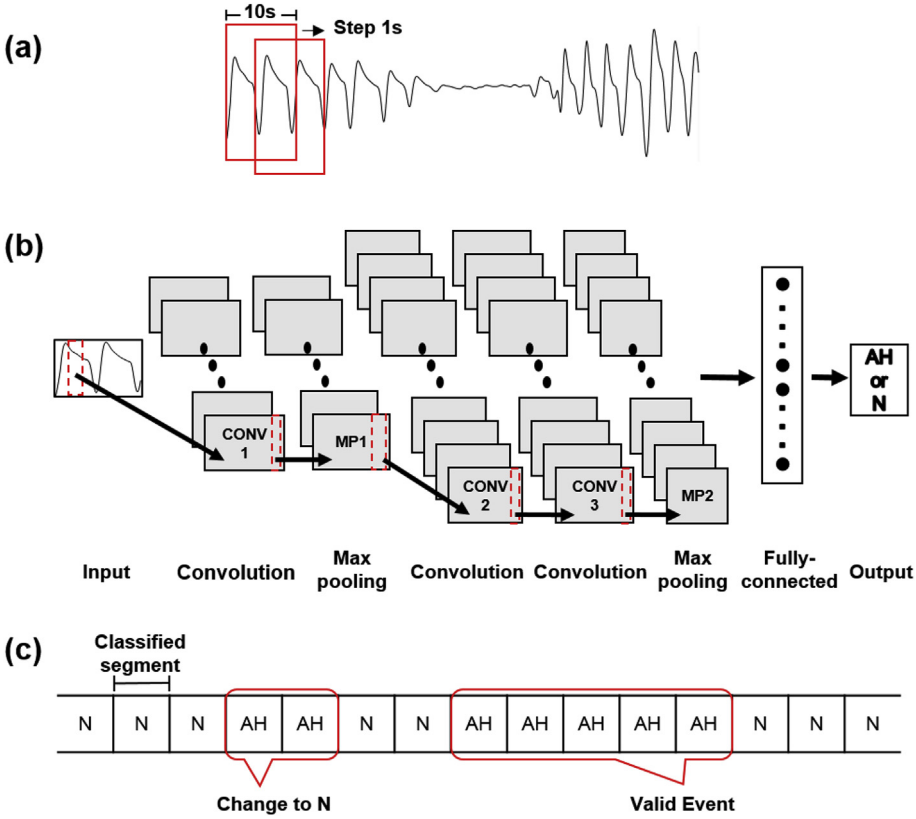

Fig. 1. Overview of proposed CNN method: (a) Data segmentation from normalized signal. (b) CNN architecture. (c) Event detector.

fully connected layers. 3) From classified events, valid events are counted, and the AHI is computed.

To train and test the CNN model, signals were preprocessed and segmented. Each nasal pressure signal recording (originally sampled at 500 or 32 Hz) was downsampled at 16 Hz to reduce model training time. Then, the signals were filtered to reduce baseline drifts and high-frequency noise. Specifically, they were high-pass filtered at 0.01 Hz, and sequentially low-pass filtered at 3 Hz using fifth-order infinite impulse response (IIR) Butterworth filters. The transfer function coefficients were extracted from a low and high-pass Butterworth filter. Then, real-time filtering was carried out with sample-by-sample processing using these coefficients. During lengthy respiration recording times, signals are affected by body posture and movement. In this study, adaptive normalization method [35] was applied in order to save the part where the amplitude of respiration is small owing to the sleeping posture for a long time. From the filtered signal, the area and standard deviation of the signal were obtained for each second by using Equations (1) and (2). Then, the normalization factor was calculated using Equation (3), and the normalized signals obtained by dividing the signals by the normalization factor. The adaptive normalization method does not affect short breathing change. In this case, it functions like z-score normalization amplifying the amplitude because the adaptive normalization factor is calculated by using 5% of the new 1-s respiration parameters and 95% of the previous factor value, as shown in Equation (3).

$$A(k) = \frac{1}{fs} \sum_{i=k*fs}^{(k+1)*fs-1} abs(x(i)) \tag{1}$$

$$\sigma(k) = \sqrt{\frac{1}{fs-1} \sum_{i=k*fs}^{(k+1)*fs-1} (x(i) - \bar{x}(k))^2} \tag{2}$$

$$F_{Norm}(k) = \min\{0.95F_{Norm}(k-1) + 0.05A(k), 0.95F_{Norm}(k-1) + 0.05\sigma(k)\} \tag{3}$$

fs: number of samples in 1 s; $\bar{x}(k)$: average of the signal in 1 s.

The preprocessed signal was segmented from overlapping windows. Comparing the event detection performance with adjusting window size from 5 s to 10 s, 10 s showed the best performance. Therefore, we set the segment length to 10 s. Segments were categorized into two classes: AH and N. If at least 80% of a segment occurred within an apnea or hypopnea event period, it was labeled as class AH. Other cases were labeled as class N. Event covering percent was also optimized by comparing the results of percent adjustment from 50 to 100. The best event detection performance was obtained at 80%. For a training dataset, the recorded data were segmented using windows shifted by 1 s. Then, each segment was categorized by the aforementioned rules. Previous studies showed higher classification performance for CNNs trained using a class-balanced dataset rather than a class-imbalanced dataset [36]. In order to prevent the model from overfitting to the majority number of the class, the training set consisted of a balanced number for each class segment. For each subject segment, class segments with more segments were randomly subsampled, in accordance with the number of other class segments. The original imbalance ratio was 6.9, which is the number of N class segments divided by the number of AH class segments.

For real-time AH event estimation, validation and test set segments were extracted using windows shifted by 1 s at a time from the overall recording time. The segments were labeled using the same rules as the training data segments.

We developed and verified the model with the hold-out method because the number of segments was sufficient for model training and developing a generalized model. Of the 179 PSG recordings, 50 were used for training, 25 for model validation, and the remaining 104 for independent testing. The number of non-SAHS, mild SAHS, moderate SAHS, and severe SAHS recordings were, respectively, (8/10/16/16) for the training set, (4/5/8/8) for the validation set, and (26/26/26/26)

Table 2
Network configuration summary.

Layer	No. of units	Activation function	Kernel size	Stride	Output size
Input					(160, 1)
Convolutional1	15	ReLU	4	2	(79, 15)
Max-pooling1			2	1	(78, 15)
Convolutional2	30	ReLU	4	2	(38, 30)
Convolutional3	30	ReLU	4	2	(18, 30)
Max-pooling2			2	1	(17,30)
Fully connected1	50	ReLU			50
Output	1	Sigmoid			1

for the test set. Each recording was randomly selected from the group to make training/validation/test sets. For the training, validation, and test sets, the numbers of extracted segments were, respectively 358762 (AH: N = 179381: 179381), 699885 (AH: N = 84574: 615311), and 3008004 (AH: N = 334482: 2673522).

2.3. CNN background

In this study, we used one-dimensional CNNs to classify the time series respiration signals. The CNN layers consisted of convolutional, pooling, and fully connected layers.

1 Convolutional layers: These layers learn filters (kernels) that activate to a specific pattern in the input signal. The kernel is convolved with the input signal while sliding across the input signal. A feature map is output by the convolution operation, and the computation of one-dimensional convolutional layers is as follows:

$$x_k^l = f\left(b_k^l + \sum_{i=1}^{N} conv(w_k^{l-1}, y_i^{l-1})\right) \quad (4)$$

where $x_k{}^l$ is the kth feature map in layer l, $b_k{}^l$ is the bias of the kth feature map in layer l, $w_k{}^{l-1}$ is the kth convolutional kernel from all features in layer l-1 to the kth feature map in layer l, $y_i{}^{l-1}$ is the output of the ith feature map in layer l-1, and N is the number of elements in layer l-1. The symbol *conv* represents the vector convolution and $f(.)$ is the activation function [33].

2 Pooling layers: These layers are typically placed after convolutional layers, and simplify the information in the convolutional layers. They decrease the spatial dimensions and reduce the computational cost for successive layers. In addition, they ensure that the network learns features invariant of scale or orientation changes. The pooling operations slide a window over the previous feature map, taking either the maximum value from the values in the window (max-pooling) or the average of the values (average-pooling).
3 Fully connected layers: These layers are fully connected to the outputs of the previous layers. These types of layers are generally used in the last stages of CNNs.

2.4. CNN architecture

As shown in Fig. 1(b), our CNN architecture consists of three convolutional layers, two max-pooling layers, and two fully connected layers. Layer CONV1 comprises 15 filters; thus, its output comprises 15 filtered input signals. The filtered signals are then subsampled in layer MP1. These processes are repeated for the CONV2-CONV3-MP2 layers, and 30 output signals are obtained. These signals are then connected to the 50 units of the fully connected layer. Finally, this layer connects to a second fully connected layer, which classifies binary outputs (AH, N).

We conducted several experiments to optimize hyperparameters over predetermined ranges. We set the candidate hyperparameters as follows, and then searched for the optimal combination of parameters that produced the best results in the validation set: number of CNN layers {1, 2, 3}; number of convolution filters {5, 15, 30}; kernel size for convolution {4, 8, 16, 32}; stride for convolution {1, 2, 4, 8, 16}; and stride for pooling {1, 2}. The pooling kernel size was set at two, and the number of units in the fully connected layer was set to 50. To prevent overfitting, dropout layers [37] were added after every max-pooling and fully connected layer. The probability of the dropout layers was set to 0.2. To increase learning speed, a batch normalization process [38] was performed before applying the activation function. A rectified linear unit (ReLU) was used as an activation function, and binary cross-entropy was used as the loss function. We applied the He normal initializer to the kernel [39] and trained the model using Adam as a gradient descent optimizer, with the default parameters provided in the original paper [40]. The learning rate, batch size, and epoch were set to 0.001, 100, and 100, respectively.

All models were trained until overfitting began to occur, and an early stopping method was applied to the validation set. If validation loss did not improve for more than 50 epochs, training was halted, and the weights were reverted to the values that had achieved the best performance in the validation set. To test for classification repeatability and reduce the influence of weight initialization, training processes were repeated five times for each combination of hyperparameters. From the five trials, we selected the hyperparameter combination that produced the highest average kappa value in the validation set. Fig. 1(b) shows our final CNN architecture, and Table 2 presents our model configuration.

For model implementations, the Keras (version 2.0.2) library [41] was used with TensorFlow (version 1.0.1) as the backend [42]. Experiments were carried out on a workstation with a 3.4 GHz Intel i7-6700 CPU, 16 GB RAM, and an NVIDIA GeForce GTX1080 8 GB GPU.

2.5. Event detector and performance evaluation

After optimizing the hyperparameters, we evaluated the AH detection performance of the final model by using the test set. Test set segments were applied to the CNN model, and the segments were classified. These classified events were then fed to the event detector to identify valid AH events. An event was considered to be valid if at least five consecutive segments were classified as AH. We labeled the segment as class AH if at least 80% of a segment occurred within an apnea or hypopnea event period. For example, if a real AH event occurred in 20–30 s, five consecutive segments (18–28 s, 19–29 s, 20–30 s, 21–31 s, and 22–32 s) would be labeled as class AH because more than 80% of a segment contained an event. Otherwise, the segments were marked as class N. After processing was performed according to this rule, the number of AH events was obtained by considering the beginning-to-end of consecutive AH segments to be one event.

To evaluate the event detection performance, we computed Cohen's kappa coefficient value (KAPPA), a robust statistical measure of inter-rater agreement. In addition, the sensitivity (SENS), specificity (SPEC), accuracy (ACC), positive predictive value (PPV), and negative predictive value (NPV) were calculated by comparing the CNN model results to the reference PSG results, according to a segment-by-segment analysis. AHI was obtained using the number of valid events from the event detector. The Pearson's correlation coefficient between the estimated AHI and reference PSG AHI was calculated, and Bland-Altman analysis was performed. Further, SAHS diagnostic performance for AHI cutoffs $\geqq$ 5, 15, and 30 events/h was evaluated with SENS values, SPEC values, PPV values, NPV values, ACCs, F1-scores, and KAPPAs.

3. Results

This section describes the AH event detection performance of the CNN model for the test dataset. In this study, we used two methods to evaluate event detection performance: per-segment (segment-by-segment) analysis and per-recording analysis.

Table 3
Confusion matrix across all test dataset segments.

		Reference		KAPPA	SENS (%)	SPEC (%)	ACC (%)	PPV (%)	NPV (%)
		AH	N						
Estimated	AH	270372	40518	0.82	81.1	98.5	96.6	87.0	97.7
	N	63160	2633954						

3.1. Segment-by-segment analysis

The test set consisted of segments for 104 subjects, extracted as 10 s windows shifted 1 s at a time from the overall recording time, and 3008004 extracted test segments were applied to the model. Table 3 shows the performance of the proposed model over the test set by segment analysis. The AH events detected by the CNN model were compared with the scoring results from the reference PSG. For the 3008004 test segments, we obtained a KAPPA of 0.82, SENS of 81.1%, SPEC of 98.5%, ACC of 96.6%, PPV of 87.0%, and NPV of 97.7%. Furthermore, we computed the AH event detection performance for each group. Table 4 summarizes the statistical results for the SAHS severity group. KAPPA value gradually increased from the Non-SAHS group (AHI $\leq$5 events/h) to the severe SAHS group (AHI $\geq$ 30 events/h). From the non-SAHS to the moderate SAHS group, the KAPPA holds that k$\in$[0.6, 0.8], which is regarded as substantial agreement. For the severe SAHS group, the KAPPA showed almost perfect agreement (k$\in$[0.81, 1.00]) [43].

Fig. 2 displays the AH event estimation results for the best non-SAHS group case (see Fig. 2(a)) and the best severe SAHS group case (see Fig. 2(b)). For the best non-SAHS group case, which had an AHI reference value of 2.4, the KAPPA value was 0.78, SENS was 84.3%, SPEC was 99.8%, and ACC was 99.7%. In the severe SAHS group, the corresponding values were 44.2, 0.91, 97.3%, 94.8%, and 95.7%, respectively.

3.2. AHI estimation analysis

44. After classifying AH events in each segment, we analyzed AHI estimation performance. The estimated AHI (AHI_{Estim}) value was calculated from the ratio of the number of AH events (extracted by the event detector) to total sleep time. Fig. 3(a) shows a scatter plot of the AHI_{Estim} and reference AHI (AHI_{Refer}) values determined from PSG. The solid regression fitting line shows a significant correlation (Pearson's correlation coefficient = 0.99, $p < 0.001$) between AHI_{Estim} and AHI_{Refer}. In addition, the absolute error between the two AHI values was 3.1 $\pm$ 2.9 events/h (mean $\pm$ SD). A Bland-Altman plot was used to analyze the agreement between the two AHIs, as shown in Fig. 3(b). This figure represents the difference between the AHI_{Refer} and AHI_{Estim} values against the averages of the two AHI values. This figure shows the mean difference of the two AHI values as -2.42 events/h (Fig. 3(b), solid line), with a 95% confidential interval ranging from -9.42 to 4.59 events/h (Fig. 3(b), dashed line). Table 5 summarizes the SAHS severity classification and diagnostic performance for the test set. SAHS diagnostic performance was evaluated for AHI thresholds of 5, 15, and 30 events/h; the mean values for SENS, SPEC, PPV, NPV, ACC, F1-score, KAPPA and ROC-AUC were 98.1%, 89.1%, 90.8%, 98.8%, 94.9%, 0.94, 0.88 and 0.99, respectively.

Table 4
Event detection performance in the SAHS groups.

Group	N	KAPPA	SENS (%)	SPEC (%)	ACC (%)
Non-SAHS	26	0.69	69.8	99.7	99.4
Mild SAHS	26	0.71	67.2	99.3	98.2
Moderate SAHS	26	0.78	76.0	98.4	96.3
Severe SAHS	26	0.82	84.5	95.9	92.5

4. Discussion

In this paper, we proposed an optimal CNN model to detect AH events precisely during sleep in real time. A single channel (nasal pressure signal) was used, and real-time AH events were detected using an overlapping window method.

Real-time AH events were automatically detected using the following procedure: 1) normalizing data using an adaptive normalization method, 2) segmenting data by 10 s windows using 1 s shifts, 3) applying the CNNs model and classifying events, and 4) applying an event detector. Normal respiration signals can be affected by a subject's body position. When a subject moves from their back onto their left or right side, the amplitude of the respiration signal decreases, causing a hypopnea-like event to occur in the signal, despite the subject's normal breathing (see gray box in Fig. 4). When the signal is normalized by a conventional z-scoring method and then input into the CNN model, such periods are misclassified as AH class. In contrast, the adaptive normalization method amplifies the normal breathing period affected by changes in body position, while maintaining apnea and hypopnea event patterns (see Fig. 4(d)). In a segment-by-segment analysis of the test set, we found a KAPPA value of 0.71 for the z-score normalization method, and 0.82 for the adaptive normalization method.

In the segment-by-segment analysis, the KAPPA value was 0.82, which is regarded as almost perfect agreement. A segment-by-segment comparison between the proposed CNN model and PSG scoring results is presented as a confusion matrix in Table 3. For approximately 49.2% of the false positives (FPs), the amplitude of the nasal pressure signal decreased by more than 30% from the pre-event baseline, which is one of the criteria for determining hypopnea according to the AASM guidelines. However, there was no additional information for identifying hypopneas, such as the arousal or desaturation of SpO_2. Therefore, these segments may be misclassified as AH segments. Other FPs appeared before or after true positive segments, meaning that the event was detected to be longer than the reference event. Because of these FPs segments, AH events were overestimated relative to the reference. Table 6 compares the number of AH events and AHI values for the CNN and PSG methods. As already mentioned, because there was no additional information, such as the arousal or desaturation of SpO_2, AHI and the number of AH events were overestimated. Although the mean absolute error between AHI_{Estim} and AHI_{Refer} was 3.1 event/h, the SAHS diagnostic performance exhibited a high value, as shown in Table 5. The average KAPPA value of diagnostic performance was 0.88, which means that our proposed method can be used to screen for SAHS severity. Table 7 shows the SAHS severity diagnostic performance found in previous studies. As you can see our study achieved better performance than previous studies.

In the segment analysis for all SAHS severity groups (see Table 4), the KAPPA values exceeded 0.6, indicating substantial agreement. Performance gradually increased from the Non-SAHS to severe SAHS groups. The reference hypopnea event percent, which is the number of hypopnea events divided by apnea-hypopnea events, were 85.0%, 85.7%, 71.4%, and 57.0% for each group. The model showed relatively low performance with the Non-SAHA and mild SAHS group because of its higher hypopnea event ratio.

In the per-recording analysis, the AHI calculated by our proposed model showed a significant correlation with the reference AHI. Using a

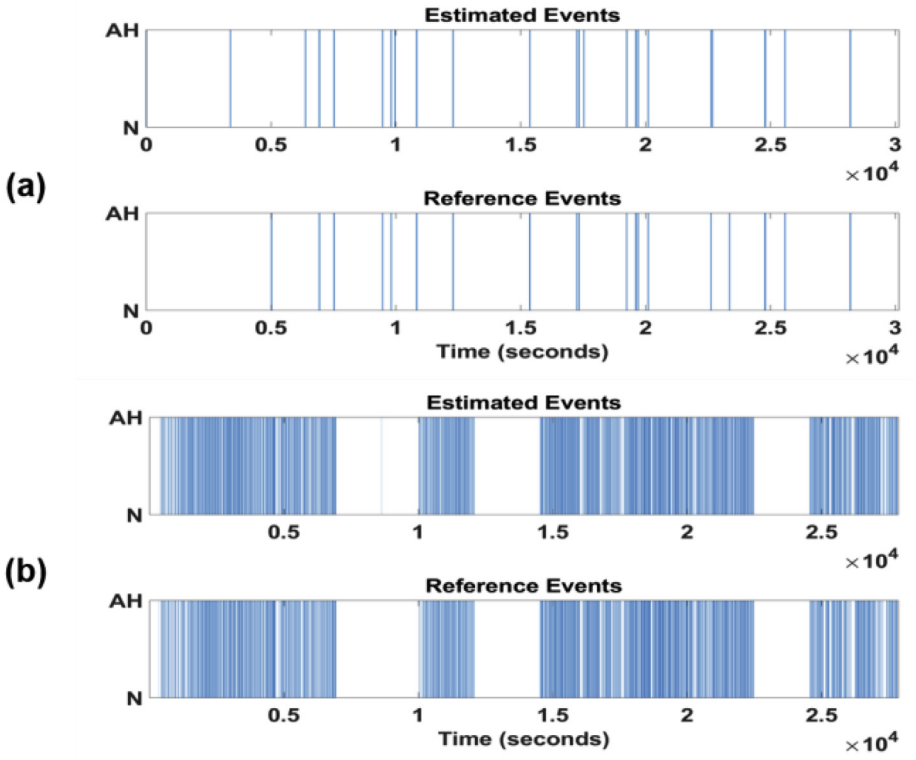

Fig. 2. (A) Best estimation result in non-SAHS group. (b) Best estimation result in severe SAHS group. (Each bar in the plots indicates an AH segment.)

Bland-Altman analysis, we found a mean difference between AHI_{Refer} and AHI_{Estim} of −2.42 events/h, meaning that the proposed model overestimated AHI. Events were overestimated because additional information (such as SpO_2 desaturation or arousal) was not available for more accurately detecting hypopnea. By using such information, we may be able to improve AH event detection performance in future studies.

We also checked the time required for classifying test segments for real-time applications. From the trained CNN model, it took 43 s to classify 3008004 test segments. The classification time for one segment is approximately 0.000014 s, which is less than 1 s. This implies that there is no problem in real-time events identification.

Unlike previous studies [10,11,13,16,17,23–26], we did not use hand-engineered features. The CNN model is fully automatic; thus, no additional feature extraction and selection processes are required. Another important aspect of the proposed method is that AH events can be detected in real time. Most previous works estimated events from non-overlapped epochs/episodes [13,21,22]. The number of events detected by a non-overlapping method may not be accurate, because a single episode can contain multiple events, and a long event can occur across multiple episodes. In our PSG recordings, when AH events are scored by minute, the mean values for AH events number and AHI are 138.3 and 21.0 events/h, respectively. The original mean values for AH events number and AHI were 158.6 and 24.1 events/h, respectively. These values indicate that events are underestimated if events are scored by minute, which can lead to an inaccurate diagnosis of SAHS severity. Therefore, we identified actual AH events using a sliding window of 10 s, shifted by 1 s at a time. With this approach AH events can be

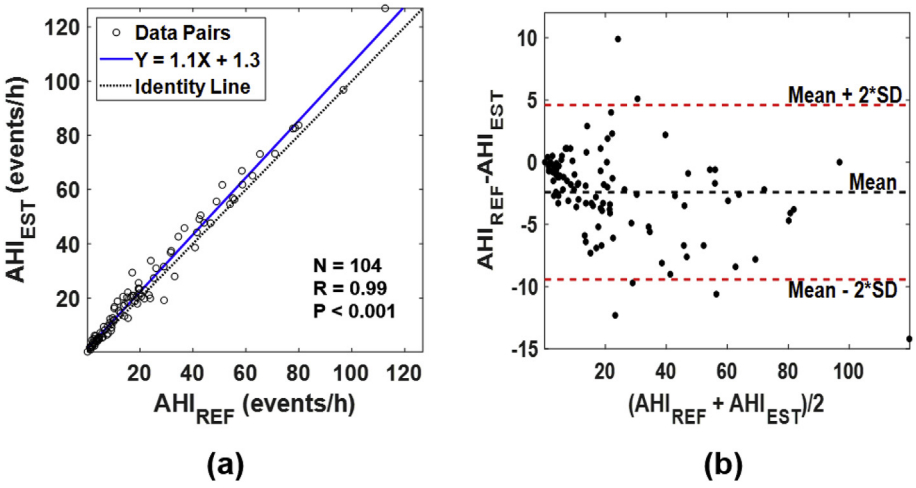

Fig. 3. (A) Scatter plot of the correlation between AHI_{Estim} and AHI_{Refer}. (b) Bland-Altman plot of AHI_{Estim} and AHI_{Refer}.

Table 5
SAHS severity classification and diagnostic performance for the test set.
ROC-AUC, area under the receiver operating characteristics curve.

		Estimated SAHS severity					AHI cutoff (events/h)			
		Non-SAHS	Mild	Moderate	Severe		≥5	≥15	≥30	Average
SAHS severity determined from PSG	Non-SAHS	22	4	0	0	SENS (%)	100.0	98.1	96.2	98.1
						SPEC (%)	84.6	86.5	96.2	89.1
	Mild	0	19	7	0	PPV (%)	95.1	87.9	89.3	90.8
						NPV (%)	100.0	97.8	98.7	98.8
	Moderate	0	1	22	3	ACC (%)	96.2	92.3	96.2	94.9
						F1-score	0.98	0.93	0.93	0.94
	Severe	0	0	1	25	KAPPA	0.89	0.85	0.90	0.88
						ROC-AUC	0.99	0.99	1.00	0.99

detected more precisely than with the non-overlapped minute method.

This study had some limitations. First, we did not differentiate between the three types of sleep apnea. The respiration change characteristics of the nasal pressure signal were not significantly different for obstructive, central, and mixed sleep apnea. Therefore, it was difficult for us to categorize the three types of sleep apnea. To differentiate the three types of apnea, we need to collect three types of apnea data and develop a new model in future studies. Second, measuring nasal pressure signal using cannula is uncomfortable for the subjects. To reduce the discomfort, polyvinylidene (PVDF) sensor signals or abdominal/thoracic belt signals could be used as alternative signals. We will perform event detection research by applying the CNN model to these signals in future studies. Third, the nasal pressure signal is weak to

Table 6
Comparison of number of AH events and AHI values from the proposed model and PSG.

Group	No. of AH events		AHI (events/h)	
	Estimated	Reference	Estimated	Reference
Non-SAHS	20.8 ± 11.8	17.8 ± 9.0	3.5 ± 1.6	2.7 ± 1.2
Mild SAHS	68.9 ± 26.8	65.5 ± 20.0	11.7 ± 4.3	9.5 ± 2.7
Moderate SAHS	149.5 ± 40.9	139.6 ± 32.7	22.6 ± 4.7	20.3 ± 3.9
Severe SAHS	374.3 ± 126.5	359.7 ± 121.3	60.2 ± 21.2	55.9 ± 20.0

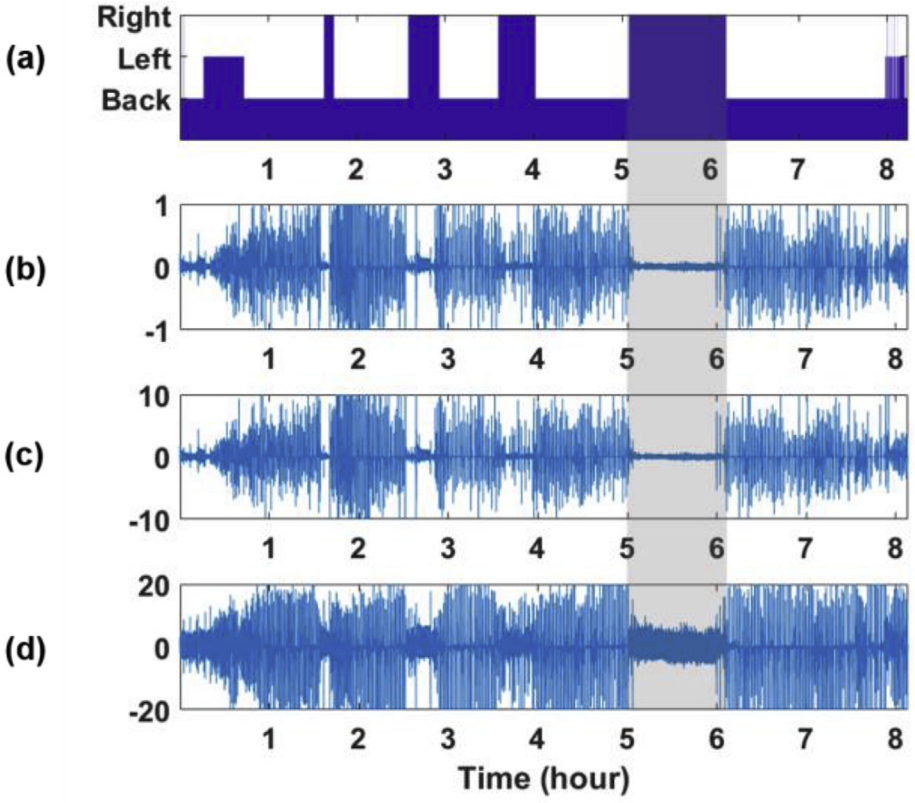

Fig. 4. Comparison of normalized methods effect on raw signal. (a) Body position. (b) Raw signal recorded from nasal pressure sensor. (c) Normalized signal using z-score method. (d) Normalized signal using adaptive normalization method.

Table 7
SAHS severity diagnostic performance comparisons with existing studies.

Research work	Signal	AHI cutoff	SENS (%)	SPEC (%)	ACC (%)
H. Nakano et al. [16]	Nasal pressure	5	97.0	77.0	–
		15	97.0	73.0	–
Nigro et al. [45]	Nasal pressure	5	89.3	60.0	–
		15	76.7	83.0	–
		30	88.5	95.3	–
Gutierrez-Tobal et al. [46]	Oronasal	5	87.1	80.0	86.5
	thermal and	15	85.9	72.9	81.0
	nasal pressure	30	74.2	90.6	82.5
Our study	Nasal pressure	5	**100.0**	**84.6**	**96.2**
		15	**98.1**	**86.5**	**92.3**
		30	**96.2**	**96.2**	**96.2**

detect mouth breathing, which may decrease our event detection performance. To verify this issue, evaluation of the proposed algorithm will be performed using oronasal pressure signals in a future study. Finally, AH event detection was not tested in an online environment. We evaluated detection performance offline. To confirm usability, our model needs to be validated at sleep laboratories and homes in future studies.

5. Conclusion

We developed a real-time AH event detection method using CNNs and detected events based on overlapping segments of a nasal pressure signal, allowing us to identify actual events in real time. We also implemented a CNN model that did not use hand-engineered features. In this model, a feature extraction and selection process is not required. Our proposed method could detect AH events more precisely and achieved comparable performance to those of previous studies for assessing SAHS severity. Because our model automatically extracts features from the signal, the inclusion of additional SpO_2 or EEG data could further improve performance. Our approach can be used to reduce AH events scoring time in sleep laboratories, and it can be applied to screen SAHS severity before PSG tests. In addition, long-term home-based SAHS monitoring, which is useful for follow up studies on sleep apnea-hypopnea, could be conducted using our proposed method.

Conflicts-of-interest statement

None Declared.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science, ICT & Future Planning) (No. 2017R1A2B2004061).

References

[1] W.W. Flemons, et al., Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research, Sleep 22 (5) (1999) 667–689.
[2] R.D. Chervin, Sleepiness, fatigue, tired, and lack of energy in obstructive sleep apnea, Chest 118 (2000) 372–379.
[3] W.W. Flemons, et al., Sleep apnea and cardiac arrhythmias. Is there a relationship? Am. Rev. Respir. Dis. 148 (1993) 618–621.
[4] Y. Peker, et al., An independent association between obstructive sleep apnoea and coronary artery disease, Eur. Respir. J. 14 (1) (1999) 179–184.
[5] H.K. Yaggi, et al., Obstructive sleep apnea as a risk factor for stroke and death, N. Engl. J. Med. 353 (19) (2005) 2034–2041.
[6] D. Brooks, et al., Obstructive sleep-apnea as a cause of systemic hypertension: evidence from a canine model, J. Clin. Invest. 99 (1) (1997) 106–109.
[7] P.E. Peppard, et al., Longitudinal association of sleep-related breathing disorder and depression, Arch. Intern. Med 166 (16) (2006) 1709–1715.
[8] M.S.M. Ip, et al., Obstructive sleep apnea is independently associated with insulin resistance, Am. J. Respir. Crit. Care Med. 165 (2002) 670–676.
[9] R.B. Berry, et al., The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, Version 2.0, Am. Acad. Sleep Med (2012).
[10] A.H. Khandoker, et al., Automated scoring of obstructive sleep apnea and hypopnea events using short-term electrocardiogram recordings, IEEE Trans. Inf. Technol. Biomed 13 (6) (2009) 1057–1067.
[11] A.H. Khandoker, et al., Support vector machines for automated recognition of obstructive sleep apnea syndrome from electrocardiogram recordings, IEEE Trans. Inf. Technol. Biomed. 13 (1) (2008) 1–32.
[12] M.O. Mendez, et al., Sleep apnea screening by autoregressive models from a single ECG lead, IEEE Trans. Biomed. Eng 56 (12) (2009) 2838–2850.
[13] M. Bsoul, et al., Apnea MedAssist: real-time sleep apnea monitor using single-lead ECG, IEEE Trans. Inf. Technol. Biomed 15 (3) (2011) 416–427.
[14] T. Penzel, et al., Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings, Med. Biol. Eng. Comput 40 (4) (2002) 402–407.
[15] P. De Chazal, et al., Automated processing of the single-lead electrocardiogram for the detection of obstructive sleep apnoea, IEEE Trans. Biomed. Eng. 50 (6) (2003) 686–696.
[16] H. Nakano, et al., Automatic detection of sleep-disordered breathing from a single-channel airflow record, Eur. Respir. J 29 (4) (2007) 728–736.
[17] J. Han, et al., Detection of apneic events from single channel nasal airflow using 2nd derivative method, Comput. Meth. Progr. Biomed. 91 (3) (2008) 199–207.
[18] R. Ragette, et al., Diagnostic performance of single airflow channel recording (ApneaLink) in home diagnosis of sleep apnea, Sleep Breath. 14 (2) (2010) 109–114.
[19] S.I. Rathnayake, et al., Nonlinear features for single-channel diagnosis of sleep-disordered breathing diseases, IEEE Trans. Biomed. Eng. 57 (8) (2010) 1973–1981.
[20] D. Alvarez, et al., Multivariate analysis of blood oxygen saturation recordings in obstructive sleep apnea diagnosis, IEEE Trans. Biomed. Eng. 57 (12) (2010) 2816–2824.
[21] A. Burgos, et al., Real-time detection of apneas on a PDA, IEEE Trans. Inf. Technol. Biomed. 14 (4) (2010) 995–1002.
[22] B. Xie, H. Minn, Real-time sleep apnea detection by classifier combination, IEEE Trans. Inf. Technol. Biomed 16 (3) (2012) 469–477.
[23] S. Reisch, et al., Detection of sleep apnea with the forced oscillation technique compared to three standard polysomnographic signals, Respiration 67 (5) (2000) 518–525.
[24] J.V. Marcos, et al., Assessment of four statistical pattern recognition techniques to assist in obstructive sleep apnoea diagnosis from nocturnal oximetry, Med. Eng. Phys. 31 (8) (2009) 971–978.
[25] H.M. Al-Angari, A.V. Sahakian, Automated recognition of obstructive sleep apnea syndrome using support vector machine classifier, IEEE Trans. Inf. Technol. Biomed 16 (3) (Apr. 2012) 463–468.
[26] P. Várady, et al., A novel method for the detection of apnea and hypopnea events in respiration signals, IEEE Trans. Biomed. Eng. 49 (9) (2002) 936–942.
[27] Y. LeCun, et al., Deep learning, Nature 521 (7553) (2015) 436–444.
[28] A. Krizhevsky, et al., ImageNet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst (2012) 1–9.
[29] J.J. Tompson, et al., Joint training of a convolutional network and a graphical model for human pose estimation, Adv. Neural Inf. Process. Syst. (2014) 1799–1807.
[30] Y. Taigman, et al., Deepface: closing the gap to human-level performance in face verification, CVPR IEEE Conf (2014) 1701–1708.

[31] Y. Ren, Y. Wu, Convolutional deep belief networks for feature extraction of EEG signal, Int. Jt. Conf. Neural Network. (2014) 2850–2853 2014.
[32] X. Zhu, et al., EOG-based drowsiness detection using convolutional neural networks," 2014 Int, Jt. Conf. Neural Networks (2014) 128–134.
[33] W. Yin, et al., ECG monitoring system integrated with IR-UWB radar based on CNN, IEEE Access 4 (2016) 6344–6351.
[34] S. Kiranyaz, et al., Real-time patient-specific ECG classification by 1-D convolutional neural networks, IEEE Trans. Biomed. Eng 63 (3) (2016) 664–675.
[35] J. Tian, J. Liu, Apnea detection based on time delay neural network, Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 3 (2005) 2571–2574.
[36] P. Hensman, D. Masko, The impact of imbalanced training data for convolutional neural networks, Degree Project in Computer Science, KTH Royal Institute of Technology (2015).
[37] N. Srivastava, et al., Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929–1958.
[38] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, International Conference on Machine Learning, 2015.
[39] K. He, et al., Delving deep into rectifiers: surpassing human-level performance on imagenet classification, Proc. IEEE Int. Conf. Comput. Vis (2015) 1024–1034.
[40] D.P. Kingma, J. Ba, Adam: a Method for Stochastic Optimization, (2014), pp. 1–15 arXiv Prepr. arXiv1412.6980.
[41] F. Chollet, Keras, GitHub (2015) https://github.com/fchollet/keras.
[42] M. Abadi, et al., TensorFlow: a System for Large-scale Machine Learning TensorFlow: a System for Large-scale Machine Learning, (2016), pp. 265–284.
[43] J.R. Landis, G.G. Koch, Measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.
[44] R. Haidar, et al., Sleep apnea event detection from nasal airflow using convolutional neural networks, Int. Conf. on Neural Information Processing (2017) 819–827.
[45] C.A. Nigro, et al., Comparison of the automatic analysis versus the manual scoring from ApneaLink™ device for the diagnosis of obstructive sleep apnoea syndrome, Sleep Breath. 15 (2011) 679–686.
[46] G.C. Gutiérrez-Tobal, et al., Utility of AdaBoost to detect sleep apnea-hypopnea syndrome from single-channel airflow, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 63 (2016) 636–646.
[47] D.A. Dean, Scaling up scientific discovery in sleep medicine: the national sleep research resource, Sleep 39 (5) (2016) 1151–1164.
[48] G.Q. Zhang, The national sleep research resource: towards a sleep data commons, J. Am. Med. Inf. Assoc. (2018).