Computers in Biology and Medicine 147 (2022) 105784

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

A 2D convolutional neural network to detect sleep apnea in children using airflow and oximetry

Jorge Jiménez-García [a,b,*], María García [a,b], Gonzalo C. Gutiérrez-Tobal [a,b], Leila Kheirandish-Gozal [c], Fernando Vaquerizo-Villar [a,b], Daniel Álvarez [a,b,d], Félix del Campo [a,b,d], David Gozal [c,e], Roberto Hornero [a,b]

[a] Biomedical Engineering Group, University of Valladolid, Valladolid, Spain
[b] CIBER-BBN, Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina, Instituto de Salud Carlos III, Valladolid, Spain
[c] Department of Child Health and Child Health Research Institute, The University of Missouri School of Medicine, Columbia, MO, USA
[d] Sleep-Ventilation Unit, Pneumology Department, Río Hortega University Hospital, Valladolid, Spain
[e] Department of Medical Pharmacology and Physiology, The University of Missouri School of Medicine, Columbia, MO, USA

ARTICLE INFO

Keywords:
Obstructive sleep apnea
Children
Airflow
Oximetry
Convolutional neural network
Deep learning

ABSTRACT

The gold standard approach to diagnose obstructive sleep apnea (OSA) in children is overnight in-lab polysomnography (PSG), which is labor-intensive for clinicians and onerous to healthcare systems and families. Simplification of PSG should enhance availability and comfort, and reduce complexity and waitlists. Airflow (AF) and oximetry (SpO_2) signals summarize most of the information needed to detect apneas and hypopneas, but automatic analysis of these signals using deep-learning algorithms has not been extensively investigated in the pediatric context. The aim of this study was to evaluate a convolutional neural network (CNN) architecture based on these two signals to estimate the severity of pediatric OSA. PSG-derived AF and SpO_2 signals from the Childhood Adenotonsillectomy Trial (CHAT) database (1638 recordings), as well as from a clinical database (974 recordings), were analyzed. A 2D CNN fed with AF and SpO_2 signals was implemented to estimate the number of apneic events, and the total apnea-hypopnea index (AHI) was estimated. A training-validation-test strategy was used to train the CNN, adjust the hyperparameters, and assess the diagnostic ability of the algorithm, respectively. Classification into four OSA severity levels (no OSA, mild, moderate, or severe) reached 4-class accuracy and Cohen's Kappa of 72.55% and 0.6011 in the CHAT test set, and 61.79% and 0.4469 in the clinical dataset, respectively. Binary classification accuracy using AHI cutoffs 1, 5 and 10 events/h ranged between 84.64% and 94.44% in CHAT, and 84.10%–90.26% in the clinical database. The proposed CNN-based architecture achieved high diagnostic ability in two independent databases, outperforming previous approaches that employed SpO_2 signals alone, or other classical feature-engineering approaches. Therefore, analysis of AF and SpO_2 signals using deep learning can be useful to deploy reliable computer-aided diagnostic tools for childhood OSA.

1. Introduction

Obstructive Sleep Apnea (OSA) syndrome is a common sleep disorder in which the increased resistance within the upper airway induces decreases and cessation of respiratory airflow during sleep [1,2]. When OSA affects children and remains untreated, several negative consequences arise, including neurocognitive and behavioral deficits, and cardiovascular and metabolic morbidities [2]. The prevalence of OSA ranges between approximately 1-5% in the pediatric population, and it is likely that many additional cases remain undiagnosed due to the lack of resources and extensive waitlists in sleep laboratories [1]. The standard approach to diagnose OSA consist in overnight in-lab polysomnography (PSG) testing, a complex multichannel recording during which sleep stages are characterized, and cardiorespiratory and other physiological parameters are concurrently measured [1,3]. These signals are subsequently analyzed, and the number of respiratory flow interruptions (apnea) and airflow reductions (hypopnea) during sleep is computed to provide an estimate of the disease severity if such is present

* Corresponding author. Biomedical Engineering Group, E.T.S. Ingenieros de Telecomunicación, Universidad de Valladolid, Campus Miguel Delibes, Paseo Belén 15, 47011, Valladolid, Spain.
E-mail address: jorge.jimenez@gib.tel.uva.es (J. Jiménez-García).

https://doi.org/10.1016/j.compbiomed.2022.105784
Received 12 January 2022; Received in revised form 19 April 2022; Accepted 26 June 2022
Available online 28 June 2022

[3]. When evaluating children, the American Academy of Sleep Medicine (AASM) guidelines define apneas as a decrease ≥90% in the airflow (AF) signal during at least two respiratory cycles, while hypopneas are defined as a decrease ≥30% in AF accompanied by a reduction of at least 3% in the blood oxygen saturation (SpO_2) signal or an electroencephalographic arousal [3]. The severity of OSA is primarily determined by the Apnea Hypopnea Index (AHI), which is the rate of apnea and hypopnea events per hour of sleep (e/h).

Despite the fact that the PSG is the preferred and gold standard diagnostic method in children with symptoms suggestive of OSA, simpler alternatives like nocturnal oximetry and respiratory polygraphy have been proposed to simplify the assessments and to reduce costs [4–6]. These alternatives mainly focus on AF and/or SpO_2 signals, which provide the most relevant information to detect respiratory events and the associated oxyhemoglobin desaturations [7]. In addition, the analysis of cardiorespiratory signals to automatically detect pediatric OSA using feature-engineering approaches has gained relevance and popularity in the last several years [7,8]. Most of these approaches have assessed AF, SpO_2, electrocardiogram (ECG), pulse rate variability (PRV), and heart rate variability (HRV) signals among others, and applied temporal, spectral, and nonlinear methods to objectively quantify and characterize them while deriving their unique and distinguishable characteristics in the context of OSA. Typically, these analyses were then followed by application of a variety of machine-learning (ML) algorithms to enable detection of OSA severity [8]. Classical ML algorithms encompassed logistic regression [9–13], ensemble learning [10,14], and neural networks [15–19]. However, the latest approaches such as deep learning (DL) have been scarcely investigated in the context of childhood OSA despite their successful application in the adult population [8,20,21]. Feature-engineering approaches require an exhaustive characterization of the signals to obtain relevant features, whereas DL methods automatically extract the relevant contextual information from the signals to detect OSA [21]. DL applications in adult OSA primarily focused on the analysis of ECG, SpO_2, and AF signals [21]. Convolutional neural networks (CNN), recurrent neural networks (RNN), and a combination of CNN and RNN have been primarily used to detect OSA using the ECG signal [22–24]. In addition, spectrograms and scalograms obtained from the ECG were used to evaluate bidimensional (2D) CNN models [22,25]. Other approaches encompassed the analysis of HRV and ECG-derived respiration using deep architectures [26]. Regarding respiration signals, CNN and RNN-based approaches have been also investigated using oronasal AF or thoracic/abdominal sensors [27–32]. The SpO_2 signal has been employed to detect OSA by means of deep neural networks (DNN) [33] and CNNs [34–36]. In the same regard, other studies combined AF and SpO_2 signals using these DL algorithms [37,38]. These DL-based methodologies have been developed and tested for the adult population and may not be suited for the particularities of pediatric OSA. The main differences between adult and pediatric OSA are related to the symptoms, negative consequences, and the PSG findings, where both the scoring of apneas and hypopneas and the criteria to assess OSA severity are more restrictive for children [2,3]. Among the numerous DL algorithms that have been developed, the use of CNNs has emerged as a relevant and useful DL technique in image and signal analysis, including the medical field [21,39,40]. Despite CNN success in biomedical applications, only one recent study proposed a CNN architecture for detection of pediatric OSA using SpO_2 signals to estimate AHI [20]. AF and SpO_2 signals have not been jointly evaluated using DL approaches despite the complementarity of these signals as shown in previous studies using feature-engineering approaches [14,18,19]. Here, we propose the joint use of AF and SpO_2 signals to detect pediatric OSA using a CNN-based approach. The main novelty of this study relies on the combination of AF and SpO_2 data to derive a CNN model aimed at estimating pediatric OSA severity.

We hypothesized that a CNN can take advantage of the complementarity of the information from AF and SpO_2 signals to estimate pediatric OSA severity. That information can then be useful to improve the automated identification of apneas and hypopneas with a reduced subset of signals. Accordingly, the objective of this study was to evaluate a trained CNN model with these two overnight signals to estimate the AHI. The major contributions of our proposal are (i) the combination of AF and SpO_2 signals to evaluate a DL algorithm aimed at detecting OSA in children, and (ii) the use of a 2D CNN approach to automatically estimate the total sleep AHI from these two signals jointly.

2. Subjects and signals

Two databases, one public and one private, were used in this study. The first one, the Childhood Adenotonsillectomy Trial (CHAT) database, comprises 1638 PSG recordings from 1232 subjects recruited between 2007 and 2012 (Number of Clinical Trial: NCT00560859) [41,42]. This database is publicly available through the National Sleep Research Resource website (https://sleepdata.org/datasets/chat) under request. The sleep studies were divided into three subgroups: baseline (453 subjects), follow-up (406 subjects) and nonrandomized (779 subjects). Sleep studies from each of the 3 subgroups were randomly split into training (60%), validation (20%), and test (20%) sets (Table 1). The division conducted was subject-wise, so that no subject was represented in two sets simultaneously. The sets assigned to the subjects in the baseline group were transferred to the follow-up subgroup, so each follow-up subject has both baseline and follow-up recordings in the same training/validation/test set. No statistically significant differences were found in age, sex, body mass index (BMI), and AHI between the sets ($p \geq 0.01$, Mann-Whitney U test for numeric variables, Chi-square test for sex). PSG data from CHAT database included annotations of the beginning and duration of apneic events according to the AASM guidelines in Ref. [43], which is needed to train the proposed DL algorithm.

A second private database from the University of Chicago Medicine (UofC) includes PSG data from 974 pediatric subjects referred for polysomnographic evaluation between 2012 and 2014. The legal caretakers were informed and gave their consent, and the Ethics Committee of the UofC approved the study protocol (see Ethical Approval section). The PSGs of the subjects in the UofC database were evaluated according to the current AASM guidelines [3]. Since this database did not contain annotations of individual time stamp of respiratory events, no data from UofC were used during the training stage of the CNN. However, subjects in the UofC database were used to validate and test the AHI estimation algorithm, so each subject was labeled with the PSG-derived AHI. The 974 subjects in the UofC database were randomly split into the validation (60%) and test (40%) sets, with no statistically significant differences in age, sex, BMI and AHI among them ($p \geq 0.01$, Mann-Whitney U test for numeric variables, Chi-square test for sex). Table 1 summarizes the demographic and clinical data of the subjects involved in this study. Statistically significant differences were observed in age, sex, and AHI between CHAT and UofC sets, which were noted in Table 1. As these differences may affect the generalizability of the proposed model, a joint validation set was formed in order to minimize the chances of overfitting towards the CHAT training and validation data. The subjects from CHAT and UofC validation sets were merged to form a larger, heterogeneous, and generic validation set involving 910 subjects. This set was used to find the optimal hyperparameters of the CNN and the AHI estimation algorithms. Finally, the test sets were used to evaluate the diagnostic ability of the proposed algorithm.

AF signals were obtained from the PSG and were recorded using an oronasal thermistor with sampling frequencies (f_s) ranging 20–512 Hz, whereas SpO_2 signals were recorded with a pulse oximeter finger probe with f_s in the range 1–512 Hz. Fig. 1 shows an example of AF and SpO_2 signals with apnea/hypopnea events and their respective desaturations. As can be seen in the AF signal, the apneas and hypopneas produce reductions of the amplitude of AF, as well as sudden drops in the oxyhemoglobin saturation due to reduced or interrupted gas exchange.

Table 1
Demographic and clinical characteristics of the children in the CHAT and UofC databases.

	CHAT-Training	CHAT-Validation	UofC-Validation	CHAT-Test	UofC-Test
Subjects (*n*)	1006 (61.42%)	326 (19.90%)	584 (59.96%)	306 (18.68%)	390 (40.04%)
Age (years)	7 [6; 8] [a,b]	7 [6; 8] [c,d]	6 [3; 8] [a,c,e]	6.9 [6; 8] [e,f]	5.5 [3; 9] [b,d,f]
Females (*n*)	520 (51.7%)[a,b]	168 (51.5%)[d]	238 (40.8%)[a,e]	168 (54.9%)[e,f)]	137 (35.1%)[b,d,f]
Males (*n*)	471 (46.8%)[a,b]	156 (47.9%)[d]	346 (59.2%)[a,e]	134 (43.8%)[e,f]	253 (64.9%)[b,d,f]
BMI (kg/m^2)	17.4 [15.6; 21.7]	17.1 [15.4; 21.8]	17.7 [16.1; 22.7]	17.6 [15.7; 21.7]	18.2 [16.0; 21.9]
AHI (events/h)	2.6 [1.1; 5.9][a]	2.4 [1.2; 5.8][c]	4.1 [1.7; 10.0][a,c,e]	2.3 [1.1; 6.2][e]	3.3 [1.4; 7.9]
No OSA[g] (*n*)	219 (21.8%)	69 (21.2%)	96 (16.4%)	67 (21.9%)	75 (19.2%)
Mild OSA[h] (*n*)	496 (49.3%)	168 (51.5%)	229 (39.2%)	148 (48.4%)	169 (43.3%)
Moderate OSA[i] (*n*)	160 (15.9%)	44 (13.5%)	113 (19.4%)	49 (16.0%)	63 (16.2%)
Severe OSA[j] (*n*)	131 (13.0%)	45 (13.8%)	146 (25.0%)	42 (13.7%)	83 (21.3%)
Segments (*n*)	227,101	73,451	114,588	68,727	76,896

Data presented as median [interquartile range] or n (%).
BMI = body mass index, AHI = apnea-hypopnea index; OSA = Obstructive Sleep Apnea.
[a] Statistically significant differences ($p < 0.01$, Bonferroni correction) between CHAT-Training and UofC-Validation.
[b] Statistically significant differences ($p < 0.01$, Bonferroni correction) between CHAT-Training and UofC-Test.
[c] Statistically significant differences ($p < 0.01$, Bonferroni correction) between CHAT-Validation and UofC-Validation.
[d] Statistically significant differences ($p < 0.01$, Bonferroni correction) between CHAT-Validation and UofC-Test.
[e] Statistically significant differences ($p < 0.01$, Bonferroni correction) between CHAT-Test and UofC-Validation.
[f] Statistically significant differences ($p < 0.01$, Bonferroni correction) between CHAT-Test and UofC-Test.
[g] AHI<1 event/h.
[h] 1 ≤ AHI<5 events/h.
[i] 5 ≤ AHI<10 events/h.
[j] AHI≥10 events/h.

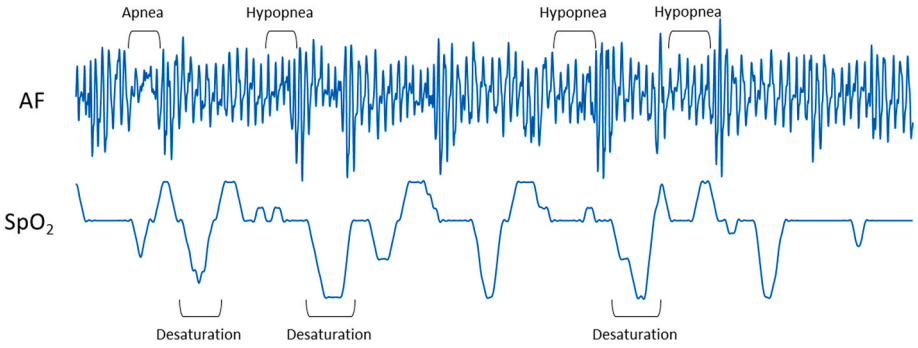


Fig. 1. Example of 5-min segments of airflow (AF) and oximetry (SpO_2) signals with apneas, hypopneas, and desaturations associated to apneic events.

3. Methods

In this study, AF and SpO_2 signals were analyzed by means of a CNN aimed at estimating the AHI. Fig. 2 summarizes the workflow of the methodology, including the methods that have been developed and applied to process the overnight AF and SpO_2. The signals were minimally preprocessed and were segmented into 5-min epochs before feeding the CNN. The AHI was estimated by counting the number of apneic events detected throughout the segments. In addition, the datasets employed in this study are also represented in Fig. 2. The CHAT-Training and CHAT-Validation datasets were used to train and monitor the convergence of the proposed CNN at segment level, while the joint validation set was used to derive the optimal hyperparameters of the CNN as well as to train and validate the AHI estimation algorithm at subject level. Finally, both test sets were used to evaluate the diagnostic performance of the proposed CNN-based algorithm.

3.1. Signal preprocessing and segmentation

The signals were preprocessed to normalize their characteristics before they were presented to the CNN. Both AF and SpO_2 were resampled to a common $f_s = 10$ Hz to ensure that the signals had the same length and to reduce the computational cost. Afterwards, the AF signal was low-pass filtered to reduce noise. This was achieved using a Kaiser window finite impulse response (FIR) filter. The application of a FIR filter ensures linear phase filtering, and the Kaiser window low-pass filter allowed us to establish a trade-off between the main-lobe width and the side-lobe amplitude, independently defining the transition bandwidth and the stopband attenuation. The design parameters of the filter were: cutoff frequency 1.5 Hz, stop frequency 2 Hz, and stopband attenuation 100 dB. The amplitude of AF was adaptively normalized as in previous studies [14,44]. Then, both signals were standardized prior to being presented to the CNN [36]. The signals were joined and segmented into 5-min epochs (300 s), so segments were shaped as 2D tensors of size 3000 × 2 (Fig. 2). The length of the segments was selected as a tradeoff between a period that encompasses consecutive apneic events and their respective desaturations and the number of available segments to train and validate the CNN. Data augmentation was performed in the training stage by using 50% overlapped segments, so the total number of training instances was increased. In addition, overlapped segments can enhance the detection of those apneic events truncated at the beginning or the end of a segment. The annotation of each segment is the number of apneic events that begin and end in the period that encompasses the segment, which was obtained from the PSG annotations in the CHAT database [20,33,36]. The number of extracted segments from each database is specified in Table 1.

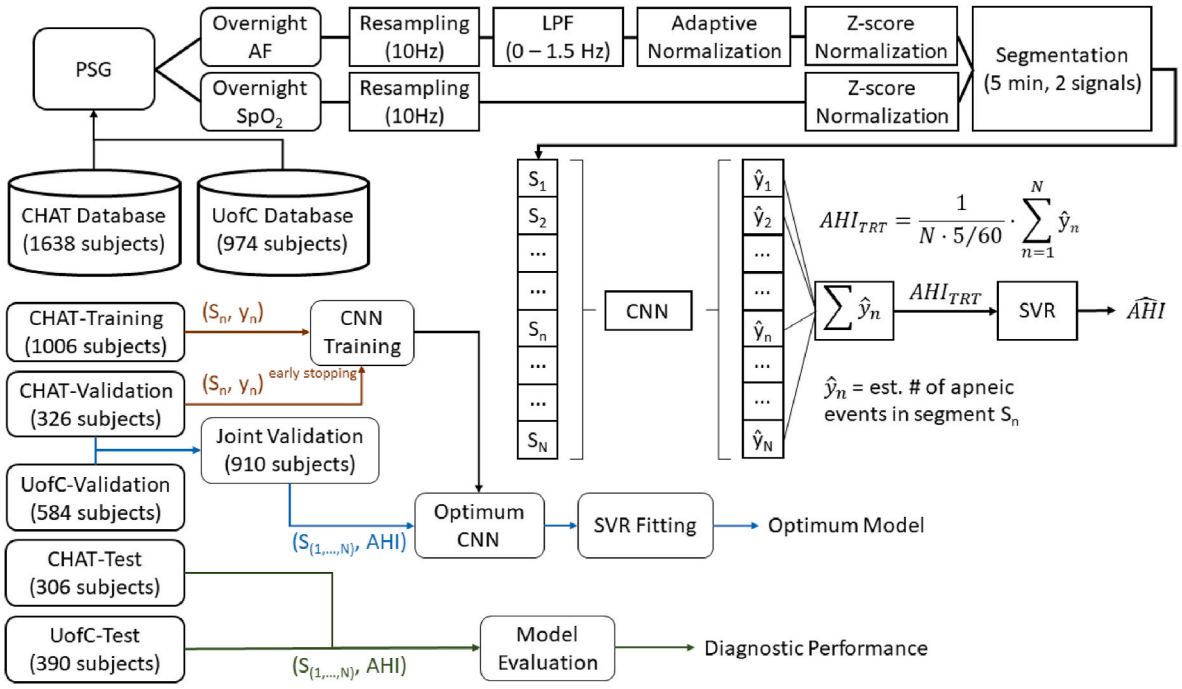


Fig. 2. Workflow of the methodology and databases employed in this study. Preprocessing of overnight airflow (AF) and oximetry (SpO_2) signals from polysomnography (PSG) was performed prior to the application of the Convolutional Neural Network (CNN) based apnea-hypopnea index (AHI) estimation algorithm. The Childhood Adenotonsillectomy Trial (CHAT) and the University of Chicago (UofC) databases were employed to train, validate, and test the proposed methodology. LPF: low-pass filter; TRT: total recording time; SVR: Support Vector Regression; S_n: segment n; y_n: label n.

3.2. Design of the CNN architecture

In this study, a CNN model was trained and validated with the extracted segments of AF and SpO_2 signals to estimate the number of apneic events. The CNN architecture was arranged into *Nlay* stacked convolutional blocks that sequentially processed the input data (Fig. 3). Each convolutional block was composed of 5 consecutive layers: convolution, batch normalization, activation, max pooling, and dropout. First, a convolutional layer generated 3D feature maps using the 2D convolution operation [39,45]:

$$x_i^j[m,n] = \sum_{k=1}^{ksize} \sum_{l=1}^{2} w_i^j[k,l] \cdot a_i[m-k+1, n-l+1] + b_i^j, \quad (1)$$

where x_i^j is the feature map generated in the convolutional block i ($i = 1, \ldots, Nlay$), with the filter with weights w_i^j and bias b_i^j ($j = 1, \ldots, Nfilt$) and a_i as the input to the i-th convolutional block. The convolutional layer was composed of a bank of *Nfilt* 2D filters with kernel size *ksize* × 2, stride 1, and zero padding to ensure that the input and output lengths are the same. Next, the batch normalization layer applied a normalization of the *Nfilt* feature maps generated by the previous layer [46]. The rectified linear unit (ReLU) activation was then applied [45]:

$$ReLU\left(x_i^j\right) = \max\left(0, x_i^j\right), \quad (2)$$

where x_i^j is the value of each sample of the feature map. ReLU activation is standard in deep architectures to accelerate the training process of the CNN without using previously optimized weights [39]. Dimensionality reduction was applied to the activations using a max pooling layer with a pool factor 2 × 1 to reduce the length of the feature maps while the width and depth are kept. The last layer of the convolutional blocks is a dropout layer that randomly removed a small fraction of the activations during each training batch to reduce overfitting. In this study, we have tested drop probabilities ranging from $P = 0$ to $P = 0.5$ [20].

Next to the convolutional blocks, a flattening layer was employed to reshape the 3D feature maps into 1D feature vectors. The output layer was a linear activation unit that yielded the estimation of the number of apneic events present in each epoch [36].

3.3. CNN training and optimization

The training stage of the CNN was initialized with random weights in all layers. The optimizer employed in this study was the adaptive momentum estimation (Adam). The initial learning rate was sought among various values between $5.0 \cdot 10^{-5}$ and 0.02, while the momentum-related parameters were set to their default values $\beta 1 = 0.9$ and $\beta 2 = 0.999$ [45,47]. Training data was divided into mini-batches, which were shuffled before each training epoch to ensure independence between consecutive batches and improve convergence [45]. The size of the mini-batches was sought between 64 and 512. The loss function used in the Adam optimizer was the Huber loss, that is widely used for robust regression if large outliers are present [48]. The delta (δ) parameter of the Huber loss was fixed to $\delta = 1$, an intermediate value that is suitable for arbitrary data distributions [48]. The validation set was used during the learning stage to monitor the convergence of the CNN optimization process with the validation loss. The learning rate was reduced by a factor of 2 during training when the loss in the validation set did not improve in the last 10 epochs [45,49]. In addition, an early stopping criterion was applied to avoid overfitting. The training was stopped after 30 epochs with no reduction of the loss in the validation set, restoring the weights to those obtained in the epoch with minimum validation loss [45,49]. Otherwise, the maximum number of epochs was 100. Keras 2.3 framework with Tensorflow-gpu 2.0 backend was used to implement, train, and evaluate the CNN models [49].

3.4. Apnea-hypopnea index estimation

After the estimated number of apneic events in each 5-min segment was obtained, the rate of respiratory events was calculated as the sum of the detected apneic and hypopneic events divided by the total recording

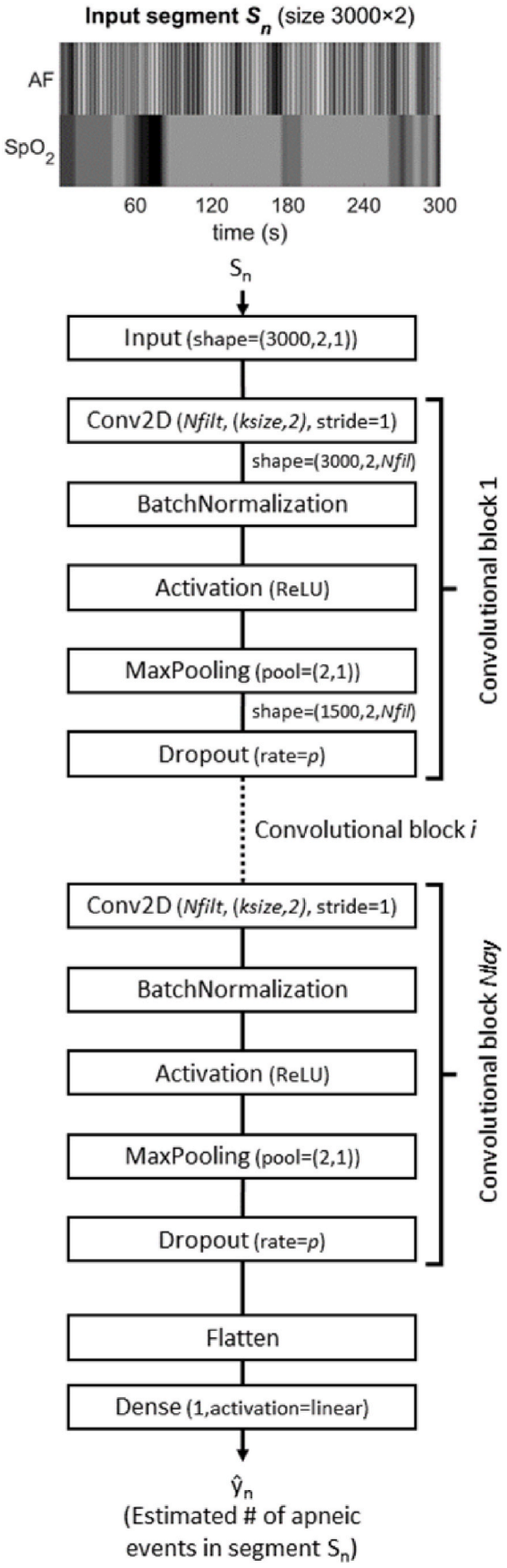


Fig. 3. Block diagram of the Convolutional Neural Network developed in this study. Conv2D: bidimensional (2D) convolution layer; *Nfil*: number of filters in the convolution layer; *ksize*: kernel size of the convolution filters; ReLU: Rectified Linear Unit activation function; *Nlay*: number of blocks of convolutional layers.

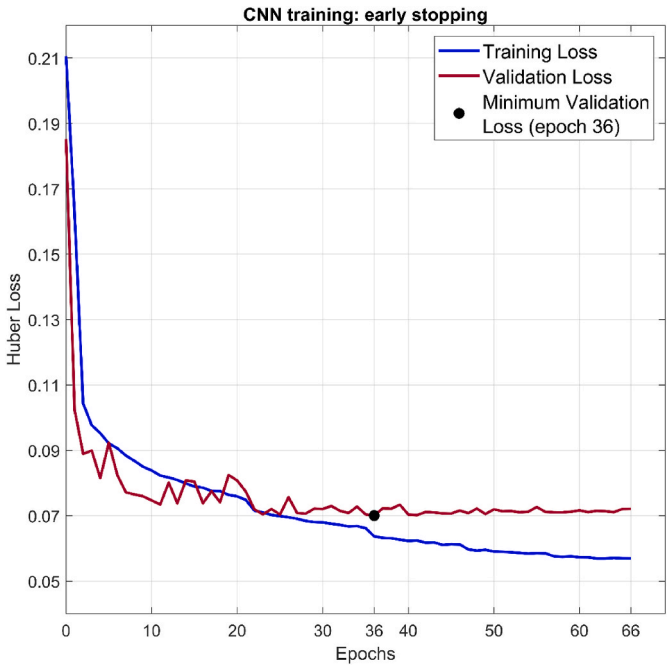


Fig. 4. Evolution of the Huber Loss in the training and validation sets during the CNN optimization alongside the training epochs.

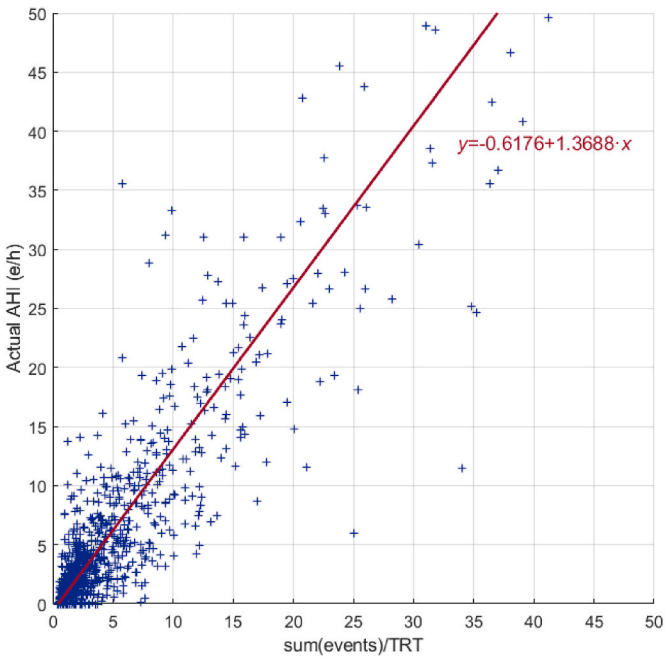


Fig. 5. Scatter plot and regression function of the Apnea-Hypopnea Index (AHI) from the rate of apneic events in the total recording time (TRT) in the validation set.

time (Fig. 2). This rate underestimates the actual AHI, since the total recording time is usually larger than the total sleep time. To correct this tendency, the final estimated AHI was calculated by means of a Support Vector Regression (SVR) model [50,51]. This method conducts a more robust regression than ordinary least squares in presence of noisy data [51]. The SVR optimization method finds the optimal coefficients of the linear equation $f(x) = a_0 + a_1 \cdot x$ that minimizes the error function. This error function is a ε-insensitive loss [50]:

$$L(y,f(x)) = \begin{cases} 0 & , |y - f(x)| \leq \varepsilon \\ |y - f(x)| - \varepsilon, & |y - f(x)| \geq \varepsilon \end{cases} . \quad (3)$$

The margin $\varepsilon \geq 0$ has to be fixed to define a zone around $f(x)$ where the difference between the actual AHI and the estimated AHI does not contribute to the error. In the present work, this positive margin was empirically set to $\varepsilon = 0.25$ to minimize misclassifications into different OSA severity levels, assuming that an error in the AHI estimation lower than the margin is very unlikely to contribute to the misclassification of OSA severity. Therefore, the margin was chosen to be lower than the most restrictive AHI cutoff that define the presence of OSA. Three accepted and child-specific AHI cutoffs were used to classify the severity of OSA into four levels: no OSA (AHI<1 e/h), mild OSA (1 ≤ AHI<5 e/h), moderate OSA (5 ≤ AHI<10 e/h) and severe OSA (AHI≥10 e/h) [2,4].

Table 2
Results of the CNN hyperparameters optimization process in the validation set.

Nlay	*Nfilt*	*ksize*	Cohen's κ
5	8	3	0.2561
5	8	5	0.3759
5	16	5	0.3784
6	16	5	0.4132
6	32	7	0.4457
6	32	9	0.4697
7	32	9	0.4533
7	32	17	0.4702
8	32	9	0.4711
8	64	9	0.4673
8	**64**	**17**	**0.4807**
8	64	33	0.4697
8	96	17	0.4624
9	64	17	0.4533
9	64	33	0.4387
9	96	17	0.4617

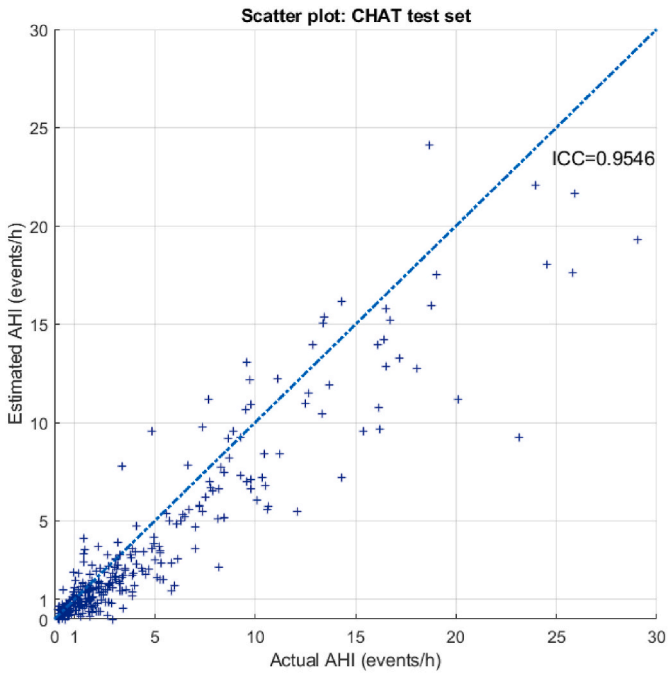
Nlay = Number of convolutional layers, *Nfilt* = Number of filters in each convolutional layer, *ksize* = length of the filters, κ = Cohen's kappa coefficient.

3.5. Hyperparameter search and algorithm evaluation

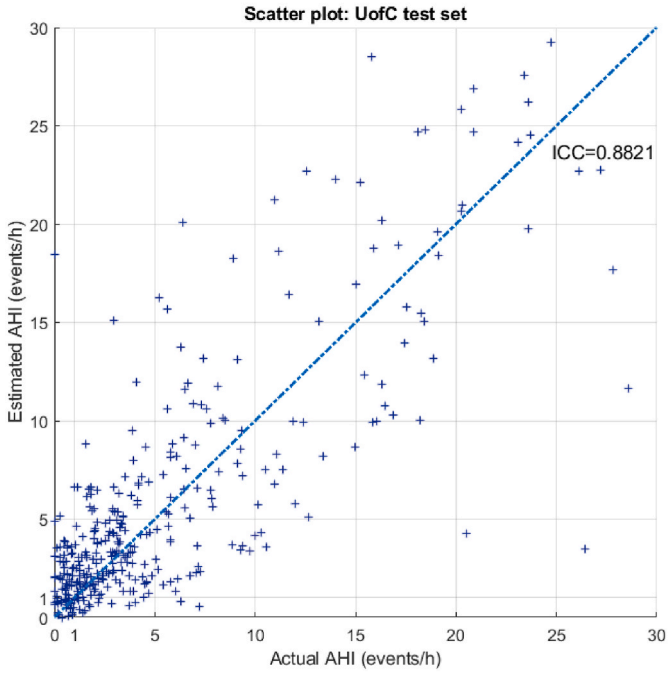
Three CNN hyperparameters were optimized in this study: *Nlay*, *Nfilt*, and *ksize*. These hyperparameters are directly involved in controlling the size of the network and the number of trainable weights. We evaluated the performance of the CNNs with varying values of *Nlay* = {5, 6, 7, 8, 9}, *Nfilt* = {8, 16, 32, 64, 96}, and *ksize* = {3, 5, 7, 9, 17, 33}. Based on preliminary analyses, these values were selected to define a CNN with intermediate complexity, i.e., lower values led to underfitting, and higher values resulted in a CNN that was hard to train and prone to overfitting. The optimal hyperparameters were those that maximized the 4-class Cohen's Kappa coefficient (κ) when classifying OSA severity in the validation set using stratified 10-fold cross-validation. The Cohen's kappa is a measure of agreement useful in classification tasks when the classes are not equally distributed, because it considers the probability of agreement by chance and is less biased towards the majority class as the accuracy is [52]. A search over the hyperparameter values was performed by training the CNNs and evaluating them with the validation data.

3.6. Statistical analysis

The agreement between the AHI estimated using the proposed methodology and the actual AHI derived from the PSG was evaluated by means of the intraclass correlation coefficient (*ICC*), scatter plots, and Bland-Altman plots [53]. The classification into the four severity levels was assessed using confusion matrices, 4-class accuracy (Acc_4), and κ [52]. The diagnostic ability of the algorithm in the AHI cutoffs (1, 5 and 10 e/h) was evaluated by means of sensitivity (*Se*), specificity (*Sp*), accuracy (*Acc*), positive and negative predictive values (*PPV*, *NPV*), and positive and negative likelihood ratios (*LR*+, *LR*−).

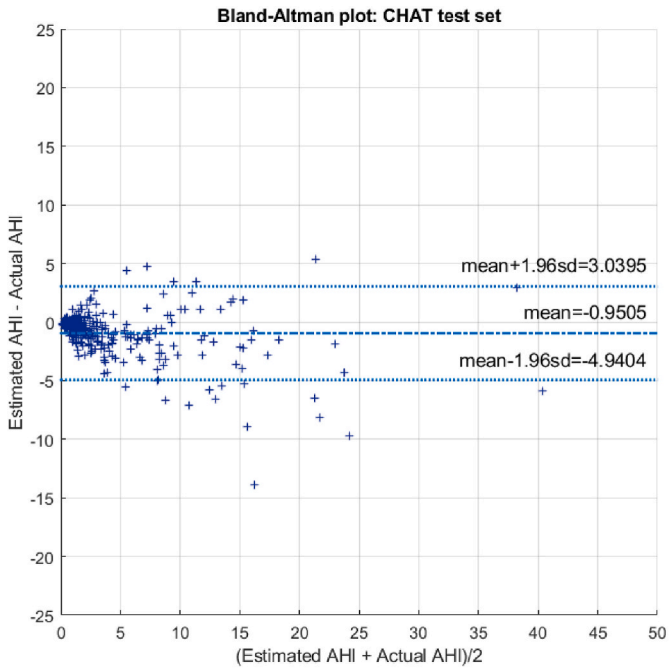


(a)

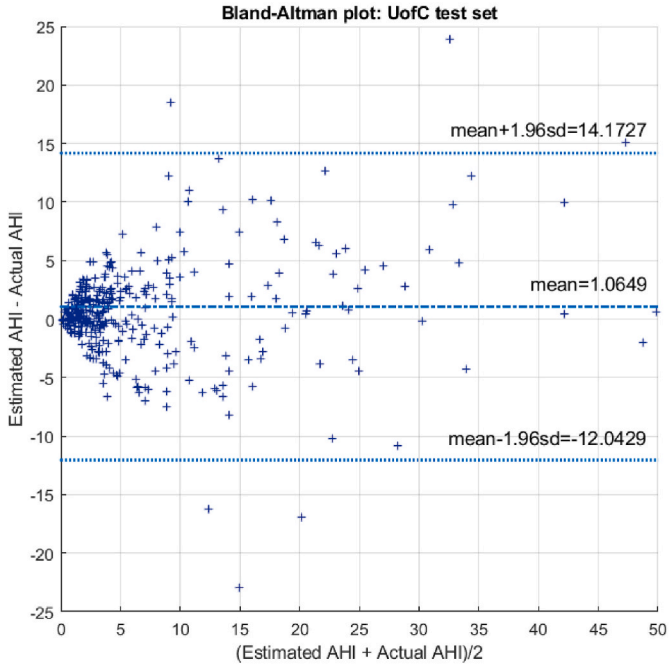


(b)

Fig. 6. Scatter plots of the actual Apnea-Hypopnea Index (AHI) *vs*. the estimated AHI in (a) CHAT and (b) UofC test sets.



(a)



(b)

Fig. 7. Bland-Altman plots of actual and estimated Apnea-Hypopnea Index (AHI) in (a) CHAT and (b) UofC test sets.

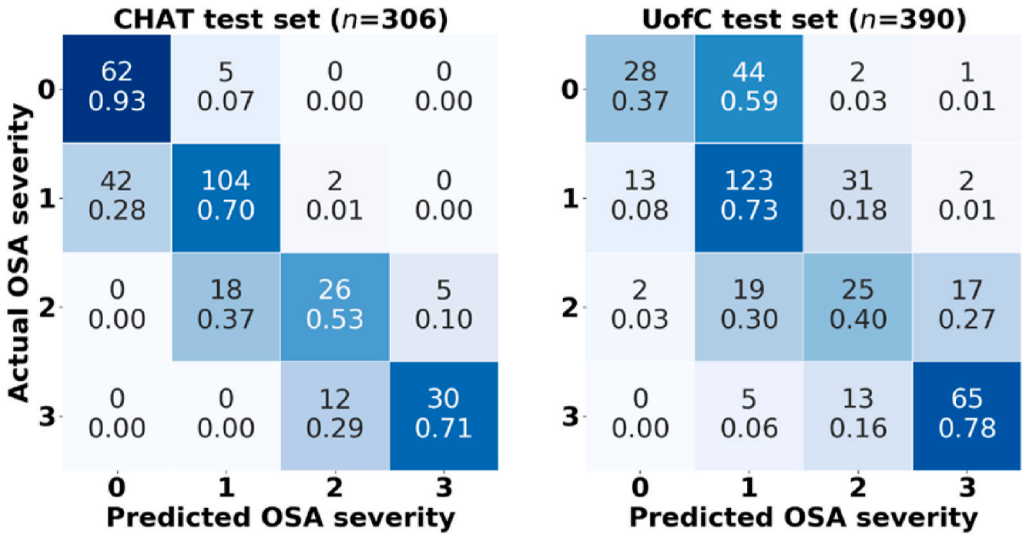


Fig. 8. Confusion matrices of the predicted OSA severity against the actual OSA severity in CHAT and UofC test sets. 0: no OSA; 1: mild; 2: moderate; 3: severe.

Table 3
Performance of the proposed algorithm in the test sets of the CHAT and UofC databases.

Test set (*n*)	ICC	Acc_4 (%)	Cohen's κ
CHAT-baseline (*n*=86)	0.8793	66.28	0.4978
CHAT-followup (*n*=78)	0.9786	78.21	0.6711
CHAT-nonrandomized (*n*=142)	0.9537	73.24	0.5652
CHAT-all (*n*=306)	0.9546	72.55	0.6011
UofC (*n*=390)	0.8821	61.79	0.4469

n = Number of subjects in the set, ICC = intra-class correlation coefficient, Acc_4 = four-class accuracy, κ = Cohen's kappa coefficient, CHAT = Childhood Adenotonsillectomy Trial, UofC = University of Chicago.

4. Results

4.1. Optimum CNN architecture

In this study, a CNN model was trained and evaluated, and the optimal combination of the hyperparameters *Nlay*, *Nfilt* and *ksize* was sought. The result of the optimization carried out during the training stage is shown in Fig. 4, where it can be seen that the Huber loss decreases alongside the training epochs. It can also be observed that the training was halted 30 epochs after the minimum validation loss was reached. At that epoch, the validation loss is minimum and similar to the training loss, indicating that the CNN does not overfit the training data. Table 2 shows the Cohen's κ values estimated in the validation set, with subjects from both CHAT and UofC databases. For the sake of simplicity, only the configurations that sensibly differ from similar combinations of hyperparameters are shown. The configuration that reached the highest $\kappa = 0.4807$ was $Nlay = 8$, $Nfilt = 64$, and $ksize = 17$ (Table 2). This optimum model was obtained with the following hyperparameters in the training stage: drop probability $P = 0.1$, learning rate 0.01, and batch size 128.

Fig. 5 shows the scatter plot of the estimated rate of respiratory events per hour of recording against the actual AHI using the optimal CNN configuration, and the regression function derived from the SVR model to estimate the AHI in the validation set. The equation corrects the tendency to underestimate the AHI when using the total recording time instead of the total sleep time.

4.2. Diagnostic ability in the test set

Once the optimal configuration of the algorithm was achieved from data in the training and validation sets, the model was applied to the test dataset. The scatter plots shown in Fig. 6 reveal that the agreement between the actual and estimated AHI is higher in the CHAT test set in comparison with the UofC data. Similarly, Fig. 7 Bland-Altman plots show that the difference between actual and estimated AHI is more dispersed in the UofC test set. The proposed algorithm showed a slight tendency to underestimate the AHI in the CHAT test set, contrarily to that observed in the UofC test set. Fig. 8 shows the confusion matrices obtained in both test sets, and Tables 3 and 4 display the diagnostic performances of our proposed CNN-based approach. In accordance with the tendencies shown in the Bland-Altman plots (Fig. 7), the results of binary classification show that *Sp* and *PPV* were higher in the CHAT test set at the cost of reducing *Se*, which was higher in the UofC test set. Overall, the highest *Acc* and *LR* + values were achieved in CHAT test set.

Table 4
Diagnostic ability of the algorithm for the AHI cutoffs 1, 5, and 10 events/h in the test sets of CHAT and UofC databases.

AHI cutoff	Test set	Se (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	LR+	LR-
1 e/h	**CHAT**	82.43	92.54	84.64	97.52	59.62	11.0452	0.1899
	UofC	95.24	37.33	84.10	86.46	65.12	1.5198	0.1276
5 e/h	**CHAT**	80.22	99.07	93.46	97.33	92.21	86.2363	0.1997
	UofC	82.19	85.25	84.10	76.92	88.89	5.5708	0.2089
10 e/h	**CHAT**	71.43	98.11	94.44	85.71	95.57	37.7143	0.2912
	UofC	78.31	93.49	90.26	76.47	94.10	12.0211	0.2320

AHI = apnea-hypopnea index, Se = sensitivity, Sp = specificity, Acc = accuracy, PPV = positive predictive value, NPV = negative predictive value, LR+ = positive likelihood ratio, LR- = negative likelihood ratio, e/h = events/hour CHAT = Childhood Adenotonsillectomy Trial, UofC = University of Chicago.

Table 5
Diagnostic performance of state-of-the-art approaches in the childhood OSA context.

Study	Signal	Methods (Extraction/Selection/Classification/Validation)	Nº Subjects	Cutoff (events/h)	Se (%)	Sp (%)	Acc (%)
Choi et al. (2018) [30]	AF	CNN/Holdout	179 (adults)	5	100.0	84.6	96.2
				15	98.1	86.5	92.3
				30	96.2	96.2	96.2
Nikkonen et al. (2019) [33]	SpO_2	ANN/Holdout	1959 (adults)	5	96.1	87.5	92.1
				15	94.7	97.0	96.4
				30	88.2	99.1	97.5
Wu et al. (2017) [9]	–	Clinical Parameters/-/Stepwise LR/Holdout	311	5	**94.8**	25.0	78.2
Garde et al. (2014) [11]	SpO_2, PRV	Time statistics, spectral, nonlinear/AUC optimization/LDA/Four-fold cross validation	146	5	**88.4**	83.6	84.9
Garde et al. (2019) [12]	SpO_2, PRV	Time statistics, spectral, ODI (SpO_2), spectral (PRV)/Stepwise LR/Binary LR (for each cutoff)/Holdout	207	1	68.0	86.0	71.0
				5	58.0	89.0	78.0
				10	90.0	87.0	88.0
Alvarez et al. (2018) [13]	SpO_2	Time statistics, ODI, symbolic dynamics/FSLR/LR/Bootstrap	142	5	73.5	89.5	83.3
Calderón et al. (2020) [10]	SpO_2	Oxygen desaturations, ODI/-/LR/15-fold cross validation	453 (CHAT)	5	62.0	96.0	79.0
Hornero et al. (2017) [15]	SpO_2	Time statistics, spectral, nonlinear, ODI/FCBF/MLP regression/Holdout	4191	1	84.0	53.2	75.2
				5	68.2	87.2	81.7
				10	68.7	94.1	90.2
Xu et al. (2019) [16]	SpO_2	ODI and M3f/FCBF/MLP regression/Direct validation	432	1	**95.3**	19.1	79.6
				5	77.8	80.5	79.4
				10	73.5	92.7	88.2
Vaquerizo-Villar et al. (2018) [17]	SpO_2	DFA, ODI/FCBF/MLP regression/Holdout	981 (UofC)	1	**97.1**	23.3	82.7
				5	78.8	83.7	81.9
				10	77.1	**94.8**	**91.1**
Barroso-García (2021a) [18]	AF, SpO_2	Bispectral (AF), ODI (SpO_2)/FCBF/MLP regression/Bootstrap	946 (UofC)	1	**98.0**	15.3	82.2
				5	81.6	83.0	82.5
				10	72.3	**95.0**	90.2
Barroso-García (2021b) [19]	AF, SpO_2	Wavelet (AF), ODI (SpO_2)/FCBF/BY-MLP regression/Bootstrap	946 (UofC)	1	91.2	**43.3**	82.0
				5	79.3	83.8	82.1
				10	74.9	**95.0**	90.7
Jiménez-García et al. (2020) [14]	AF, SpO_2	Time statistics, spectral, nonlinear, ODI/FCBF/Multiclass AdaBoost/Holdout	974 (UofC)	1	92.1	36.0	81.3
				5	76.0	**85.7**	82.1
				10	62.7	**97.7**	90.3
Vaquerizo-Villar et al. (2021) [20]	SpO_2	CNN/Holdout	1638 (CHAT)	1	71.2	81.8	77.6
				5	**83.7**	**100.0**	**97.4**
				10	**83.9**	**99.3**	**97.8**
			980 (UofC)	1	90.8	36.4	80.1
				5	76.0	**88.6**	83.9
				10	**79.5**	**95.8**	**92.3**
This study	**AF, SpO_2**	**CNN/Holdout**	**1638 (CHAT)**	**1**	**82.4**	**92.5**	**84.6**
				5	**80.2**	**99.1**	**93.5**
				10	**71.4**	**98.1**	**94.4**
			974 (UofC)	**1**	**95.2**	**37.3**	**84.1**
				5	**82.2**	**85.3**	**84.1**
				10	**78.3**	**93.5**	**90.3**

Acc = Accuracy, AF = Airflow signal, ANN = Artificial Neural Network, AUC = Area under the receiver operating characteristic curve, BY-MLP = Multilayer perceptron neural network with Bayesian approach, CHAT = Childhood Adenotonsillectomy Trial, CNN = Convolutional neural network, DFA = Detrended fluctuation analysis, FCBF = Fast correlation-based filter, FSLR = Forward stepwise logistic regression, LDA = Linear discriminant analysis, LR = Logistic regression, M3f = 3rd order statistical moment in the frequency band, MLP = Multilayer perceptron neural network, ODI = Oxygen desaturation index, PRV = Pulse rate variability Se = Sensitivity, Sp = Specificity, SpO_2 = Oxygen saturation signal, UofC = University of Chicago.

5. Discussion

The diagnostic ability of a CNN architecture fed with minimally preprocessed AF and SpO_2 signals to detect OSA in the pediatric population was evaluated in the present study. The databases covered a total of 2612 pediatric overnight sleep studies, which were used to train, validate, and evaluate the proposed algorithm. The AHI estimated by our CNN-based approach reached a high agreement with the actual AHI, as well as a remarkable diagnostic ability to detect pediatric OSA severity in both CHAT and UofC datasets.

5.1. Proposed CNN-based model

To the best of our knowledge, CNNs have been proposed in the context of pediatric OSA detection only once [20]. In that study, OSA severity was predicted using only SpO_2 data by means of a single channel 1D CNN. The present study is the first study that proposes the joint use of AF and SpO_2 with a DL architecture in pediatric OSA. AF and SpO_2 data were combined using a 2D CNN approach that outperformed the previous approach in terms of Acc_4, κ, and *Acc* in 1 and 5 e/h (Tables 3 and 4) [20]. These results may indicate that the contribution of AF enhances the identification of no OSA and mild OSA subjects, i.e., those categories in which the largest discrepancies in accuracy occur. Previous studies in the context of adult OSA employed DL-based methodologies [21], with a large proportion of these studies also using CNNs. The investigation of respiratory signals, including AF, in DL-based approaches is also frequent [27–32], as well as using SpO_2 alone [33,34,36], and a combination of signals [37]. Preprocessing of AF is also common in those approaches [29,30], and usually includes resampling or low pass filtering due to the noisy nature of the AF signal [29]. On the other hand, SpO_2 required minimal preprocessing, and is often limited to data resampling [33,36]. Our algorithm was aimed at estimating the AHI [20,33,36], in contrast with other approaches that have focused on detecting single apnea or hypopnea events. To accomplish such goal, the algorithm focused on estimating the number of respiratory events on a 5-min basis and calculating the AHI. Previous studies aimed at event detection usually classify every time step as apneic/hypopneic or normal using short windows that must be adapted to the duration of the apneic events. On the contrary, our approach took advantage of employing large segments that can encompass both various apneic and hypopneic events in the AF signal, and the possible oxygen desaturations in the SpO_2 signal, to perform a regression of the AHI by predicting the number of respiratory events in each segment. These apneic events have much shorter duration and delay between apneas/hypopneas and oxygen desaturations than the duration of the segment and can often occur repeatedly in the period that encompasses a segment, so the regression over large segments is a straightforward approach to perform AHI estimation [20]. We have also tested our CNN-based approach using 1-min segments, obtaining remarkably lower performance ($\kappa = 0.3963$ vs. $\kappa = 0.4807$ in the validation set). This may indicate that the regression-based approach employed in this study performs better with large segments, which is consistent with the fact that apneic events are typically grouped in clusters. The choice of 5-min segments with 50% overlap allowed us to increase the number of examples in our dataset to train and validate the CNN, thus reducing the choices of overfitting, while preventing any decreases in the accuracy of the estimation. The results of the hyperparameters optimization in the validation set pointed that only slight differences in the values of κ can be observed with *Nlay* = 7–8 (Table 2). The top performing configuration was *Nlay* = 8, *Nfilt* = 64 and *ksize* = 17, although similar performance was achieved with *Nlay* = 6–8, *Nfil* = 32–64 and *ksize* = 9–17. The simplest configurations (i.e., *Nlay* = 5, *Nfil* = 8–16 or *ksize* = 3–7) showed lower performance, suggesting that those CNN architectures lack of capacity to identify the patterns associated to apneas/hypopneas. The diagnostic ability did not improve using larger values of these hyperparameters, suggesting that more complex CNNs have less generalization ability with higher computational cost (see Table 2).

5.2. Diagnostic performance of the algorithm

Our CNN-based approach reached high agreement to predict OSA severity from AF and SpO_2 compared with standard manually-scored PSGs (Table 3), thus achieving remarkable diagnostic ability for every AHI cutoff in both CHAT and UofC datasets (Table 4). The values of *ICC*, Acc_4 and κ were higher in the CHAT test set. In preliminary analyses of this study, these differences were even higher when the algorithm was trained and validated in the CHAT dataset, and finally tested in the UofC database ($Acc_4 = 75.49\%$, $\kappa = 0.6253$ in the CHAT test set, and $Acc_4 = 56.92\%$, $\kappa = 0.3693$ in the UofC test set). The CNN architecture was only trained with CHAT data, possibly causing a bias that limited its generalization ability. Of note, a possible reason for the discrepancies between the two datasets resides in the methodology employed for scoring the CHAT dataset, i.e., always the same scorers for all recordings, as opposed to the more real life UofC dataset in which there were multiple scorers as dictated by the clinical operations within the sleep center, as well as different scoring criteria among databases. We therefore included subjects from the two databases in the validation set to minimize this issue, resulting in an enhanced diagnostic performance in the UofC database (Table 3). These 4-class classification results also seem not being affected by the uneven proportion of subjects in each OSA severity class. As indicated in Table 1, the proportion of subjects in each severity group is unbalanced, but consistent across the training, validation, and test datasets. After balancing the training and validation data to have 25% of instances of each severity class, the diagnostic ability of the AHI estimation algorithm did not leverage the effect of class balancing ($Acc_4 = 72.22\%$, $\kappa = 0.5969$ in CHAT test set, and $Acc_4 = 61.28\%$, $\kappa = 0.4402$ in UofC test set). This suggests the robustness of our AHI estimation algorithm, which was not influenced by the distribution of OSA severity. Diagnostic performances were also different in the CHAT subgroups. *ICC*, Acc_4 and κ were lower in the baseline subgroup. All subjects in the baseline subgroup had AHI >1 e/h, and the proportion of moderate and severe OSA subjects was larger than those in other subgroups. Figs. 6(a) and 7(a) show that the AHI estimation is more precise for no OSA and mild OSA subjects, i.e., the conditions that are more prevalent in follow-up and nonrandomized subgroups. In addition, the slight tendency to underestimate the AHI in the CHAT test set may have increased the number of false negatives among baseline sleep studies. In the case of subjects with higher AHI in both CHAT and UofC tests sets, the precision of the algorithm is lower. This is an issue of lesser importance in the case of severe OSA patients because the predicted AHI is likely to be greater than 10 e/h, as shown in the confusion matrices (Fig. 8), such that there should not be any severity category mislabeling. As mentioned, the differences in the results between the two databases may be caused by the interrater variability differences in the databases due to differences in PSG scoring methodologies. Of note, this is a consistent issue when evaluating PSGs, and illustrates how a priori such issues might influence the diagnostic performance of the algorithms [54].

5.3. Comparison with previous studies

Other studies have also combined AF and SpO_2 data by means of feature-engineering approaches [14,18,19]. These algorithms employed the UofC database in their development, and the algorithm in the present study outperformed them in terms of Acc_4 (61.79% vs 57.95%–58.57%) and κ (0.4469 vs 0.3800–0.4088) using roughly the same subjects in the test set [14,18,19]. In this sense, the application of DL-based algorithms seems to be advantageous in two key aspects: they substitute the development of specifically tailored feature extraction methods by an automatic feature learning methodology, and they demonstrate superior diagnostic ability. With regard to similar approaches in adult OSA, Choi et al. reached *Acc* = 96.20%, 92.30% and

96.20% in the common adult AHI cutoffs (5, 15 and 30 e/h) using the nasal pressure AF signal [30], while Nikkonen et al. reached *Acc* = 92.14%, 96.38% and 97.50% in the same thresholds using SpO_2 data (see Table 5) [33]. Regarding Acc_4, the performance ranged from 80.20% to 88.3% [30,33,36,37]. These performances are higher than those reached in studies focusing on pediatric OSA, mainly due to the less stringent rules for both scoring apneas and hypopneas and defining the AHI cutoffs employed to determine OSA severity in adults [3]. Regarding the results obtained using the ECG signal in the adult context, some previous studies reached *Acc* = 100% in subject-based OSA classification [25,26], although the test set of the Apnea-ECG database employed in these studies only encompasses 35 subjects from a total of 70, which does not follow standard rules for the annotations of apneas/hypopneas.

ML methods in general, and DL in particular have been seldom developed and tested in childhood OSA [8]. A recent review and meta-analysis on the reliability of ML to aid in the diagnosis of pediatric OSA revealed that the robustness of OSA detection algorithms increases as the AHI cutoff becomes higher. In that meta-analysis, it was observed that *Se* decreases (84.9%, 71.4%, and 65.2% in 1, 5, and 10 e/h) and *Sp* increases for higher AHI cutoffs (49.9%, 83.2%, and 93.1% in 1, 5, and 10 e/h). The results achieved in this study showed the same trend, but surpassed previous performance in 1, 5 and 10 e/h, indicating that the current algorithm is more robust than previous approaches relying on classical ML and/or SpO_2 only.

Table 5 summarizes the results achieved in previous studies in the context of childhood OSA detection using ML-based approaches. Our results in the CHAT database, one of the biggest databases aimed at the study of childhood OSA, clearly outperformed those reported by Vaquerizo-Villar et al. [20] in 1 e/h in terms of *Se*, *Sp* and *Acc*, and were similar when using 5 e/h and 10 e/h as diagnostic cut-offs. Regarding the results in 1 e/h in the UofC database, our results were also the highest among the studies that evaluated that database, reaching the highest *Acc* while both *Se* and *Sp* were close to the highest. Only Barroso-Garcia et al. [19] obtained a higher *Sp*, but with lower *Se*. Our results remained among the highest in 5 and 10 e/h when considering data other than CHAT. The *Acc* in 5 e/h was close to the highest, showing the most balanced *Se-Sp* pair as well. Only Garde et al. [11] reached slightly higher *Acc* in 5 e/h using a binary approach that was tested in only 142 subjects. We obtained similar *Acc* in 10 e/h than previous approaches in the UofC database, but the balance between *Se* and *Sp* was enhanced in the present study. Only the CNN-based algorithm trained with single-channel SpO_2 [20] outperformed our two-channel approach in 10 e/h. These findings suggest that the use of SpO_2 with a CNN algorithm would suffice to assess the presence of OSA in 10 e/h, but the combination of AF and SpO_2 enhances the diagnostic ability of the CNN in 1 and 5 e/h. Therefore, the algorithm proposed in the present study is more suitable than previous approaches that rely on the analysis of SpO_2. In summary, this study confirms the suitability of novel CNN-based algorithms fed with AF and SpO_2 to solve the challenging problem of the development of a simplified and minimally invasive computer aided diagnostic tool for childhood OSA.

5.4. Limitations and future work

It is necessary to note some limitations of our study. The sequential structure of the CNNs employed in this study can be substituted by alternative CNN structures, such as Inception or ResNet, or combined with RNN architectures. An interesting future goal could be the identification of the specific AF and SpO_2 information that our model uses to conduct its predictions, thereby offering the possibility of providing clinically relevant knowledge. The differences observed between the databases employed in this study might have limited the generalization ability of our approach. Although the present study included the validation of the methodology in a mixed dataset, as well as the assessment of the algorithm in two independent tests sets, an even more exhaustive validation process would increase the robustness of our results, for example by including recordings from a wider range of pediatric sleep centers.

6. Conclusion

A CNN architecture fed with AF and SpO_2 data demonstrated high diagnostic ability in the identification of OSA severity in children. The performance of our proposed algorithm surpassed other previous automated approaches in pediatric OSA, and showed enhanced differentiation of no OSA against mild and moderate OSA patients with respect to state-of-the-art approaches. The results of this study suggest that a CNN-based algorithm can reliably be used to assess the presence of OSA in children. Therefore, the usefulness of this approach allows advances in the development of computer aided diagnostic tools to automatically detect pediatric OSA using a reduced set of signals.

Ethical approval

This work has been carried out according to the Declaration of Helsinki. The informed consents of all children caretakers were obtained, and the Ethics Committee of the Comer Children's Hospital approved the protocol of the study (#11-0268-AM017, #09-115-B-AM031, and #IRB14-1241).

Authorship contribution statement

Data collection: L. Kheirandish-Gozal and D. Gozal; Medical diagnostic: L. Kheirandish-Gozal and D. Gozal; Study design: J. Jiménez-García, M. García, G.C. Gutiérrez-Tobal, F. del Campo, and R. Hornero. Implementation: J. Jiménez-García and F. Vaquerizo-Villar. Data analysis: J. Jiménez-García, M. García, G. C. Gutiérrez-Tobal, F. Vaquerizo-Villar, and D. Álvarez. Manuscript writing: J. Jiménez-García, M. García, G.C. Gutiérrez-Tobal, L. Kheirandish-Gozal. F. Vaquerizo-Villar, D. Álvarez, F. Del Campo, D. Gozal, and R. Hornero. Manuscript review: J. Jiménez-García, M. García, G.C. Gutiérrez-Tobal, L. Kheirandish-Gozal, F. Vaquerizo-Villar, D. Álvarez, F del Campo, D. Gozal and R. Hornero. Funding acquisition: R. Hornero, F del Campo, D. Álvarez, L. Kheirandish-Gozal, and D. Gozal. All authors gave their final approval of this version of the manuscript.

Declaration of competing interest

There are no conflicts of interest that could inappropriately influence this research work.

Acknowledgements

This research was supported by 'Ministerio de Ciencia, Innovación y Universidades - Agencia Estatal de Investigación' 10.13039/501100011033, 'European Regional Development Fund A way of making Europe' (ERDF), and the "European Union NextGenerationEU/PRTR" under projects PID2020-115468RB-I00 and PDC2021-120775-I00, 'Sociedad Española de Neumología y Cirugía Torácica (SEPAR)' under project 649/2018, 'Sociedad Española de Sueño (SES)' under project "Beca de Investigación SES 2019", and by 'CIBER - Consorcio Centro de Investigación Biomédica en Red- en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN) (CB19/01/00012)' through 'Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación'. The Childhood Adenotonsillectomy Trial (CHAT) was supported by the National Institutes of Health (HL083075, HL083129, UL1-RR-024134, UL1 RR024989). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002). J. Jiménez-García was in receipt of a PIF-UVa grant of the University of Valladolid. F. Vaquerizo-Villar was in receipt of a "Ayuda para contratos predoctorales para la Formación de Profesorado

Universitario (FPU)" grant from the "Ministerio de Educación, Cultura y Deporte" (FPU16/02938). D. Álvarez is supported by a "Ramón y Cajal" grant (RYC2019-028566-I) from the 'Ministerio de Ciencia e Innovación - Agencia Estatal de Investigación' co-funded by the European Social Fund. Leila Kheirandish-Gozal and David Gozal are supported by 'National Institutes of Health (NIH)' grants HL130984, HL140548, and AG061824, and by the Leda J Sears Foundation for Pediatric Research.

References

[1] C.L. Marcus, L.J. Brooks, S.D. Ward, K.A. Draper, D. Gozal, A.C. Halbower, J. Jones, C. Lehmann, M.S. Schechter, S. Sheldon, R.N. Shiffman, K. Spruyt, Diagnosis and management of childhood obstructive sleep apnea syndrome, Pediatrics 130 (2012) e714–e755, https://doi.org/10.1542/peds.2012-1672.

[2] E. Dehlink, H.-L. Tan, Update on paediatric obstructive sleep apnoea, J. Thorac. Dis. 8 (2016) 224–235, https://doi.org/10.3978/j.issn.2072-1439.2015.12.04.

[3] R.B. Berry, S.F. Quan, A. Abreu, et al., for the A.A. of S. Medicine, The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 2.6, 2020. Darien, IL, www.aasmnet.org.

[4] H.-L. Tan, H.P.R. Bandla, H.M. Ramirez, D. Gozal, L. Kheirandish-Gozal, Overnight polysomnography versus respiratory polygraphy in the diagnosis of pediatric obstructive sleep apnea, Sleep 37 (2014) 255–260, https://doi.org/10.5665/sleep.3392.

[5] H.-L. Tan, L. Kheirandish-Gozal, D. Gozal, Pediatric home sleep apnea testing, Chest 148 (2015) 1382–1395, https://doi.org/10.1378/chest.15-1365.

[6] K.F. Joosten, H. Larramona, S. Miano, D. Van Waardenburg, A.G. Kaditis, N. Vandenbussche, R. Ersu, How do we recognize the child with OSAS? Pediatr. Pulmonol. 52 (2017) 260–271, https://doi.org/10.1002/ppul.23639.

[7] D. Bertoni, A. Isaiah, Towards patient-centered diagnosis of pediatric obstructive sleep apnea—a review of biomedical engineering strategies, Expet Rev. Med. Dev. 16 (2019) 617–629, https://doi.org/10.1080/17434440.2019.1626233.

[8] G.C. Gutiérrez-Tobal, D. Álvarez, L. Kheirandish-Gozal, F. Campo, D. Gozal, R. Hornero, Reliability of machine learning to diagnose pediatric obstructive sleep apnea: systematic review and meta-analysis, Pediatr. Pulmonol. (2021), 25423, https://doi.org/10.1002/ppul.25423 ppul.

[9] D. Wu, X. Li, X. Guo, J. Qin, S. Li, A simple diagnostic scale based on the analysis and screening of clinical parameters in paediatric obstructive sleep apnoea hypopnea syndrome, J. Laryngol. Otol. 131 (2017) 363–367, https://doi.org/10.1017/S0022215117000238.

[10] J.M. Calderón, J. Álvarez-Pitti, I. Cuenca, F. Ponce, P. Redon, Development of a minimally invasive screening tool to identify obese Pediatric population at risk of obstructive sleep Apnea/Hypopnea syndrome, Bioengineering 7 (2020) 1–13, https://doi.org/10.3390/bioengineering7040131.

[11] A. Garde, P. Dehkordi, W. Karlen, D. Wensley, J.M. Ansermino, G.A. Dumont, Development of a screening tool for sleep disordered breathing in children using the phone Oximeter™, PLoS One 9 (2014), e112959, https://doi.org/10.1371/journal.pone.0112959.

[12] A. Garde, X. Hoppenbrouwer, P. Dehkordi, G. Zhou, A.U. Rollinson, D. Wensley, G. A. Dumont, J.M. Ansermino, Pediatric pulse oximetry-based OSA screening at different thresholds of the apnea-hypopnea index with an expression of uncertainty for inconclusive classifications, Sleep Med. 60 (2019) 45–52, https://doi.org/10.1016/j.sleep.2018.08.027.

[13] D. Álvarez, A. Crespo, F. Vaquerizo-Villar, G.C. Gutiérrez-Tobal, A. Cerezo-Hernández, V. Barroso-García, J.M. Ansermino, G.A. Dumont, R. Hornero, F. del Campo, A. Garde, Symbolic dynamics to enhance diagnostic ability of portable oximetry from the Phone Oximeter in the detection of paediatric sleep apnoea, Physiol. Meas. 39 (2018), 104002, https://doi.org/10.1088/1361-6579/aae2a8.

[14] J. Jiménez-García, G.C. Gutiérrez-Tobal, M. García, L. Kheirandish-Gozal, A. Martín-Montero, D. Álvarez, F. del Campo, D. Gozal, R. Hornero, Assessment of airflow and oximetry signals to detect pediatric sleep apnea-hypopnea syndrome using AdaBoost, Entropy 22 (2020) 670, https://doi.org/10.3390/e22060670.

[15] R. Hornero, L. Kheirandish-Gozal, G.C. Gutiérrez-Tobal, M.F. Philby, M.L. Alonso-Álvarez, D. Álvarez, E.A. Dayyat, Z. Xu, Y.-S. Huang, M. Tamae Kakazu, A.M. Li, A. Van Eyck, P.E. Brockmann, Z. Ehsan, N. Simakajornboon, A.G. Kaditis, F. Vaquerizo-Villar, A. Crespo Sedano, O. Sans Capdevila, M. von Lukowicz, J. Terán-Santos, F. Del Campo, C.F. Poets, R. Ferreira, K. Bertran, Y. Zhang, J. Schuen, S. Verhulst, D. Gozal, Nocturnal oximetry–based evaluation of habitually snoring children, Am. J. Respir. Crit. Care Med. 196 (2017) 1591–1598, https://doi.org/10.1164/rccm.201705-0930OC.

[16] Z. Xu, G.C. Gutiérrez-Tobal, Y. Wu, L. Kheirandish-Gozal, X. Ni, R. Hornero, D. Gozal, Cloud algorithm-driven oximetry-based diagnosis of obstructive sleep apnoea in symptomatic habitually snoring children, Eur. Respir. J. 53 (2019), 1801788, https://doi.org/10.1183/13993003.01788-2018.

[17] F. Vaquerizo-Villar, D. Álvarez, L. Kheirandish-Gozal, G.C. Gutiérrez-Tobal, V. Barroso-García, A. Crespo, F. del Campo, D. Gozal, R. Hornero, Detrended fluctuation analysis of the oximetry signal to assist in paediatric sleep apnoea–hypopnoea syndrome diagnosis, Physiol. Meas. 39 (2018), 114006, https://doi.org/10.1088/1361-6579/aae66a.

[18] V. Barroso-García, G.C. Gutiérrez-Tobal, L. Kheirandish-Gozal, F. Vaquerizo-Villar, D. Álvarez, F. del Campo, D. Gozal, R. Hornero, Bispectral analysis of overnight airflow to improve the pediatric sleep apnea diagnosis, Comput. Biol. Med. 129 (2021), https://doi.org/10.1016/j.compbiomed.2020.104167.

[19] V. Barroso-García, G.C. Gutiérrez-Tobal, D. Gozal, F. Vaquerizo-Villar, D. Álvarez, F. Del Campo, L. Kheirandish-Gozal, R. Hornero, Wavelet analysis of overnight airflow to detect obstructive sleep apnea in children, Sensors 21 (2021) 1–19, https://doi.org/10.3390/s21041491.

[20] F. Vaquerizo-Villar, D. Alvarez, L. Kheirandish-Gozal, G.C. Gutierrez-Tobal, V. Barroso-Garcia, E. Santamaria-Vazquez, F. del Campo, D. Gozal, R. Hornero, A convolutional neural network architecture to enhance oximetry ability to diagnose pediatric obstructive sleep apnea, IEEE J. Biomed. Heal. Informatics. 25 (2021) 2906–2916, https://doi.org/10.1109/JBHI.2020.3048901.

[21] S.S. Mostafa, F. Mendonça, A.G. Ravelo-García, F. Morgado-Dias, A systematic review of detecting sleep apnea using deep learning, Sensors 19 (2019) 1–26, https://doi.org/10.3390/s19224934.

[22] U. Erdenebayar, Y.J. Kim, J.U. Park, E.Y. Joo, K.J. Lee, Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram, Comput. Methods Progr. Biomed. 180 (2019), 105001, https://doi.org/10.1016/j.cmpb.2019.105001.

[23] D. Dey, S. Chaudhuri, S. Munshi, Obstructive sleep apnoea detection using convolutional neural network based deep learning framework, Biomed. Eng. Lett. 8 (2018) 95–100, https://doi.org/10.1007/s13534-017-0055-y.

[24] A. Zarei, H. Beheshti, B.M. Asl, Detection of sleep apnea using deep neural networks and single-lead ECG signals, Biomed. Signal Process Control 71 (2022), 103125, https://doi.org/10.1016/j.bspc.2021.103125.

[25] F.R. Mashrur, M.S. Islam, D.K. Saha, S.M.R. Islam, M.A. Moni, SCNN: scalogram-based convolutional neural network to detect obstructive sleep apnea using single-lead electrocardiogram signals, Comput. Biol. Med. 134 (2021), 104532, https://doi.org/10.1016/j.compbiomed.2021.104532.

[26] Q. Yang, L. Zou, K. Wei, G. Liu, Obstructive sleep apnea detection from single-lead electrocardiogram signals using one-dimensional squeeze-and-excitation residual group network, Comput. Biol. Med. 140 (2022), 105124, https://doi.org/10.1016/j.compbiomed.2021.105124.

[27] R. Haidar, I. Koprinska, B. Jeffries, Sleep apnea event detection from nasal airflow using convolutional neural networks, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2017, pp. 819–827, https://doi.org/10.1007/978-3-319-70139-4_83.

[28] S. McCloskey, R. Haidar, I. Koprinska, B. Jeffries, Detecting Hypopnea and Obstructive Apnea Events Using Convolutional Neural Networks on Wavelet Spectrograms of Nasal Airflow, in: D. Phung, V.S. Tseng, G.I. Webb, B. Ho, M. Ganji, L. Rashidi (Eds.), Springer International Publishing, Cham, 2018, pp. 361–372, https://doi.org/10.1007/978-3-319-93034-3_29.

[29] T. Van Steenkiste, W. Groenendaal, Di Deschrijver, T. Dhaene, Automated sleep apnea detection in raw respiratory signals using long short-term memory neural networks, IEEE J. Biomed. Heal. Informatics. 23 (2019) 2354–2364, https://doi.org/10.1109/JBHI.2018.2886064.

[30] S.H. Choi, H. Yoon, H.S. Kim, H.B. Kim, H. Bin Kwon, S.M. Oh, Y.J. Lee, K.S. Park, Real-time apnea-hypopnea event detection during sleep by convolutional neural networks, Comput. Biol. Med. 100 (2018) 123–131, https://doi.org/10.1016/j.compbiomed.2018.06.028.

[31] H. Yue, Y. Lin, Y. Wu, Y. Wang, Y. Li, X. Guo, Y. Huang, W. Wen, G. Zhao, X. Pang, W. Lei, Deep learning for diagnosis and classification of obstructive sleep apnea: a nasal airflow-based multi-resolution residual network, Nat. Sci. Sleep 13 (2021) 361–373, https://doi.org/10.2147/NSS.S297856.

[32] H. Elmoaqet, M. Eid, M. Glos, M. Ryalat, T. Penzel, Deep recurrent neural networks for automatic detection of sleep apnea from single channel respiration signals, Sensors 20 (2020) 1–19, https://doi.org/10.3390/s20185037.

[33] S. Nikkonen, I.O. Afara, T. Leppänen, J. Töyräs, Artificial neural network analysis of the oxygen saturation signal enables accurate diagnostics of sleep apnea, Sci. Rep. 9 (2019) 1–9, https://doi.org/10.1038/s41598-019-49330-7.

[34] S.S. Mostafa, F. Mendonca, A.G. Ravelo-Garcia, G. Julia-Serda, F. Morgado-Dias, Multi-objective hyperparameter optimization of convolutional neural network for obstructive sleep apnea detection, IEEE Access 8 (2020) 129586–129599, https://doi.org/10.1109/ACCESS.2020.3009149.

[35] S.S. Mostafa, D. Baptista, A.G. Ravelo-García, G. Juliá-Serdá, F. Morgado-Dias, Greedy based convolutional neural network optimization for detecting apnea, Comput. Methods Progr. Biomed. 197 (2020), 105640, https://doi.org/10.1016/j.cmpb.2020.105640.

[36] A. Leino, S. Nikkonen, S. Kainulainen, H. Korkalainen, J. Töyräs, S. Myllymaa, T. Leppänen, S. Ylä-Herttuala, S. Westeren-Punnonen, A. Muraja-Murro, P. Jäkälä, E. Mervaala, K. Myllymaa, Neural network analysis of nocturnal SpO 2 signal enables easy screening of sleep apnea in patients with acute cerebrovascular disease, Sleep Med. 79 (2021) 71–78, https://doi.org/10.1016/j.sleep.2020.12.032.

[37] S. Biswal, H. Sun, B. Goparaju, M. Brandon Westover, J. Sun, M.T. Bianchi, Expert-level sleep scoring with deep neural networks, J. Am. Med. Inf. Assoc. 25 (2018) 1643–1650, https://doi.org/10.1093/jamia/ocy131.

[38] M. Piorecky, M. Bartoň, V. Koudelka, J. Buskova, J. Koprivova, M. Brunovsky, V. Piorecka, Apnea detection in polysomnographic recordings using machine learning techniques, Diagnostics 11 (2021) 1–21, https://doi.org/10.3390/diagnostics11122302.

[39] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, https://doi.org/10.1038/nature14539.

[40] O. Faust, Y. Hagiwara, T.J. Hong, O.S. Lih, U.R. Acharya, Deep learning for healthcare applications based on physiological signals: a review, Comput. Methods Progr. Biomed. 161 (2018) 1–13, https://doi.org/10.1016/j.cmpb.2018.04.005.

[41] C.L. Marcus, R.H. Moore, C.L. Rosen, B. Giordani, S.L. Garetz, H.G. Taylor, R. B. Mitchell, R. Amin, E.S. Katz, R. Arens, S. Paruthi, H. Muzumdar, D. Gozal, N. H. Thomas, J. Ware, D. Beebe, K. Snyder, L. Elden, R.C. Sprecher, P. Willging,

D. Jones, J.P. Bent, T. Hoban, R.D. Chervin, S.S. Ellenberg, S. Redline, A randomized trial of adenotonsillectomy for childhood sleep apnea, N. Engl. J. Med. 368 (2013) 2366–2376, https://doi.org/10.1056/nejmoa1215881.

[42] S. Redline, R. Amin, D. Beebe, R.D. Chervin, S.L. Garetz, B. Giordani, C.L. Marcus, R.H. Moore, C.L. Rosen, R. Arens, D. Gozal, E.S. Katz, R.B. Mitchell, H. Muzumdar, H.G. Taylor, N. Thomas, S. Ellenberg, The Childhood Adenotonsillectomy Trial (CHAT): rationale, design, and challenges of a randomized controlled trial evaluating a standard surgical procedure in a pediatric population, Sleep 34 (2011) 1509–1517, https://doi.org/10.5665/sleep.1388.

[43] C. Iber, S. Ancoli-Israel, A.L. Chesson, S.F. Quan, The AASM Manual for the Scoring of Sleep and Associated Events: Rules Terminology and Technical Specification, American academy of sleep medicine, Westchester, IL, 2007.

[44] V. Barroso-García, G.C. Gutiérrez-Tobal, L. Kheirandish-Gozal, D. Álvarez, F. Vaquerizo-Villar, P. Núñez, F. del Campo, D. Gozal, R. Hornero, Usefulness of recurrence plots from airflow recordings to aid in paediatric sleep apnoea diagnosis, Comput. Methods Progr. Biomed. 183 (2020), 105083, https://doi.org/10.1016/j.cmpb.2019.105083.

[45] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, 2016. MIT Press.

[46] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: Proc. 32nd Int. Conf. Int. Conf. Mach. Learn. vol. 37, 2015, pp. 448–456. JMLR.org.

[47] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2014. http://arxiv.org/abs/1412.6980.

[48] P.J. Huber, Robust estimation of a location parameter, Ann. Math. Stat. 35 (1964) 73–101, https://doi.org/10.1214/aoms/1177703732.

[49] F. Chollet, Keras, 2015.

[50] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer New York, New York, NY, 2000, https://doi.org/10.1007/978-1-4757-3264-1.

[51] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Third, Morgan Kaufmann/Elsevier, Burlington, 2011, https://doi.org/10.1016/C2009-0-19715-5.

[52] J. Cohen, A coefficient of agreement for nominal scales, Educ. Psychol. Meas. 20 (1960) 37–46, https://doi.org/10.1177/001316446002000104.

[53] J.M. Bland, D.G. Altman, Statistical methods for assessing agreement between two methods of clinical measurement, Lancet 327 (1986) 307–310, https://doi.org/10.1016/S0140-6736(86)90837-8.

[54] N.A. Collop, Scoring variability between polysomnography technologists in different sleep laboratories, Sleep Med. 3 (2002) 43–47, https://doi.org/10.1016/S1389-9457(01)00115-0.