Biomedical Signal Processing and Control 87 (2024) 105490

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

An explainable deep-learning architecture for pediatric sleep apnea identification from overnight airflow and oximetry signals

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 [e], Roberto Hornero [a,b]

[a] Biomedical Engineering Group, University of Valladolid, Valladolid, Spain
[b] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina, Instituto de Salud Carlos III, Valladolid, Spain
[c] Department of Neurology, University of Missouri School of Medicine, Columbia, MO, USA
[d] Sleep-Ventilation Unit, Pneumology Department, Río Hortega University Hospital, Valladolid, Spain
[e] Joan C. Edwards School of Medicine, Marshall University, Huntington, WV, USA

ARTICLE INFO

Keywords:
Obstructive sleep apnea
Children
Airflow
Oximetry
Explainable artificial intelligence
Deep learning

ABSTRACT

Deep-learning algorithms have been proposed to analyze overnight airflow (AF) and oximetry (SpO_2) signals to simplify the diagnosis of pediatric obstructive sleep apnea (OSA), but current algorithms are hardly interpretable. Explainable artificial intelligence (XAI) algorithms can clarify the models-derived predictions on these signals, enhancing their diagnostic trustworthiness. Here, we assess an explainable architecture that combines convolutional and recurrent neural networks (CNN + RNN) to detect pediatric OSA and its severity. AF and SpO_2 were obtained from the Childhood Adenotonsillectomy Trial (CHAT) public database (n = 1,638) and a proprietary database (n = 974). These signals were arranged in 30-min segments and processed by the CNN + RNN architecture to derive the number of apneic events per segment. The apnea-hypopnea index (AHI) was computed from the CNN + RNN-derived estimates and grouped into four OSA severity levels. The Gradient-weighted Class Activation Mapping (Grad-CAM) XAI algorithm was used to identify and interpret novel OSA-related patterns of interest. The AHI regression reached very high agreement (intraclass correlation coefficient > 0.9), while OSA severity classification achieved 4-class accuracies 74.51% and 62.31%, and 4-class Cohen's Kappa 0.6231 and 0.4495, in CHAT and the private datasets, respectively. All diagnostic accuracies on increasing AHI cutoffs (1, 5 and 10 events/h) surpassed 84%. The Grad-CAM heatmaps revealed that the model focuses on sudden AF cessations and SpO_2 drops to detect apneas and hypopneas with desaturations, and often discards patterns of hypopneas linked to arousals. Therefore, an interpretable CNN + RNN model to analyze AF and SpO_2 can be helpful as a diagnostic alternative in symptomatic children at risk of OSA.

1. Introduction

Obstructive Sleep Apnea (OSA) syndrome is a prevalent sleep disorder that affects 1–5% of children worldwide [1,2]. Increased upper airway resistance and intermittent collapsibility during sleep result in respiratory flow pauses (apneas) and decreased airflow (hypopneas), which lead to a fragmented and restless sleep along with gas exchange abnormalities [1,2]. Undiagnosed and untreated OSA is associated with deleterious neurocognitive, developmental, and behavioral effects, as well as cardiovascular and metabolic morbidities [2]. The gold standard approach to diagnose OSA in children is the overnight in-lab polysomnogram (PSG), in which sleep is evaluated using multiple sensors that record neurological, cardiorespiratory, and other biomedical signals [1,3]. These signals are then analyzed to calculate the number of apneas and hypopneas during sleep [3]. The American Academy of Sleep Medicine (AASM) guidelines define apneas as a reduction greater than 90% in the airflow (AF) signal during at least two respiratory cycles, and hypopneas as a reduction greater than 30% in the AF during the same number of cycles followed by a drop of at least 3% in the blood oxygen saturation signal (SpO_2) or an electroencephalographic arousal [3]. Among several indices derived from such analysis of the PSG, the apnea-hypopnea index (AHI) is most frequently the major index used to

* Corresponding author.
E-mail address: jorge.jimenez.garcia@uva.es (J. Jiménez-García).

https://doi.org/10.1016/j.bspc.2023.105490
Received 12 May 2023; Received in revised form 28 August 2023; Accepted 12 September 2023
Available online 29 September 2023

diagnose OSA and establish its severity, and it is defined as the number of apneas and hypopneas per hour of sleep (e/h) [2]. The complexity of the PSG testing and the scarcity of accredited sleep laboratories with expertise in pediatric sleep result in inordinately lengthy waiting lists and therefore lead to a large proportion of children not properly or timely diagnosed with OSA [1]. Accordingly, children with clinical symptoms of OSA would benefit from being assessed with simpler alternatives that could facilitate and speed up their access to treatment [2].

Simplified diagnostic tests have been developed and tested as an alternative to in-lab PSG in the past 20 years. Among these, the most popular approaches focus on nocturnal oximetry or respiratory polygraphy [4,5]. AF and SpO_2 have been commonly evaluated within these approaches due to their simpler acquisition and because these signals are involved in the apnea and hypopnea definitions [3]. An additional way to simplify signal analyses is to employ automatic signal processing algorithms based on machine learning (ML) [6,7]. Most of the previous approaches to detect pediatric OSA focused on ML techniques that relied on comprehensive signal characterization and feature engineering [7], but emergent deep-learning (DL) algorithms eliminate the need of feature extraction as they automatically find relevant patterns in the recorded data [8]. DL methods have been extensively developed and applied in the context of adult sleep apnea detection using different signals [8], but have not been properly translated into the pediatric population [7]. In contrast, common ML methods have been extensively assessed using AF or SpO_2 data, like shallow neural networks [9–13], and ensemble learning [14–16]. These methods have been typically assessed in children regarding to their diagnostic ability in terms of sensitivity (*Se*), specificity (*Sp*), and accuracy (*Acc*) considering different AHI cutoffs. Noticeably, *Acc* increases with the threshold and normally ranges between 75.2% and 90.7% according to previous studies [9,13]. A recent *meta*-analysis reported *Se*-*Sp* pairs of 84.9%-49.9%, 71.4%-83.2%, and 65.2%-93.1% in the respective cutoffs 1, 5, and 10 e/h [7]. Only two recent studies developed and tested DL methods based on Convolutional Neural Networks (CNN) in the pediatric population using AF and/or SpO_2 [17,18], but other architectures have still not been assessed. Regarding DL-based methodologies applied in adult OSA, CNNs and Recurrent Neural Networks (RNN) have been proposed to analyze different respiratory signals including oronasal AF or effort sensors [19–22]. Also, CNNs and deep neural networks (DNN) were selected to process SpO_2 data [23–25], and other studies included diverse signals to detect sleep apnea and/or sleep stages with DL [26–30].

CNNs and RNNs have demonstrated their usefulness in dealing with cardiorespiratory and neurally-derived signals from sleep studies [8]. While CNNs are useful to automatically extract complex patterns from the biomedical signals, RNNs leverage the temporal distribution of the signal information [28,31]. The combination of a CNN with a RNN (CNN + RNN) exploits the benefits of both architectures, with CNNs being used as time-independent feature extractors formed by convolution filters, and RNNs being useful to model the temporal structure of the patterns extracted in the previous layers [28]. This CNN + RNN architecture can be suitable to detect OSA as it can model the recurrence of apneas within normal breathing patterns in respiratory signals. To the best of our knowledge, CNN + RNN models have not been assessed using AF and SpO_2 signals with the aim of detecting pediatric OSA. Only one recent work proposed a 2D CNN to detect OSA in children using both signals, which has demonstrated their usefulness and complementarity [18].

Although the above-mentioned DL algorithms have reached remarkable diagnostic performances none of them provide explanations of their derivation. Indeed, extant algorithms act as black boxes that only provide accurate predictions without a reasoning of what patterns influence in the detection of the disease. This issue can be solved using Explainable Artificial Intelligence (XAI) methodologies, which are aimed at making complex ML/DL models more transparent and interpretable [32]. XAI methods have been scarcely applied in the context of OSA detection, and has relied on feature engineering rather than DL over raw signals [16,33,34]. Explainable DL approaches that relied on a CNN to analyze biomedical images or signals frequently employed the Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm to identify the regions of the inputs that contribute for a certain prediction [35]. In this study, we propose an explainable CNN + RNN algorithm to estimate pediatric OSA severity jointly from AF and SpO_2 that uses Grad-CAM to provide an identification of the characteristics of the signals that drive the model to detect the disease.

We hypothesized that a combination CNN + RNN can leverage information of AF and SpO_2 data to detect pediatric OSA. Moreover, Grad-CAM can contribute to understand the OSA detection process inside the DL architecture by revealing relevant patterns in these two signals. Accordingly, the objective of this study was two-fold: (i) to assess the diagnostic performance of an estimated AHI derived from a CNN + RNN algorithm fed with AF and SpO_2 signals; and (ii) to identify the patterns that contribute to the detection of OSA in these signals. The novel contributions of this article are: (i) the development of a refined DL architecture based on the combination of a CNN and a RNN to detect OSA from overnight AF and SpO_2 data, which has never been tested in pediatric OSA; and (ii) the novel use of Grad-CAM as XAI algorithm to reveal relevant patterns of the signals discovered by the network and used to detect the disease.

2. Subjects and signals

In the present study, a public database and a proprietary database were used. The Childhood Adenotonsillectomy Trial (CHAT) is a publicly available database provided by the National Sleep Research Resource through its repository: https://sleepdata.org/datasets/chat [36,37]. This database contains 1,638 PSGs of children between 5 and 10 years old with OSA symptoms. All the PSG studies included several overnight biomedical signals such as electroencephalogram, electrocardiogram, electromyogram, respiratory movements and AF, pulse-oximetry, body position, etc., and also contained the annotations of the apneic events according to the AASM guidelines [38], enabling us to generate the labels of the signal segments. The recordings were randomly separated into training (60%), validation (20%) and test (20%) sets, ensuring that no statistically significant differences ($p > 0.01$) were present among sets in age, sex, normalized to the age body mass index (BMI z-score), and AHI (Table 1). This division was subject-wise, so individual data were assigned exclusively to one of the sets. The training set was used to train the CNN + RNN model, the validation set was used to optimize the AHI detection algorithm, and the test set was used to evaluate the estimated AHI and interpret the decisions made by the CNN + RNN model.

A proprietary database from the University of Chicago (UofC), with 974 sleep studies from subjects aged 0–13 years with clinical suspicion of OSA, was also analyzed [9,14]. These recordings were used to externally validate and test the proposed AHI estimation algorithm. The Ethics Committee of the UofC approved the study protocol, and the legal caretakers were informed and gave their signed consent (see Ethical Approval section). Children were diagnosed using the PSG according to the current AASM rules [39]. All PSGs contained neuronal, cardiorespiratory, muscular, body position, etc. signals according to the AASM recommendations [38]. The time locations of the apneic events were not provided in the UofC database, so data from this database could not be used to train the CNN + RNN architecture. These data were used only to estimate the total AHI. The subjects in the UofC database were divided into validation (60%) and test (40%) sets, with no statistically significant differences in age, sex, BMI z-score and AHI ($p > 0.01$).

Demographic and clinical data of the subjects that formed the databases of this study, as well as signal characteristics (duration, sampling frequency), are shown in Tables 1 and 2. Some of these variables showed statistically significant differences between the CHAT and UofC

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*)	1,006 (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 z-score	−0.21 [−0.66; 0.49]	−0.28 [−0.66; 0.46]	−0.24 [−0.61; 0.43]	−0.26 [−0.60; 0.47]	−0.17 [−0.58; 0.28]
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[1] (*n*)	219 (21.8%)	69 (21.2%)	96 (16.4%)	67 (21.9%)	75 (19.2%)
Mild OSA[2] (*n*)	496 (49.3%)	168 (51.5%)	229 (39.2%)	148 (48.4%)	169 (43.3%)
Moderate OSA[3] (*n*)	160 (15.9%)	44 (13.5%)	113 (19.4%)	49 (16.0%)	63 (16.2%)
Severe OSA[4] (*n*)	131 (13.0%)	45 (13.8%)	146 (25.0%)	42 (13.7%)	83 (21.3%)
Segments (*n*)	114,873	37,155	58,985	34,771	39,467

Data presented as median [interquartile range] or n (%).
AHI = apnea-hypopnea index; BMI z-score = normalized to the age body mass index; OSA = Obstructive Sleep Apnea; CHAT = Childhood Adenotonsillectomy Trial, UofC = University of Chicago.
[1]: AHI < 1 event/h; [2]: 1 ≤ AHI < 5 events/h; [3]: 5 ≤ AHI < 10 events/h; [4]: AHI ≥ 10 events/h.
(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.

Table 2
Characteristics of the signals included in the CHAT and UofC databases.

	CHAT	UofC
f_s AF (Hz)	20 Hz: 3 (0.18%); 25 Hz: 1 (0.06%); 32 Hz: 368 (22.47%); 50 Hz: 806 (49.21%); 125 Hz: 19 (1.16%); 128 Hz: 35 (2.14%); 200 Hz: 201 (12.27%); 256 Hz: 21 (1.28%); 512 Hz:184 (11.23%);	200 Hz: 674 (69.20%); 500 Hz: 300 (30.80%)
f_s SpO_2 (Hz)	1 Hz: 368 (22.47%); 2 Hz: 401 (24.48%); 10 Hz: 410 (25.03%); 12 Hz: 1 (0.06%); 16 Hz: 35 (2.14%); 125 Hz: 19 (1.16%); 200 Hz: 199 (12.15%); 256 Hz: 21 (1.28%); 512 Hz: 184 (11.23%)	25 Hz:297 (30.49%); 200 Hz: 525 (53.90%); 500 Hz: 152 (15.61%)
Duration (minutes)	586.39 [546.30; 645.97]	532.86 [497.10; 568.22]

Data presented as median [interquartile range] or n (%).
AF = airflow signal, CHAT = Childhood Adenotonsillectomy Trial, f_s = sampling frequency, SpO_2 = oximetry signal, UofC = University of Chicago.

databases, as shown in Table 1. We dealt with this heterogeneity by forming a joint validation dataset of 910 subjects from the CHAT and UofC validation sets. This dataset was used to obtain the optimum hyperparameter configuration of the CNN + RNN algorithm. The test sets of both databases were independently used to assess the diagnostic performance of the algorithm.

AF signals from PSG were recorded using an oronasal thermistor at sampling frequencies (f_s) ranging 20–512 Hz in the CHAT database and 200–500 Hz in the UofC database (Table 2), whereas SpO_2 signals were registered with a photoplethysmography-based pulse oximeter finger probe with f_s in the range 1–512 Hz in the CHAT database and 25–500 Hz in the UofC database (Table 2). The duration of the PSG-derived AF and SpO_2 signals were the same for each subject. An example of these signals is shown in Fig. 1, in which apneas and hypopneas are visible in the AF waveform and their respective desaturations can be observed in the SpO_2 profile.

3. Methods

An interpretable CNN + RNN model was developed in this study from AF and SpO_2 data with the aim of predicting OSA presence and severity in children. The entire architecture is shown in Fig. 2. The CNN + RNN model was fed with 30-min segments of preprocessed AF and SpO_2, which were divided into six 5-minute epochs. Each epoch was processed in the CNN blocks to form time-independent feature maps. The sequences of CNN-derived features were then analyzed in the RNN to estimate the number of apneic events present in the segment. Finally, the Grad-CAM algorithm was applied to locate the periods of time in which the model focused to predict the presence of apneic events.

3.1. Signal preprocessing and segmentation

Overnight AF and SpO_2 signals were obtained from the PSG and were preprocessed with resampling and amplitude normalization. Resampling was first applied to set a common f_s of 10 Hz, which reduces computational cost while preserving the information in both signals. This f_s was selected considering that the spectral components contained above $f_s/2$ are negligible in both signals according to the Nyquist-Shannon theorem. Moreover, the CNN blocks included in the proposed architecture were optimized in a previous study using $f_s = 10$ Hz [18]. The AF was additionally filtered with a Kaiser window low pass filter to reduce noise and preserve the respiratory oscillations. The cutoff frequency of the filter was 1.5 Hz, with a minimum stopband attenuation of 100 dB beyond 2 Hz [12,18]. AF amplitude was normalized adaptively like in previous studies using the pre-processing algorithm proposed by Varady et al. [12,13,40]. This method corrects the baseline and adjusts the scale of the AF signal segment by segment by subtracting the baseline and dividing the result by the scale [40]. Finally, both signals were standardized to have zero mean and unit standard deviation.

The signals were arranged into segments of 30-min (18,000 samples × 2 signals), and they were subsequently divided into 6 epochs of 5-min duration (3,000 samples) to adapt them to the shape of the sequences allowed by the CNN + RNN architecture. Therefore, the shape of each segment was 6 × 3,000 × 2. The duration of the segments and the epochs was selected to cover the typical duration of large clusters of apneic events and desaturations that recurrently appear in the polysomnographic records [41], as previously addressed in a study focused on the analysis of SpO_2 data [17]. After preliminary tests, we set the duration of the segments to 30 min since that duration optimized the regression problem, and the duration of the epochs was fixed to 5 min, which is suitable for the analysis of AF and SpO_2 using a CNN [18]. In addition, 25-min overlapping segments were used to perform data augmentation during training and validation. This also allowed us to feed the CNN + RNN model with segments in which the relevant epochs can be placed in any position within the sequence, thus avoiding potential bias towards certain epochs. All segments from the CHAT database were labeled with

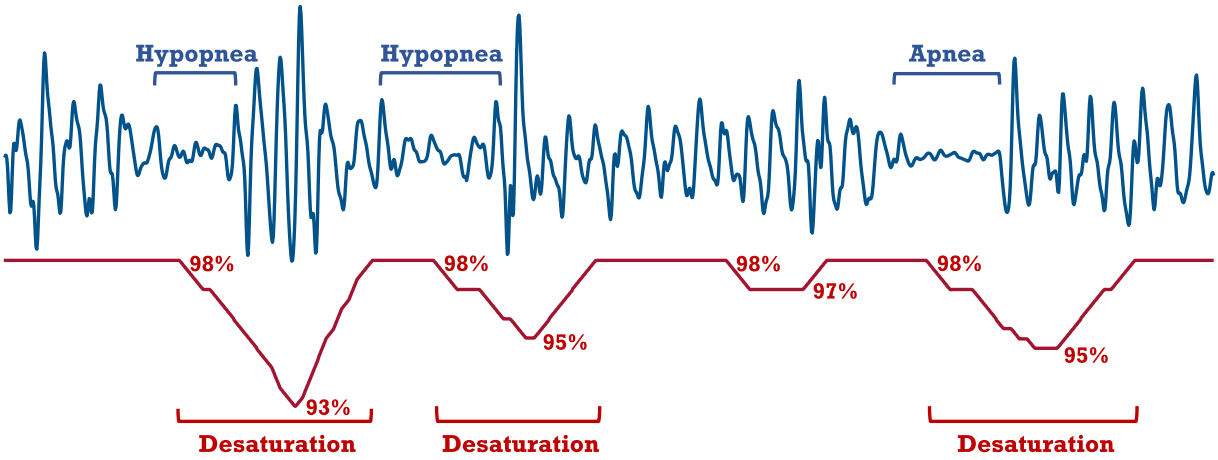

Fig. 1. Airflow (top) and oximetry (bottom) signals with apneas, hypopneas, and their respective desaturations.

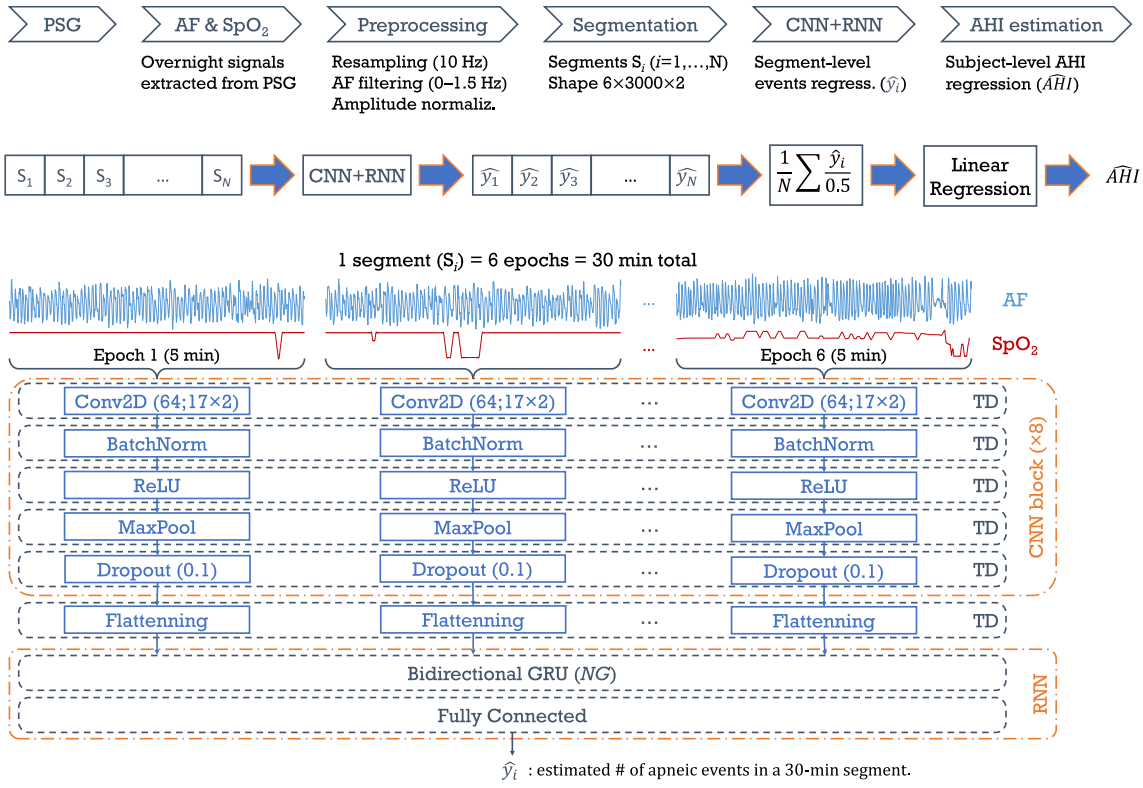

Fig. 2. Overview of the proposed methodology based on a deep-learning architecture combining convolutional and recurrent neural networks (CNN + RNN): Preprocessing of overnight airflow (AF) and oximetry (SpO_2) signals from polysomnography (PSG), segmentation into sequences of six 5-min epochs (30 min total), and estimation of the apnea-hypopnea index (AHI) through the CNN + RNN algorithm. The CNN is a stack of time distributed (TD) layers that encapsulate 2D convolutional (Conv2D), batch normalization (BatchNorm), rectified linear unit (ReLU), max pooling (MaxPool), dropout and flattening layers. Bidirectional gated recurrent unit (GRU) and fully connected layers formed the RNN. The final estimation is $\hat{y}_i$, the total number of apneic events in the segment. The estimations of all segments are used to calculate the final AHI of each subject through a linear regression model.

the number of apneic or hypopneic events that begin and end within the 30-min segment, which were used to train and validate the CNN + RNN model [18,24].

3.2. CNN + RNN architecture

A combination of CNN and RNN models was developed and trained using a dataset of 30-min segments of AF and SpO_2 signals, labeled with the number of apneas and hypopneas and distributed in six 5-min epochs. The convolutional part of the architecture was implemented as a stack of Time Distributed (TD) layers which encapsulated the layers of a previously presented CNN trained with AF and SpO_2 data [18]. The TD layers work with each epoch of the input segment independently to return a sequence of processed data with the same length (6 epochs). A total of 8 blocks of 5 consecutive CNN layers -Convolutional 2D, Batch Normalization, Rectified Linear Unit (ReLU) activation, Max Pooling, and Dropout- embedded into TD layers were arranged with the aim of learning the AF and SpO_2 features from each epoch related with apneic events [17]. The architecture is depicted in Fig. 2. Each convolutional layer generated 3D feature maps from the epochs using the 2D

convolution operation [42]:

$$x_i^j[m,n] = \sum_{k=1}^{17}\sum_{l=1}^{2} w_i^j[k,l]\cdot a_i[m-k+1, n-l+1] + b_i^j \quad (1)$$

where x_i^j is the feature map generated in the convolutional block i ($i = 1$, …, 8), with the filter with weights w_i^j and bias b_i^j ($j = 1$, …, 64) and a_i as the input to the i-th convolutional block. Each convolutional layer was composed of 64 2D filters with kernel size 17 × 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 feature maps generated by the previous layer [42]. The standard ReLU activation was then applied [42]:

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

where x_i^j is the value of each sample of the feature map. Dimensionality reduction was applied to the activations using a max pooling layer 2 × 1 to halve the length of the feature maps while the width and depth are kept. The last layer of the convolutional blocks was a dropout layer that randomly removed 10% of the activations ($P = 0.1$) during each training batch to reduce overfitting [17,18].

Next to the TD layers of the 8 convolutional blocks, three alternatives were studied to connect the output of the last CNN block, i.e., a sequence of six feature maps (4D tensor, shape 6 × 11 × 2 × 64), to the input of the RNN block (2D tensor, first dimension equal to 6): (i) a Global Average Pooling (GAP) layer inside a TD layer that calculates the mean of each channel of the feature maps, resulting in a sequence of six vectors of size 64 [28,43]; (ii) a flattening operation wrapped into a TD layer to reshape the feature maps into a sequence of six vectors of 1,408 elements; (iii) flattening followed by a fully connected layer (both encapsulated into a TD) that derive one value per sequence element [18]. The resulting sequence is then processed with a Bidirectional Gated Recurrent Unit (Bi-GRU) layer to analyze the temporal distribution of the features extracted in the CNN throughout the sequence. This layer is the main part of the RNN, which analyzes the temporal patterns of the data in both directions [42]. The GRU recurrent units were selected due to their simplicity and lower computational cost, reaching nearly the same performance compared with LSTM [44]. The number of units of the Bi-GRU (*NG*) defines the dimensionality of the output and was optimized in this study. We varied *NG* in the range {1, 2, 4, …, 64} to find the best performing configuration. The recurrent dropout and the dropout rates of the GRU layer were both empirically set to 0.1 after preliminary tests [17,18]. A fully connected layer with a linear activation unit was finally implemented to obtain the prediction of the number of apneas/hypopneas in each 30-min segment.

The proposed CNN + RNN architecture is an improved version of the CNN model developed and validated in our previous work [18]. The optimum architecture and weights of the CNN were transferred to the CNN + RNN model, so the training and optimization were carried out as a transfer-learning approach: the weights of the pretrained CNN blocks were fixed and only the RNN part was trained from scratch. Therefore, most of the structural hyperparameters of the CNN were inherited and not changed. During training, the adaptive momentum estimation (Adam) algorithm was used to optimize the model, using an initial learning rate of 10^{-3} and the default momentum-related parameters $\beta 1 = 0.9$ and $\beta 2 = 0.999$ [45]. Like in previous studies, the Huber loss with delta (δ) parameter fixed to $\delta = 1$ was employed in the Adam optimization, that has shown its robustness in regression with large outliers [46]. The validation data was used during the training process to supervise the convergence of the CNN optimization by calculating a validation loss. Additional callbacks were implemented to control the convergence of training using the validation data. The learning rate was reduced by a factor of 2 during training when the validation loss stopped decreasing during 10 epochs [42,47]. Early stopping was also implemented to avoid overfitting. If the validation loss did not improve during the 30 epochs after reaching its minimum, the training was stopped and the weights were restored to those obtained in the epoch with the best validation loss [42,47].

Once the optimized DL model was trained and validated, it was applied to the overnight segmented signals to derive an estimation of the AHI for each subject. An estimation of the number of apneic events in each 30-min segment (y_i) was obtained and divided by the segment duration (0.5 h), and the mean rate of apneic events per hour throughout the recording (total: N segments) formed a primary prediction of the AHI:

$$AHI_{PR} = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i}{0.5} \quad (3)$$

AHI_{PR} may underestimate the actual AHI because it uses all the available segments instead of the segments in which the subject was sleeping [4]. Therefore, a linear regression model was implemented to correct this bias. This linear regression was aimed at finding the optimum coefficients of a linear equation that estimates the final AHI from AHI_{PR}.

3.3. Model explainability using Grad-CAM

The CNN + RNN model was analyzed using Grad-CAM, a XAI method that generates *post-hoc* explanations using the gradients of the target into the convolutional layers of a CNN-based model [32,48]. Class Activation Mapping (CAM) was originally proposed using image classification-aimed CNN architectures using a GAP layer before the final classification layer. CAM aims to generate attribution maps that locate the discriminative regions of the input that lead to a certain classification [32,48]. Grad-CAM is a generalization of the original CAM algorithm that uses the gradients of the convolutional layers to identify the most sensitive samples of the input that influence the final prediction of the network. These gradient-based attribution maps are the heatmaps, which can be obtained from every convolutional layer in the model [48]. Firstly, the gradients of the model output $\hat{y}$ with respect to the feature maps of the i-th convolutional layer x_i^j are computed and averaged through the number of maps [48]:

$$a_i = \frac{1}{N}\sum_j \frac{\partial \hat{y}}{\partial x_{i_j}^j} \quad (4)$$

where $N = 64$ is the number of feature maps (filters) in the i-th layer. The heatmaps are then obtained as a gradient-weighted combination of the feature maps after a ReLU activation [48]:

$$L_{GradCAM} = ReLU\left(\sum_i a_i \cdot x_i^j\right) \quad (5)$$

The Grad-CAM heatmaps have the same size as the input to the i-th convolutional layer, so in each layer they have different lengths. We resized and averaged the Grad-CAM heatmaps in all layers to obtain heatmaps of the same size as the input segments in this study, as this provided us a better representation of the contribution of all convolutional layers. The Grad-CAM heatmaps are stronger in the zones that increase the final prediction, so they point to zones with apneas/hypopneas and normal breathing zones in which subtle changes may contribute to detect potential apneic events.

3.4. Model optimization and diagnostic performance

The proposed CNN + RNN architecture was aimed at predicting the total AHI from overnight signals. In order to evaluate the diagnostic performance, the subject-wise AHI was used to classify the severity of OSA into four childhood-specific 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]. The 4-class Cohen's Kappa (κ) coefficient was used

to find the optimal hyperparameters of the model in the validation set. The Cohen's κ is useful in unbalanced multiclass classification tasks, because it is less biased towards the majority class than the 4-class accuracy (Acc_4), which computes the rate of correct predictions regardless the distribution of classes [49].

The agreement between the DL-derived AHI and the reference AHI from the manually scored PSG was evaluated in the test set using Bland-Altman plots and the intraclass correlation coefficient (*ICC*) [50]. The classification into the four severity levels was assessed in the test set using confusion matrices, Acc_4, and κ [49]. The diagnostic ability of the algorithm was also evaluated in increased severity AHI-based cutoffs (1, 5 and 10 e/h) by means of sensitivity (*Se*), specificity (*Sp*), accuracy (*Acc*), positive and negative predictive values (*PPV*, *NPV*), and positive and negative likelihood ratios ($LR+$, $LR-$).

4. Results

4.1. CNN + RNN model optimization and diagnostic ability

Fig. 3 shows the Cohen's κ obtained in the validation set using different connections of the CNN with the RNN and with varying values of *NG*. The maximum performance in the validation set was $\kappa = 0.5077$ with $NG = 4$ and using a TD flattening layer between the convolutional blocks and the Bi-GRU. The performance of other configurations was slightly lower, so this optimum model was finally selected to evaluate the test data.

The regression model based on the optimum CNN + RNN was applied to estimate the AHI of the subjects in the test set. The scatter plots in Fig. 4, as well as the Bland-Altman plots in Fig. 5 show the deviation of the AHI estimates with respect to the manually scored AHI. The agreement between the actual and the estimated AHI was $ICC = 0.9465$ in the CHAT test set and $ICC = 0.9004$ in the UofC test set. The mean error (bias) was below 1 e/h in both databases and the dispersion of the error was lower in the CHAT test set than in the UofC test set, which is consistent with the subject classification results obtained in both databases. Fig. 6 shows the confusion matrices obtained from the estimated AHI, by assigning each value to one out of the 4 OSA classes (no OSA, mild, moderate, and severe OSA). The 4-class metrics obtained in the test sets were $Acc_4 = 74.51\%$, $\kappa = 0.6231$ in the CHAT database, and $Acc_4 = 62.31\%$, $\kappa = 0.4495$ in the UofC database. The slight tendency to underestimate the AHI shown in the CHAT database was also present in the confusion matrix obtained in this dataset. On the contrary, the proportion of subjects with an overestimated OSA severity was higher in the UofC data. The diagnostic performances of the proposed algorithm in the common AHI cutoffs for mild, moderate, and severe OSA are shown in Table 3. In general, the accuracies surpassed 84% in the three AHI cutoffs of both databases and remained higher in the CHAT data.

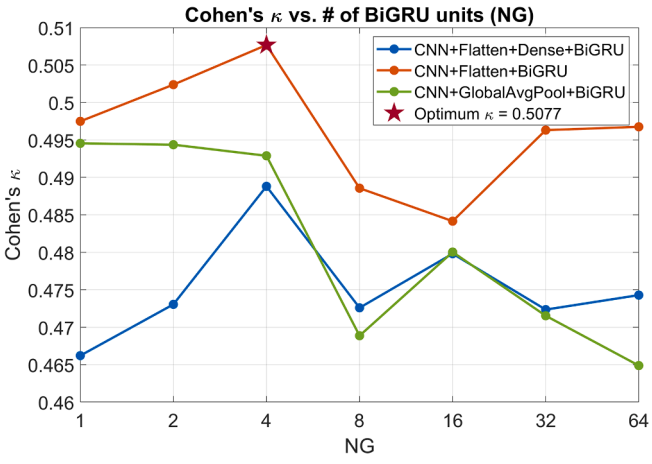

Fig. 3. Diagnostic performance of the convolutional and recurrent neural network (CNN + RNN) architecture for varying number of units in the Bidirectional GRU (Bi-GRU) layer (*NG*) and different connections of the CNN with the RNN in the validation set.

4.2. Gradient-based explanations using Grad-CAM

The final CNN + RNN model was assessed using Grad-CAM explanations to identify the patterns that led the algorithm to detect OSA events in the AF and SpO_2 signals. The heatmaps in Fig. 7 show examples of patterns that drive the model to accurately predict the number of apneic events, together with a zoom in a relevant region of the heatmap and the PSG-derived annotations of the apneic events. In the segment without apneas/hypopneas in Fig. 7 (a), the Grad-CAM heatmap indicate zones in which the algorithm could potentially have detected an apnea or hypopnea in the AF without desaturation, but both the gradients and the estimation were low. Fig. 7 (b) shows an accurately predicted segment with a single apnea, in which the zoom in the Grad-CAM heatmap points both the apnea in the AF and a consecutive desaturation in the SpO_2. The example of Fig. 7 (c) shows the Grad-CAM heatmaps of a segment with an apnea and a hypopnea, both linked to SpO_2 desaturations. The zoom in the heatmap in Fig. 7 (d) point to a group of consecutive hypopneas accompanied by desaturations and a PSG-scored arousal.

Fig. 8 shows the heatmaps of some observed inaccurate predictions accompanied with a zoom of a region where some mistakes were observed. The segment in Fig. 8 (a) shows consecutive hypopneas which are not strongly highlighted in the AF, leading to underestimation. In this case, the algorithm could not identify the hypopneas not linked to substantial desaturations. In Fig. 8 (b), the model was unable to indicate some consecutive hypopneas in the AF without evident desaturations but associated with arousals according to the PSG annotations. The artifacts in both signals and a desaturation in Fig. 8 (c) drove the model to predict more than one apneic event, one of those with a desaturation located between two periods of SpO_2 signal loss and probably after a sudden movement that worsened the AF signal quality. Finally, Fig. 8 (d) shows an example of a segment with three desaturations correctly highlighted, but not accompanied with evident AF reductions and therefore not scored as hypopneas.

5. Discussion

In this study, we have developed a DL model with a remarkable diagnostic performance that also is interpretable. This is the first study that proposes an explainable DL algorithm that not only reaches accurate diagnosis of pediatric OSA, but also provide an identification of the patterns that lead the algorithm to predict the presence of apneic events over AF and SpO_2 signals.

5.1. CNN + RNN architecture

This is the first time that a novel CNN + RNN architecture is successfully tested to assess pediatric OSA, since the previous studies only relied on CNNs [17,18]. The combination of CNN and RNN models has been previously tested to score sleep stages from electroencephalogram (EEG) and photoplethysmography signals [28,43], and is also frequent in the field of adult OSA using different cardiorespiratory signals [26,31,43]. The results of this study confirm that CNN + RNN can be also applied to analyze AF and SpO_2 data. Our architecture specifically focused on predicting the number of respiratory events in 30-min signal segments and then estimating the global AHI of each subject using these predictions. The epoch length (5 min) was suitable for the processing in the TD CNN blocks, as demonstrated in our previous work [18]. By selecting this epoch length, we could also transfer the optimized layers of the previous CNN to the TD CNN proposed in this study and train the CNN + RNN model using transfer learning, which is another novelty of

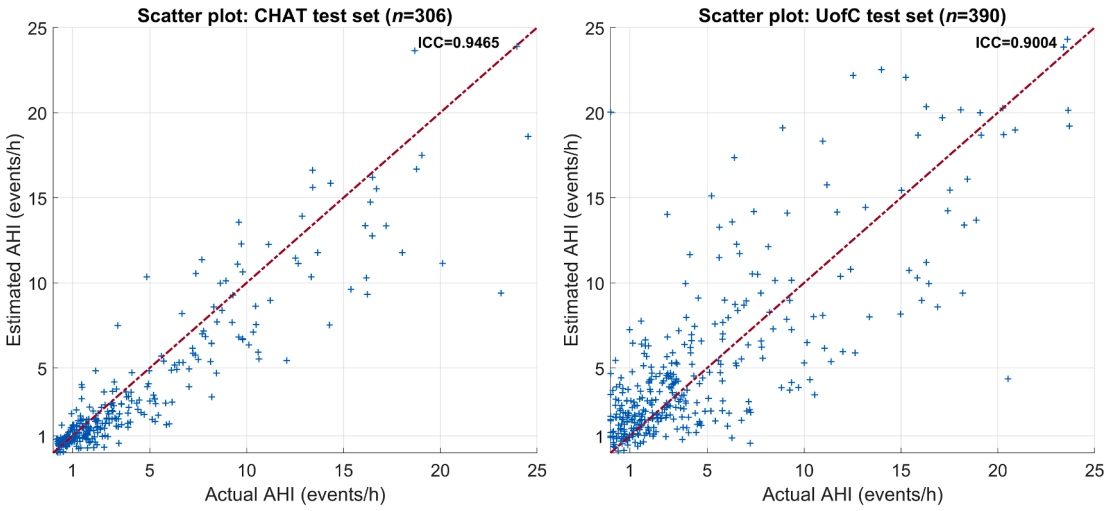

Fig. 4. Scatter plots of actual and estimated apnea-hypopnea index (AHI) in CHAT and UofC test sets.

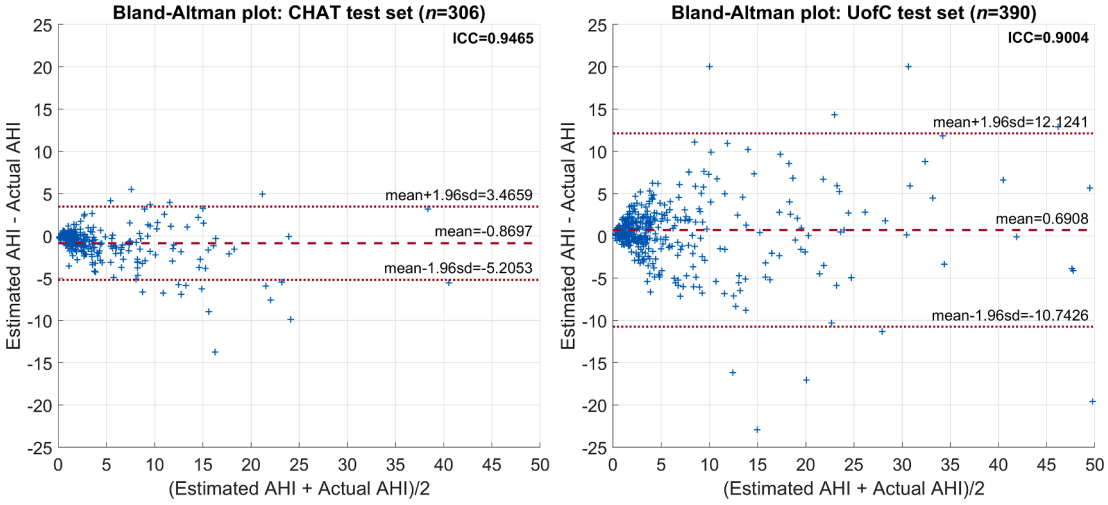

Fig. 5. Bland-Altman plots of actual and estimated apnea-hypopnea index (AHI) in CHAT and UofC test sets.

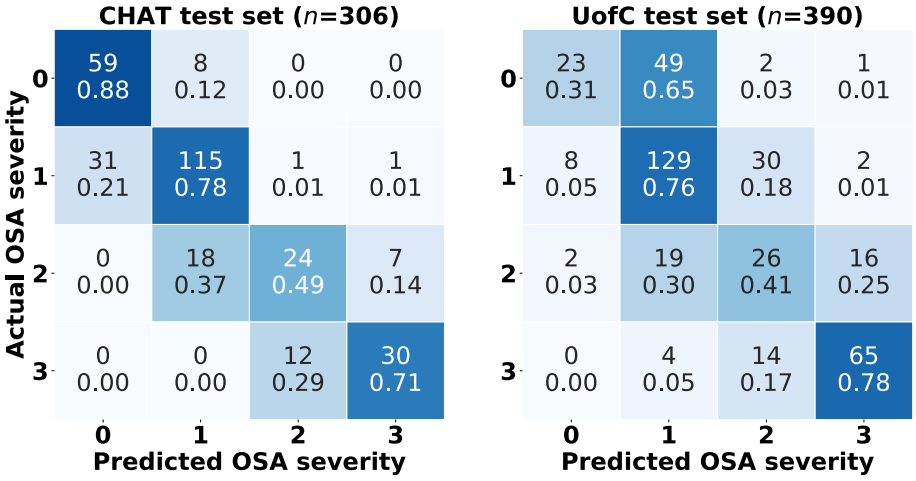

Fig. 6. Confusion matrices of the predicted obstructive sleep apnea (OSA) severity against the actual OSA severity in CHAT and UofC test sets. 0: no OSA; 1: mild OSA; 2: moderate OSA; 3: severe OSA.

Table 3
Diagnostic ability of the model 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**	87.03	88.06	87.25	96.30	65.56	7.2887	0.1473
	UofC	96.83	30.67	84.10	85.43	69.70	1.3965	0.1035
5 e/h	**CHAT**	80.22	99.07	93.46	97.33	92.21	86.2363	0.1997
	UofC	82.88	85.66	84.62	77.56	89.32	5.7777	0.1999
10 e/h	**CHAT**	71.43	96.97	93.46	78.95	95.52	23.5714	0.2946
	UofC	78.31	93.81	90.51	77.38	94.12	12.6538	0.2312

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

the study. This approach allowed us not only to simplify the training and validation of a deep architecture of more than 40 layers, but also to leverage the pattern recognition ability of the CNN that had been optimized in past studies. In addition, the sequences of six CNN-processed epochs were analyzed in the RNN using a Bi-GRU layer, which allowed us to model the recurrence of respiratory events concentrated in long-time clusters. The best configuration of the Bi-GRU layer and the optimal data representation of the TD CNN were accomplished by inserting a TD flattening layer after the CNN blocks (Fig. 3), which leverages the entire feature maps obtained in the CNN without dimensionality reduction before the GRU. In addition, the optimal $NG = 4$ was low, suggesting that modeling the recurrence of apneas/hypopneas in a large cluster using 5-min steps does not require high complexity. For the sake of completeness, the architecture obtained using a transfer learning approach was further optimized with fine tuning (i.e., by unlocking all layers of the base CNN model and training again), allowing all the layers to be trainable again after applying transfer learning. However, no improvements were observed in terms of diagnostic ability. Optimization of other hyperparameters did not produce higher performance in the validation data, so the proposed configuration seems to be nearly optimal.

5.2. Grad-CAM explanations of the model

The second source of novelty of this study is the application of XAI to clarify the mechanisms that produce the model predictions. To our knowledge, only one study employed XAI analysis in pediatric OSA detection [16]. However, the methodology applied in that study relied on feature-engineering rather than DL, applying Shapley Additive Explanations (SHAP) values to demographic, anthropometric, and heart rate-derived variables, together with the oxygen desaturation index. In our case, we focused on generating localization maps over AF and SpO_2 signals using Grad-CAM, thus allowing the identification of the most important parts of the signals for pediatric OSA detection. This technique has not been applied in the context of sleep apnea, and recent studies have only applied that to identify sleep stages from a single-channel EEG [51,52].

Figs. 7 and 8 show some cases in which the heatmaps generated using Grad-CAM highlight the most sensitive zones, where an apneic event could be present. Concretely, we can notice that the regions in which the AF amplitude suddenly changes, the algorithm tends to be more sensitive (Fig. 7 (a)). It is also noticeable that missed breaths are also highlighted (Fig. 7 (a), (b), (d)), but the algorithm also examines drops in the oxygen saturation. In general, the heatmap of SpO_2 is stronger in the case of detecting hypopneas with desaturations (Fig. 7 (c) and 8 (a)), which are the frequent respiratory abnormality in children. This is consistent with the definition of apneas and hypopneas, in which partial reduction of AF should be followed by a desaturation or an arousal to be considered a hypopnea. As hypopneas are often not easily visible in the AF signal, the associated desaturations aid in the detection of these events (Fig. 8 (a), (d)), which could help sleep technicians to review and/or improve manual scoring. In this case, the heatmaps are stronger in the SpO_2 pattern, indicating that the SpO_2 signal complements the AF when the respiratory flow reductions are not clear. Interestingly, the heatmaps of SpO_2 also highlight flat zones in the SpO_2 that correspond to normal oxygenation (Fig. 7 (a) and 8 (b)), suggesting that the model links these flat zones with zero apneas/hypopneas. However, the interpretation could be the opposite: a small variation in the SpO_2 signal could trigger the detection of an apneic event because the AF heatmap is highlighting possible amplitude variations or missed breaths (Fig. 7 (a)). Large prediction errors are also possible, and in these cases Grad-CAM can provide clues to understand why the model prediction failed. For example, artifacts due to signal loss or movements are highlighted in the AF heatmap in Fig. 8 (c). In this case, a sudden artifact may have influenced the prediction of more apneic events than were actually scored in the segment. Signal loss in the SpO_2 was also highlighted, probably indicating that a hypopnea has not been scored but was detected due to border effects. Nevertheless, artifacts do not frequently influence the detection of respiratory events, and it was observed that the model is more accurate when dealing with normal breathing segments with no apneas or hypopneas.

5.3. Diagnostic ability and comparison with previous studies

The proposed CNN + RNN algorithm clearly surpasses previous approaches in terms of diagnostic performance in the CHAT and the UofC test sets. The diagnostic ability of previous studies focused on pediatric OSA is represented in Table 4. Our study involved 2,612 pediatric subjects from two different databases, which is one of the largest cohorts to date in the literature. The study of Hornero et al. involved a multicentric database of 4,191 children [9], and Ye et al. employed a database of 3,139 subjects [16]. Regarding 4-class classification of OSA severity, our DL approach surpassed previous ML methods that relied on AF and SpO_2 and obtained $Acc_4 < 60\%$ and $\kappa < 0.4100$ in the UofC database [12–14]. In addition, the accuracies reached in this study are also higher to those reached in previous CNN-based architectures focused on pediatric OSA [17,18]. According to the extant literature, the diagnostic ability of ML-based methods to predict OSA in children increases proportionally to the AHI threshold employed to differentiate OSA, as it is frequently easier to detect severe OSA than to distinguish between healthy subjects and mild OSA patients [7]. The results achieved by our algorithm also showed that tendency, with the lowest *Acc* obtained in the most restrictive cutoff and the highest *Acc* to diagnose the most severe condition (Table 3). Our study obtained the highest *Acc* in 1 e/h among the studies that employed the CHAT and/or UofC databases, together with remarkable *NPV* and *LR-* in this cutoff in both test sets. This indicates that the proposed model is valuable to discard the presence of OSA with the most restrictive criteria. Regarding the results of the study of Ye et al., they reached very high $Acc = 90.45\%$ and $Sp = 100\%$ in 1 e/h, but the test database was limited to 12 healthy children in a total of 628 subjects [16]. Garde et al. obtained more balanced *Se* and *Sp* using a dataset of 207 pediatric subjects, but $Acc = 75\%$ was among the lowest [53]. Our previous 2D CNN reached lower *Acc* and more unbalanced *Se* and *Sp* in the CHAT database [18]. The *Sp* in 1 e/h was much lower than *Se* in 1 e/h in comparison with previous works that also employed the UofC database, due to the slight tendency of the network to overestimate the AHI in this

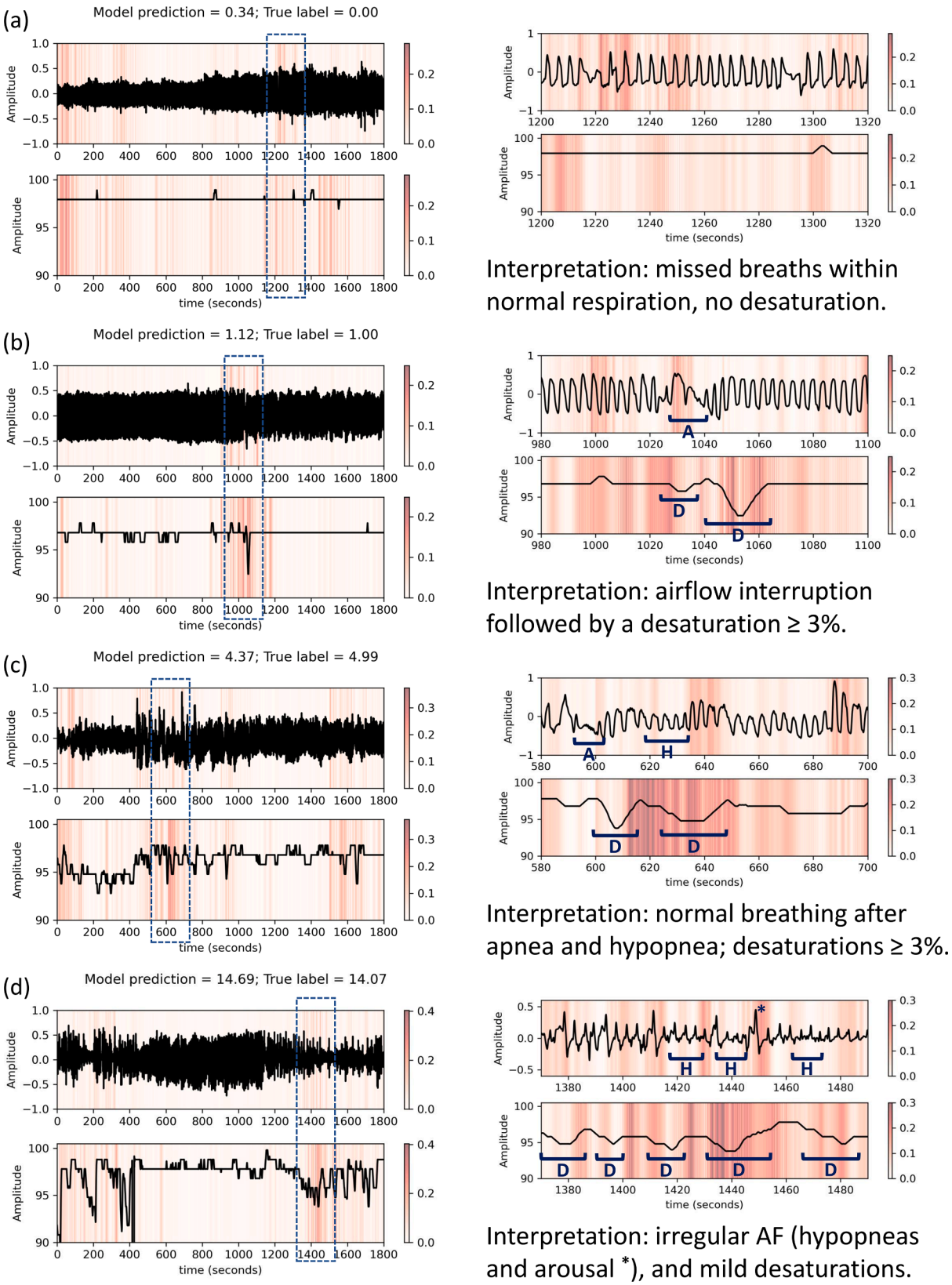

Fig. 7. Grad-CAM heatmaps generated from different segments with their corresponding estimations: (a) segment without apneas; (b) segment with an apnea [A] and desaturations [D] in which heatmaps point both the AF interruption and the SpO_2 desaturation; (c) segment with an apnea [A] followed by a hypopnea [H], both associated with desaturations [D]; (d) segment with consecutive hypopneas [H] followed by oxygen desaturations [D] lower than 3% and an arousal [*].

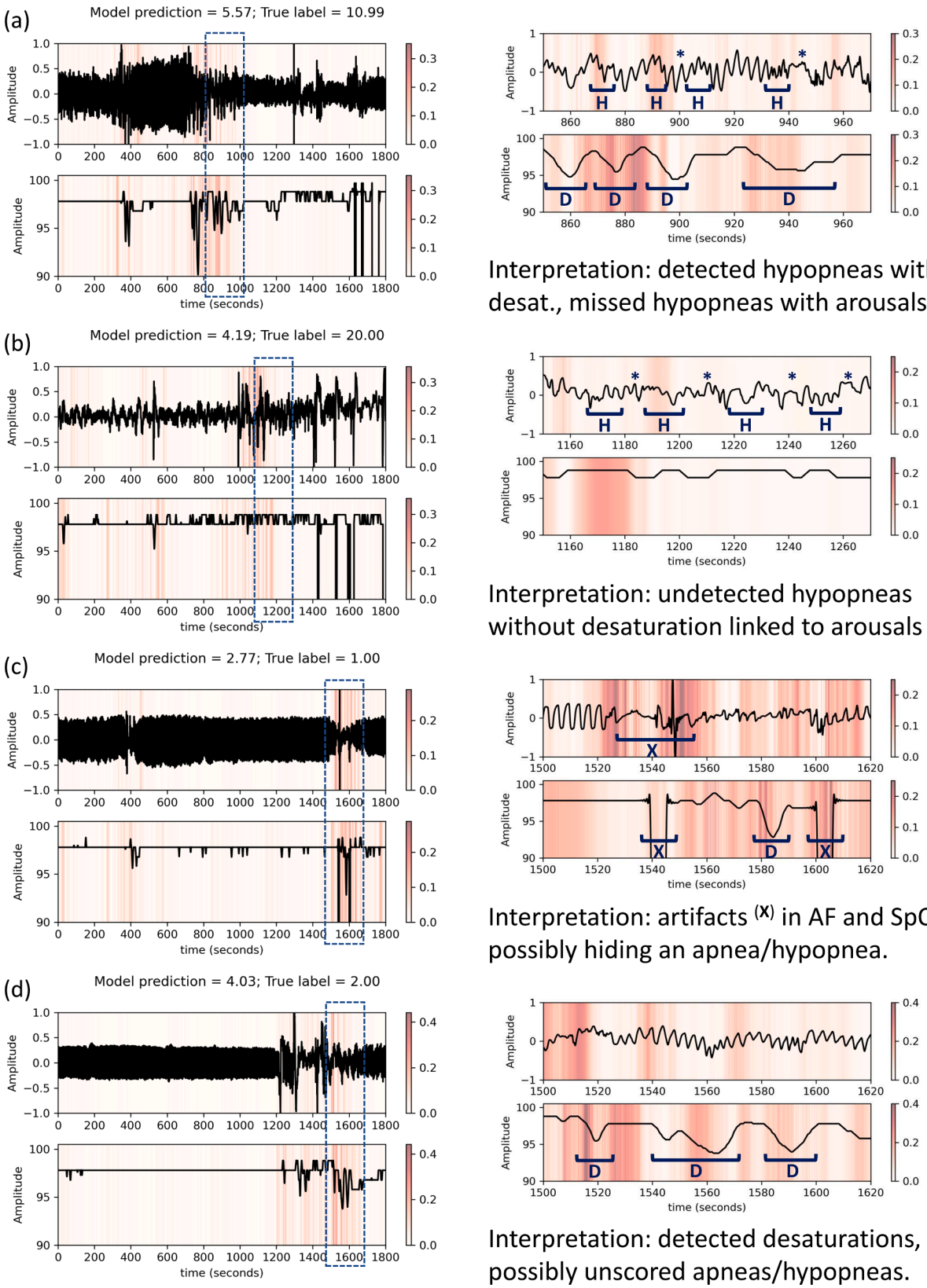

Fig. 8. Grad-CAM heatmap generated from segments with inaccurate predictions: (a) consecutive hypopneas [H] associated to desaturations [D] and arousals [*], in which the latter were not identified; (b) a group of successive hypopneas [H] associated to arousals [*] that were not highlighted; (c) overestimation caused by artifacts and signal loss [X] in both signals near to a possible apnea/hypopnea associated to a desaturation [D] that was not scored; (d) three consecutive desaturations [D] in which a previous hypopnea was not clearly visible in the AF and therefore not scored by the expert.

Table 4
Diagnostic performance of state-of-the-art approaches in the childhood OSA context.

Study	Signal	Methods: Extraction / Selection / Classification / XAI / Validation	N° Subjects	Cutoff (events/h)	Se (%)	Sp (%)	Acc (%)
Ye *et al.* (2023) [16]	SpO_2, HR	Sociodemographic, anthropometric, ODI, mean and max. HR / - / XGBoost / SHAP/Holdout	3,139	1	90.3	100.0	90.4
				5	82.1	93.8	85.7
				10	84.8	92.1	89.8
Calderón *et al.* (2020) [15]	SpO_2	Oxygen desaturations, ODI / - / LR, AdaBoost / - / 15-fold cross validation	453 (CHAT)	5	62.0	96.0	79.0
Hornero *et al.* (2017) [9]	SpO_2	Time statistics, spectral, nonlinear, ODI / FCBF / MLP regression / - / Holdout	4,191	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) [10]	SpO_2	Time statistics, spectral, nonlinear, ODI / FCBF / MLP regression / - / External validation	432	1	95.3	19.1	79.6
				5	77.8	80.5	79.4
				10	73.5	92.7	88.2
Garde et al. (2019) [53]	SpO_2, PRV	Time statistics, spectral, ODI (SpO_2), spectral (PRV) / Stepwise LR / Binary LR (for each cutoff) / - / Holdout	207	1	80.0	65.0	75.0
				5	85.0	79.0	82.0
				10	82.0	91.0	89.0
Barroso-García (2021a) [12]	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) [13]	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) [17]	SpO_2	- / - / CNN / - / Holdout	1,638 (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
Jiménez-García *et al.* (2022) [18]	AF, SpO_2	- / - / 2D 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
This study	**AF, SpO_2**	**- / - / CNN + RNN / Grad-CAM / Holdout**	**1,638 (CHAT)**	**1**	**87.0**	**88.1**	**87.3**
				5	**80.2**	**99.1**	**93.5**
				10	**71.4**	**97.0**	**93.5**
			974 (UofC)	**1**	**96.8**	**30.7**	**84.1**
				5	**82.9**	**85.7**	**84.6**
				10	**78.3**	**93.8**	**90.5**

Acc = Accuracy; AF = Airflow signal; ANN = Artificial Neural Network; BY-MLP = Multilayer perceptron neural network with Bayesian approach; CHAT = Childhood Adenotonsillectomy Trial; CNN = Convolutional neural network; FCBF = Fast correlation-based filter; Grad-CAM = Gradient-weighted Class Activation Mapping; HR = Heart rate; 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; SHAP = Shapley Additive Explanations; Sp = Specificity; SpO_2 = Oxygen saturation signal; UofC = University of Chicago; XAI = Explainable Artificial Intelligence.

dataset. Other studies that employed the UofC database also overestimated the AHI and obtained low *Sp* in 1 e/h [12,14,17]. This might indicate that these algorithms are prone to identify mild OSA patterns in healthy subjects of the UofC test database and therefore misclassify them as having mild OSA, an issue that actually may be inconsequential when considering the current clinical practice underlying the management of symptomatic children. This tendency was not observed in the CHAT dataset, maybe indicating that inter-scorer variability seriously affects the diagnostic performance of the ML/DL algorithms optimized with these databases [54]. This highlights the need for additional studies with a wide range of pediatric sleep datasets to further improve generalizability. To mitigate the difference between CHAT and UofC databases in terms of this inter-scorer variability, the validation set comprised subjects from both databases, resulting in high accuracies in the two of them compared with the literature. Our CNN + RNN approach obtained similar performance in the CHAT database in both 5 and 10 e/h with respect to the previous CNN-based approach [18]. Likewise, all the diagnostic metrics in 5 and 10 e/h improved in the UofC test set, in which our model reached *Acc* and *Se* close to the highest in all cutoffs. This might indicate that the proposed CNN + RNN architecture suits better analyzing long sequences with potential clusters of consecutive apneas/hypopneas and desaturations. Moreover, *Se*, *Sp*, and *Acc* to establish the presence of moderate-to-severe OSA, as well as *Se* and *Acc* in 10 e/h, were also higher than previous AF and/or SpO_2-derived feature-engineering approaches [9,12–14]. This reinforces the suitability of DL methods to detect pediatric OSA, as they reach high performance by automatically extracting the information from the input signals and demonstrates the generalizability of our proposal since it reached high diagnostic performance for all cutoffs in both databases. Finally, the application of XAI adds value to our proposal against opaque models. Only one of the studies in the literature included an XAI algorithm to investigate the model outcomes, so most of the previous published models lack the capability to justify the reasons that led these models to predict the presence of OSA.

In summary, the results of this study confirm that a CNN + RNN architecture has the capability to identify pediatric OSA using AF and SpO_2 and that Grad-CAM explanations are useful to provide a reasoning about the model predictions.

5.4. Limitations and future work

This study presents some limitations that should be pointed out. The

combination of CNN and RNN can be further developed using novel DL approaches such as transformers and other hybrid architectures. Another limitation arises from the exclusive use of Grad-CAM to explain our model. Future goals may include testing other XAI algorithms like SHAP. Finally, although our model was developed and tested using two different databases, it would also be convenient to employ more multicentric databases that may include at-home or ambulatory recordings to validate our proposal and increase the robustness of our results.

6. Conclusion

A combination of CNN and RNN architectures trained with AF and SpO_2 signals showed high diagnostic performance in the detection of pediatric OSA. The proposed CNN + RNN reaches accurate estimations of pediatric OSA at the same time that the Grad-CAM XAI algorithm has the capability to justify these estimates by highlighting specific OSA-characteristic patterns of both signals, facilitating the model interpretation. The desaturations that followed apneas and/or hypopneas, along with the sudden AF amplitude changes were clearly identified as relevant patterns using Grad-CAM. These explanations serve to increase the trustworthiness of the model and can be used as a tool to aid sleep physicians to analyze and interpret these signals with the objective of simplifying the diagnosis of pediatric OSA.

CRediT authorship contribution statement

Jorge Jiménez-García: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **María García:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing - original draft, Writing - review & editing. **Gonzalo C. Gutiérrez-Tobal:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Writing - original draft, Writing - review & editing. **Leila Kheirandish-Gozal:** Data curation, Funding acquisition, Writing - original draft, Writing - review & editing. **Fernando Vaquerizo-Villar:** Formal analysis, Investigation, Methodology, Software, Writing - original draft, Writing - review & editing. **Daniel Álvarez:** Formal analysis, Funding acquisition, Investigation, Methodology, Writing - original draft, Writing - review & editing. **Félix del Campo:** Conceptualization, Funding acquisition, Supervision, Writing - original draft, Writing - review & editing. **David Gozal:** Data curation, Funding acquisition, Writing - original draft, Writing - review & editing. **Roberto Hornero:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This work was funded by Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación/10.13039/501100011033/, ERDF A way of making Europe, and NextGenerationEU/PRTR under grants PID2020-115468RB-I00 and PDC2021-120775-I00, and by CIBER -Consorcio Centro de Investigación Biomédica en Red- (CB19/01/00012), Instituto de Salud Carlos III. 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. G.C. Gutiérrez-Tobal was supported by a post-docotral grant from the University of Valladolid. D. Álvarez is supported by a "Ramón y Cajal" grant RYC2019-028566-I funded by Ministerio de Ciencia e Innovación - Agencia Estatal de Investigación and European Social Fund Investing in your future.

Ethical approval

This work has been carried out according to the Declaration of Helsinki. The clinical trial identifier of the original CHAT database is NCT00560859. In all patients, a written consent for parental permission for the research, along with assent for those children over 7 years of age, was obtained as part of the research protocol, which can be found in the supplementary material of Marcus et al [36]. The informed consents of all children caretakers were obtained in the UofC database, 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).

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. 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, Darien, IL, 2020. 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] A. Kaditis, L. Kheirandish-Gozal, D. Gozal, Pediatric OSAS: Oximetry can provide answers when polysomnography is not available, Sleep Med. Rev. 27 (2016) 96–105, https://doi.org/10.1016/j.smrv.2015.05.008.

[6] D. Bertoni, A. Isaiah, Towards Patient-centered Diagnosis of Pediatric Obstructive Sleep Apnea—A Review of Biomedical Engineering Strategies, Expert Rev. Med. Devices 16 (2019) 617–629, https://doi.org/10.1080/17434440.2019.1626233.

[7] G.C. Gutiérrez-Tobal, D. Álvarez, L. Kheirandish-Gozal, F. del Campo, D. Gozal, R. Hornero, Reliability of machine learning to diagnose pediatric obstructive sleep apnea: Systematic review and meta-analysis, Pediatr. Pulmonol. 57 (2022) 1931–1943, https://doi.org/10.1002/ppul.25423.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[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] 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.
[16] P. Ye, H. Qin, X. Zhan, Z. Wang, C. Liu, B. Song, Y. Kong, X. Jia, Y. Qi, J. Ji, L. Chang, X. Ni, J. Tai, Diagnosis of obstructive sleep apnea in children based on the XGBoost algorithm using nocturnal heart rate and blood oxygen feature, Am. J. Otolaryngol. 44 (2022), 103714, https://doi.org/10.1016/j.amjoto.2022.103714.
[17] 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.
[18] J. Jiménez-García, M. García, G.C. Gutiérrez-Tobal, L. Kheirandish-Gozal, F. Vaquerizo-Villar, D. Álvarez, F. del Campo, D. Gozal, R. Hornero, A 2D convolutional neural network to detect sleep apnea in children using airflow and oximetry, Comput. Biol. Med. 147 (2022), 105784, https://doi.org/10.1016/j.compbiomed.2022.105784.
[19] T. Van Steenkiste, W. Groenendaal, D. Deschrijver, T. Dhaene, Automated Sleep Apnea Detection in Raw Respiratory Signals Using Long Short-Term Memory Neural Networks, IEEE J. Biomed. Heal. Informatics. 23 (6) (2019) 2354–2364.
[20] 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.
[21] 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.
[22] 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 (Switzerland). 20 (2020) 1–19, https://doi.org/10.3390/s20185037.
[23] 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 Programs Biomed. 197 (2020), 105640, https://doi.org/10.1016/j.cmpb.2020.105640.
[24] 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 SpO2 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.
[25] 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.
[26] 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. Informatics Assoc. 25 (2018) 1643–1650, https://doi.org/10.1093/jamia/ocy131.
[27] 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.
[28] H. Korkalainen, J. Aakko, S. Nikkonen, S. Kainulainen, A. Leino, B. Duce, I. O. Afara, S. Myllymaa, J. Toyras, T. Leppanen, Accurate Deep Learning-Based Sleep Staging in a Clinical Population with Suspected Obstructive Sleep Apnea, IEEE J. Biomed. Heal. Informatics. 24 (2020) 2073–2081, https://doi.org/10.1109/JBHI.2019.2951346.
[29] L. Cheng, S. Luo, X. Yu, H. Ghayvat, H. Zhang, Y. Zhang, EEG-CLNet: Collaborative Learning for Simultaneous Measurement of Sleep Stages and OSA Events Based on Single EEG Signal, IEEE Trans. Instrum. Meas. 72 (2023) 1–10, https://doi.org/10.1109/TIM.2023.3235436.
[30] F. Teng, D. Wang, Y. Yuan, H. Zhang, A.K. Singh, Z. Lv, Multimedia Monitoring System of Obstructive Sleep Apnea via a Deep Active Learning Model, IEEE Multimed. 29 (2022) 48–56, https://doi.org/10.1109/MMUL.2022.3146141.
[31] A. Zarei, H. Beheshti, B.M. Asl, Detection of sleep apnea using deep neural networks and single-lead ECG signals, Biomedical Signal Processing and Control 71 (2022), 103125, https://doi.org/10.1016/j.bspc.2021.103125.
[32] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion. 58 (2020) 82–115, https://doi.org/10.1016/j.inffus.2019.12.012.
[33] C. Jansen, S. Hodel, T. Penzel, M. Spott, D. Krefting, Feature relevance in physiological networks for classification of obstructive sleep apnea, Physiol. Meas. 39 (2018), https://doi.org/10.1088/1361-6579/aaf0c9.
[34] C.F. Juang, C.Y. Wen, K.M. Chang, Y.H. Chen, M.F. Wu, W.C. Huang, Explainable fuzzy neural network with easy-to-obtain physiological features for screening obstructive sleep apnea-hypopnea syndrome, Sleep Med. 85 (2021) 280–290, https://doi.org/10.1016/j.sleep.2021.07.012.
[35] H.W. Loh, C.P. Ooi, S. Seoni, P.D. Barua, F. Molinari, U.R. Acharya, Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022), Comput. Methods Programs Biomed. 226 (2022), 107161, https://doi.org/10.1016/j.cmpb.2022.107161.
[36] 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.
[37] 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.
[38] 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.
[39] R.B. Berry, R. Budhiraja, D.J. Gottlieb, D. Gozal, C. Iber, V.K. Kapur, C.L. Marcus, R. Mehra, S. Parthasarathy, S.F. Quan, others, S. Redline, K.P. Strohl, S.L.D. Ward, M.M. Tangredi, Rules for Scoring Respiratory Events in Sleep: Update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events, J. Clin. Sleep Med. 8 (2012) 597–619. 10.5664/jcsm.2172.
[40] P. Várady, T. Micsik, S. Benedek, Z. Benyó, A novel method for the detection of apnea and hypopnea events in respiration signals, I.E.E.E. Trans. Biomed. Eng. 49 (2002) 936–942, https://doi.org/10.1109/TBME.2002.802009.
[41] R.T. Brouillette, A. Morielli, A. Leimanis, K.A. Waters, R. Luciano, F.M. Ducharme, Nocturnal Pulse Oximetry as an Abbreviated Testing Modality for Pediatric Obstructive Sleep Apnea, Pediatrics 105 (2000) 405–412, https://doi.org/10.1542/peds.105.2.405.
[42] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.
[43] H. Korkalainen, J. Aakko, B. Duce, S. Kainulainen, A. Leino, S. Nikkonen, I. O. Afara, S. Myllymaa, J. Töyräs, T. Leppänen, Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea, Sleep 43 (2020) 1–10, https://doi.org/10.1093/sleep/zsaa098.
[44] 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 Programs Biomed. 180 (2019), 105001, https://doi.org/10.1016/j.cmpb.2019.105001.
[45] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, (2014). http://arxiv.org/abs/1412.6980.
[46] P.J. Huber, Robust Estimation of a Location Parameter, Ann. Math. Stat. 35 (1964) 73–101, https://doi.org/10.1214/aoms/1177703732.
[47] F. Chollet, Keras, (2015).
[48] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, Proc. IEEE Int. Conf. Comput. Vis. (2017-Octob (2017)) 618–626, https://doi.org/10.1109/ICCV.2017.74.
[49] J. Cohen, A Coefficient of Agreement for Nominal Scales, Educ. Psychol. Meas. 20 (1960) 37–46, https://doi.org/10.1177/001316446002000104.
[50] 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.
[51] M. Dutt, S. Redhu, M. Goodwin, C.W. Omlin, SleepXAI: An explainable deep learning approach for multi-class sleep stage identification, Appl. Intell. (2022), https://doi.org/10.1007/s10489-022-04357-8.
[52] C.-E. Kuo, G.-T. Chen, P.-Y. Liao, An EEG spectrogram-based automatic sleep stage scoring method via data augmentation, ensemble convolution neural network, and expert knowledge, Biomed. Signal Process. Control 70 (2021), 102981, https://doi.org/10.1016/j.bspc.2021.102981.
[53] 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.
[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.