Measurement 264 (2026) 120259

Contents lists available at ScienceDirect

Measurement

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

Combined explainable deep learning model to predict pediatric sleep apnea from ECG and SpO_2

Clara García-Vicente [a,b], Gonzalo C. Gutiérrez-Tobal [a,b], Fernando Vaquerizo-Villar [a,b], Adrián Martín-Montero [a,b,c,*], David Gozal [d], Roberto Hornero [a,b]

[a] Biomedical Engineering Group, University of Valladolid, Valladolid, Spain
[b] CIBER de Bioingeniería, Biomateriales y Nanomedicina, Instituto de Salud Carlos III Valladolid, Spain
[c] Biomedical Engineering Group, Tecnología Electrónica, Ingeniería de Sistemas y Automática (TEISA) Department, University of Cantabria, Santander, Spain
[d] Office of The Dean and Department of Pediatrics, Joan C. Edwards School of Medicine, Marshall University, 1600 Medical Center Dr, Huntington, WV 25701, United States

ARTICLE INFO

Keywords:
Obstructive sleep apnea (OSA)
Deep learning (DL)
electrocardiogram (ECG)
Oxygen saturation (SpO_2)
eXplainable Artificial Intelligence (XAI)
SHapley Additive exPlanations (SHAP)

ABSTRACT

Combining deep learning (DL) with eXplainable Artificial Intelligence (XAI) techniques has led to clinically applicable models that simplify the diagnosis of pediatric obstructive sleep apnea (OSA) using a restricted number of cardiorespiratory signals. However, no prior study has applied these techniques to concurrently analyze electrocardiogram (ECG) and oxygen saturation (SpO_2) data. Here, we present an explainable DL approach integrating convolutional neural networks with overnight SpO_2 and ECG signals to identify pediatric OSA. SHapley Additive exPlanations (SHAP) XAI technique was used to extract relevant patterns linked to pediatric OSA and explain the model decisions. Patients (n = 3,320) from the semi-public Childhood Adenotonsillectomy Trial (CHAT) and Pediatric Adenotonsillectomy Trial for Snoring (PATS), and the private University of Chicago (UofC) databases were analyzed. Performance obtained Cohen's 4-class kappa of 0.549, 0.457, and 0.378 in CHAT, PATS, and UofC, respectively. Shapley values increased with OSA severity and highlighted the complementarity of SpO_2 and ECG, with SpO_2 being more relevant in moderate and severe cases and ECG in mild or no OSA cases. SHAP visualizations identified SpO_2 desaturations linked to clusters of apneic events and those occurring independently. It also highlighted bradycardia-tachycardia and ECG cardiovascular risk patterns, including variations in P and T waves, PQ and QT intervals, and the QRS complex. Shapley values identified correlations between respiratory and cardiac patterns, showing that desaturations in OSA are linked to cardiac changes. Therefore, our interpretable DL approach may improve pediatric OSA diagnosis by integrating breathing information and accompanying cardiac changes, supporting its effective adoption in clinical settings.

1. Introduction

Pediatric obstructive sleep apnea (OSA) is a common prevalent condition affecting approximately 1 % to 5 % of children, presenting unique challenges in its etiology, diagnosis, and treatment [1]. This disorder is characterized by recurrent episodes of complete airway obstruction (apneas) and/or significant airflow reduction (hypopneas) during sleep, leading to transient hypoxemia, hypercapnia, increased respiratory efforts, and arousal events [1,2]. Consequently, heightened sympathetic activity elicited by the repeated episodes of oxygen desaturation and disrupted sleep associated with OSA have been linked to various morbid consequences [1–3]. Indeed, if left untreated, these disturbances can increase the risk of neurocognitive and behavioral impairments and reduced cardiovascular and metabolic function [3]. Cardiovascular complications may include systemic and pulmonary vascular hypertension, while the metabolic consequences may manifest as dyslipidemia and insulin resistance [1,3–5]. As a result, well-being and quality of life, academic performance, and developmental progress are diminished. Despite its widespread occurrence, pediatric OSA remains significantly underdiagnosed, with only about 10 % of affected children receiving a confirmed diagnosis [6]. Early detection and treatment are essential to reducing associated health risks, as available therapies are highly effective [7]. However, the combination of high prevalence and low diagnosis rates leaves many children at risk of the

* Corresponding author at: Biomedical Engineering Group, Facultad de Medicina, Universidad de Valladolid, Av. Ramón y Cajal 7, 47003 Valladolid, Spain.
E-mail address: adrian.martin@uva.es (A. Martín-Montero).

https://doi.org/10.1016/j.measurement.2025.120259
Received 8 April 2025; Received in revised form 26 December 2025; Accepted 27 December 2025
Available online 28 December 2025
0263-2241/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

aforementioned serious consequences.

The gold standard for diagnosing OSA is overnight polysomnography (PSG), a comprehensive test conducted in a specialized laboratory [1]. During PSG, up to 32 biomedical signals from patients are recorded and monitored during the night, including electrocardiogram (ECG), blood oxygen saturation (SpO_2), and airflow (AF), among others [2]. Sleep specialists thoroughly analyze these physiological parameters to determine the apnea-hypopnea index (AHI). AHI reflects the average number of respiratory abnormalities (apnea and hypopnea events per hour of sleep (e/h)) and serves as the most widely used metric for assessing both the presence and severity of OSA [2,8]. The effectiveness of PSG is well-known, but it involves extensive monitoring, specialized equipment, and highly trained personnel [2]. These requirements make PSG a costly, complex, and often uncomfortable procedure with limited accessibility, particularly in the pediatric population. These issues, combined with the high prevalence of OSA in children, result in long waiting times and limited access to diagnosis and treatment [1].

Over the last decade, researchers have focused on developing alternative automated methods to simplify and facilitate pediatric OSA diagnosis [9]. Most of these studies have focused on analyzing signals such as AF, SpO_2, and ECG [10,11]. Specifically, the study of SpO_2 is highly relevant as it can provide critical insights into oxygenation disturbances during sleep, often linked to the respiratory events [12]. Its analysis enables the identification of specific patterns, such as recurrent oxyhemoglobin desaturations, which are related to recurrent airway obstructions [12]. Moreover, SpO_2 has demonstrated high diagnostic yield in previous pediatric OSA-related studies, making it a valuable tool for detecting OSA in children [10]. On the other hand, it is also essential to consider the ECG signal because of the strong interdependencies between the cardiovascular and respiratory systems during apneic events [13]. These episodes have been connected to changes in the heart rate (HR) leading to characteristic bradycardia-tachycardia patterns [14,15]. Moreover, OSA has been linked to a higher risk of developing cardiovascular complications that could persist and worsen into adulthood, especially when the condition is left untreated [1,3], which emphasizes the importance and usefulness of including ECG in OSA research.

A few previous studies have explored the analysis of cardiac signals using feature-engineering (FE) techniques such as photoplethysmography and heart rate variability (HRV), among others [10,16]. Contrasting with these traditional approaches, DL methods can process the complexities of raw signals directly, eliminating the need for preprocessing and feature extraction and selection stages [17–19]. In this regard, we previously integrated nocturnal one-lead ECG signals with DL techniques using convolutional neural networks (CNN) to estimate the presence and severity of pediatric OSA [11,20]. Results highlighted the effectiveness of CNN in automatically determining the presence of OSA and its severity in children using ECG. Similarly, previous research has illustrated the usefulness of CNNs in assessing pediatric OSA severity based on SpO_2 signals [21,22]. When taken together, these previous advancements advocate for a combined assessment of ECG and SpO_2. In this regard, the present study introduces a novel and unexplored approach by integrating cardiac activity and oxygenation data. This methodology could provide crucial diagnostic information, offering a more comprehensive understanding of physiological responses during respiratory stress. Furthermore, the joint analysis of both signals could reveal temporal relationships between desaturation events and cardiac responses, capturing dynamics that might remain unnoticed when the signals are analyzed separately, while also revealing susceptibility to cardiovascular morbidities.

Although advanced DL techniques have shown potential for predicting pediatric OSA, their main limitation lies in their lack of explainability [23]. This limitation is particularly important in the medical field [24–26], where professionals need to comprehend the reasoning behind automated decisions to trust and adopt these models. In this regard, eXplainable Artificial Intelligence (XAI) approaches are crucial in enhancing the interpretability and transparency of advanced computational models [23], especially within the healthcare domain [24–26]. Specifically, we believe that the application of XAI is relevant when analyzing ECG and SpO_2 together in the context of pediatric OSA. By discerning patterns in both signals, mainly those targeted by automated algorithms, valuable insights could be gained regarding the assessment of pediatric OSA severity. Additionally, this approach may help identify relationships between the pathophysiological patterns of SpO_2 desaturations and how the cardiovascular system responds to these events in relation to OSA disease [13]. Furthermore, it could uncover novel respiratory and cardiovascular risk factors linked to these signals in pediatric OSA. One widely used XAI technique for analyzing biomedical signals is SHapley Additive exPlanations (SHAP) [24]. In the context of sleep research, SHAP has proven effective in identifying physiological features associated with sleep stages and apneic events in adult and pediatric OSA [11,22,27–30]. However, no prior study has applied any XAI method to the combined analysis of ECG and SpO_2 data. To streamline the reading of the following sections, a detailed list of the main acronyms and definitions used throughout the manuscript has been included in the Supplementary Material (Table S1 and Table S2).

This study hypothesizes that our proposal, based on integrated DL models, together with SHAP, would simplify pediatric OSA diagnosis and enhance interpretability. Accordingly, this study had two main objectives. First, we aimed to evaluate a DL approach based on a stacked generalization strategy that integrates CNNs fed with overnight SpO_2 and ECG recordings to estimate the AHI and establish pediatric OSA severity. Second, we wished to incorporate SHAP as an XAI method to enhance interpretability and identify qualitative and quantitative complementary patterns within SpO_2 and ECG signals, and their relationship with pediatric OSA. Therefore, our study introduces two significant novelties:

- Development of a novel DL regression approach based on a stacked ensemble of CNNs to directly estimate the AHI from the combination of overnight SpO2 and single-lead ECG recordings.
- Application of an XAI technique, specifically SHAP, to interpret the decisions made by the model and evaluate the joint and individual contributions of ECG and SpO2 signals to pediatric OSA estimation.

2. Subjects and signals

A total of 3,320 pediatric sleep studies involving children aged 0 to 13 years comprised the study population. Three distinct databases were used for this study. The first was the Childhood Adenotonsillectomy Trial (CHAT), a publicly available database accessible upon request, which includes 1,609 valid ECG and SpO_2 recordings from PSG studies conducted in children aged 5 to 9.9 years with symptoms of OSA [31]. The second database, the Pediatric Adenotonsillectomy Trial for Snoring (PATS), also publicly available upon request, contains 731 valid ECG and SpO_2 recordings from PSG studies performed in children between 3 and 12 years old [32]. CHAT and PATS are multicenter, randomized, and single-masked design studies conducted in compliance with the Declaration of Helsinki (CHAT clinical trial: NCT00560859; PATS clinical trial: NCT02562040) [31–35]. Written consent was obtained from children's caretakers, following the research protocols in Marcus *et al.* [33] for CHAT and Redline *et al.* [32] for PATS. Children aged 7 or older also gave their assent in both studies. Study recordings from both databases were partitioned into three sets. The training set comprised 60 % of CHAT ($n = 987$) and 60 % of PATS ($n = 426$) and was used to train the model ($n_{train} = 1{,}413$). The validation set consisted of 20 % of CHAT ($n = 323$) and 20 % of PATS ($n = 152$) and was used to adjust the optimal configuration of the model ($n_{val} = 475$). Finally, the test set included 20 % of CHAT ($n = 299$) and 20 % of PATS ($n = 153$) and was used to evaluate model performance. Each dataset (CHAT and PATS) was partitioned independently, and it was conducted so that each subject was exclusively assigned to one of the sets, avoiding duplication. The input data was labeled with AHI values.

Additionally, the study incorporated a private database from the Pediatric Sleep Unit at Comer Children's Hospital, University of Chicago (UofC), USA [36]. This dataset comprised 980 sleep studies of children aged 0 to 13 years who were referred to the pediatric sleep laboratory due to symptoms suggestive of clinically suspected OSA. The research protocol was approved by the University of Chicago (UofC) Ethics Committee (#11–0268-AM017, #09–115-B-AM031, and #IRB14-1241), and informed consent was obtained from the legal guardians of all participants. The database was de-identified and used exclusively for external validation of the model trained and validated with CHAT and PATS datasets, following the approach used in previous studies [11]. Consequently, all 980 SpO_2 and ECG recordings from the UofC dataset were designated as the test set.

Sleep specialists scored PSG recordings from all databases according to the American Academy of Sleep Medicine (AASM) guidelines [37,38]. The criterion used to diagnose the presence and severity of pediatric OSA was the AHI. Based on AHI values, children were classified into one of the four frequently used categories: AHI < 1 e/h (no OSA), $1 \leq$ AHI $<$ 5 e/h (mild OSA), $5 \leq$ AHI < 10 e/h (moderate OSA), and AHI ≥ 10 e/h (severe OSA). The demographic and clinical variables of the children included in the study are presented in Table 1.

3. Methods

Fig. 1 summarizes an overview of the methodological workflow of this study. The present study implemented and evaluated an interpretable stacked ensemble-based DL model using one channel SpO_2 and ECG recordings (S_1, …, S_n) to directly estimate the AHI per subject (y_1, …, y_n). The model was trained with minimally preprocessed ECG and SpO_2 signals, which were first used to feed independent CNNs to extract feature maps. The feature sequences extracted from these independent CNNs were then combined and fed into a higher-level model using a stacking strategy [39]. This approach enabled the model to combine features from both signals to generate the final AHI estimation while allowing appropriate sample rates for each signal. Finally, the SHAP method was applied to qualitatively and quantitatively assess the contribution of ECG and SpO_2 to the model's decision. SHAP is a post-hoc interpretability method that assigns importance values to each input of a predictive model, offering a clearer understanding of the decision-making process [23,40,41]. Additionally, SHAP was also used to identify the most relevant ECG and SpO_2 regions on which the model was fixed to perform AHI estimation, facilitating the extraction of patterns associated with pediatric OSA.

3.1. ECG and SpO_2 signals preprocessing

Following the guidelines established by the AASM, SpO_2 and ECG-II lead data were collected from the CHAT, PATS, and UofC datasets [37,38]. To ensure consistency, all databases underwent uniform preprocessing. Nevertheless, the processing of the SpO_2 signals was performed independently from the ECG signals, allowing for tailored handling of each signal type while maintaining overall uniformity in preprocessing. Consistent with previous works, raw ECG and SpO_2 signals were resampled at 100 Hz and 1 Hz, respectively [11,20,21,42–44]. No additional preprocessing steps were needed for the SpO_2 signal beyond this resampling [21,22]. In contrast, the ECG underwent further preprocessing, based on prior studies, to enhance signal quality [11,20]. First, the continuous component was adjusted by removing the mean value within 30-second windows. Next, a band-pass linear-phase finite impulse response filter with cut-off frequencies of 0.5 Hz and 50 Hz was applied to minimize noise while preserving essential signal components, particularly those related to QRS complexes [45].

SpO_2 and ECG recordings were standardized to a duration of eight hours, as this timeframe had previously demonstrated optimal performance in the validation set in previous research [11]. For signals with fewer samples, zero-padding was applied at the beginning of the recording. Conversely, for longer recordings, excess data were removed from the start of the signal, when children are more probably awake, following methodologies used in prior OSA studies that analyzed unsegmented cardiorespiratory signals [11,46–48].

Subsequently, ECG and SpO_2 signals were preprocessed to match the input dimensions required by the DL-based model. SpO_2 signals from each subject were divided into 20-minute segments. This segmentation strategy aligned with the methodology used in previous studies [21]. Various segment durations (5, 10, 20, 30, and 60 min) and overlaps (50 % and 90 %) were evaluated to find the optimal segment duration [21]. The highest kappa was obtained using 20-minute non-overlapping segments. This duration was consistent to capture desaturation clusters, which are typically 10 min long at least [49]. SpO_2 signals were resized to match the input specifications of the corresponding CNN before being fed into the model, generating matrices composed of 24 segments, each representing 20 min ($24 \times 1{,}200 \times 1 = 28{,}800$ samples). This structure was optimized for block processing within CNNs embedded in time-distributed (TD) layers. Additionally, it facilitated the transfer of the optimal architecture, including pre-trained weights and layers, from a previously developed SpO_2-based CNN model [21]. This model was optimized for estimating the number of events per segment and was integrated into the SpO_2-CNN architecture developed in this study. Similarly, ECG signals were transformed into arrays of 48 segments, each lasting 10 min ($48 \times 60{,}000 \times 1 = 2{,}880{,}000$ samples). The selection of 10-minute duration was made because 10-minute segments were identified as optimal for training the previously developed CNN [20], as they effectively captured clusters of apneic events [49]. This structure allowed for the integration of overnight ECG recordings into the model, while ensuring they matched the 8-hour duration of the SpO_2. Like SpO_2, this format was also optimized for block processing within the TD layers of the CNN. Furthermore, it allowed the adaptation of the optimal architecture, along with its pre-trained weights and

Table 1
Demographic and clinical variables of the study population.

	Training set (n_{train} = 1,413)		Validation set (n_{val} = 475)				
Variables	CHAT Training set	PATS Training set	CHAT Validation set	PATS Validation set	CHAT Test set	PATS Test set	UofC Test set
Subjects (n)	987 (61.34)	426(58.28)	323 (20.07)	152 (20.79)	299 (18.58)	153 (20.93)	980 (100)
Age (years)	7.00 [2.00]	7.69 [0.00]	7.00 [2.00]	7.50 [0.00]	6.90 [2.00]	7.56 [0.00]	6.00 [6.00]
Males (n)	510 (51.67)	221(51.88)	164 (50.77)	71(46.71 %)	161 (53.85)	86 (56.21)	379 (38.67)
BMI (kg/m^2)	17.31[5.93]	18.98 [0.0]	17.12[6.25]	18.37[0.00]	17.43[6.04]	17.95[0.00]	18.02[6.02]
AHI (e/h)	2.64 [4.78]	1.30 [2.30]	2.45 [4.77]	1.00 [2.10]	2.32 [5.11]	1.00 [2.60]	3.8 [7.78]
AHI < 1 e/h$^{(1)}$ (n)	212 (21.48)	163(38.26)	67 (20.74)	72 (47.37)	65 (21.74)	75 (49.02)	173 (17.65)
$1 \leq$ AHI < 5 e/h$^{(2)}$(n)	487 (49.34)	190(44.60)	167 (51.70)	62 (40.79)	144 (48.16)	53 (34.64)	401 (40.92)
$5 \leq$ AHI < 10e/h$^{(3)}$(n)	159 (16.11)	29 (6.81)	44 (13.62)	11 (7.24)	49 (16.39)	14 (9.15)	177 (18.06)
AHI ≥ 10 e/h$^{(4)}$ (n)	129 (13.07)	44 (10.33)	45 (13.93)	7 (4.61)	41 (13.71)	11 (7.19)	229 (23.37)

Data are presented as median [interquartile range] or n (%). BMI: body mass index; AHI: apnea-hypopnea index; e/h: events/hour; CHAT: Childhood Adenotonsillectomy Trial; PATS: Pediatric Adenotonsillectomy Trial for Snoring; UofC: University of Chicago.
AHI < 1 e/h$^{(1)}$: No OSA; $1 \leq$ AHI < 5 e/h$^{(2)}$: mild OSA; $5 \leq$ AHI < 10 e/h$^{(3)}$: moderate OSA; AHI ≥ 10 e/h$^{(4)}$: severe OSA.

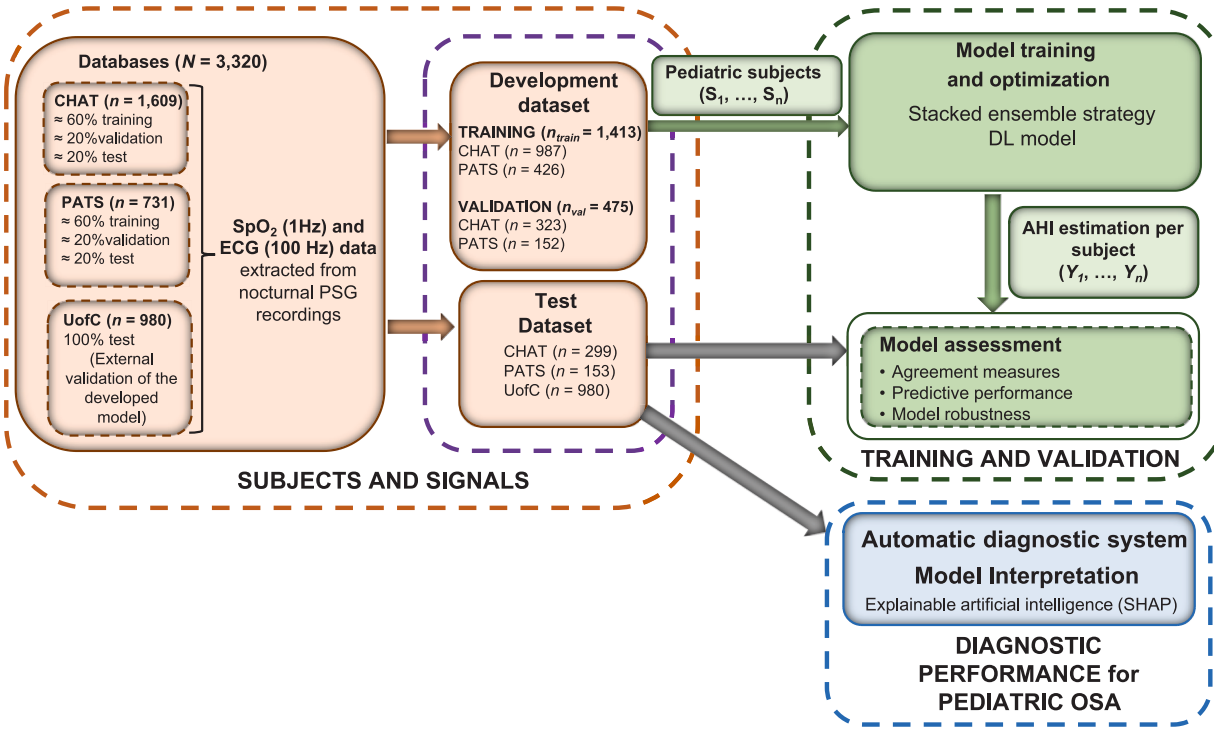

Fig. 1. Proposed workflow for developing, validating, and explaining the DL-based model enabling prediction and interpretation of pediatric OSA severity using SpO_2 and ECG recordings. S_n: subject n; y_n: estimation of AHI in subject n.

layers, from a previously developed ECG-based CNN approach into the ECG-CNN architecture implemented in this research [20]. A significant advantage of this implementation lies in its ability to process the ECG and SpO_2 signals independently, without the need for a uniform sampling rate. This approach streamlines the workflow by allowing the use of data with varying temporal resolutions while ensuring effective

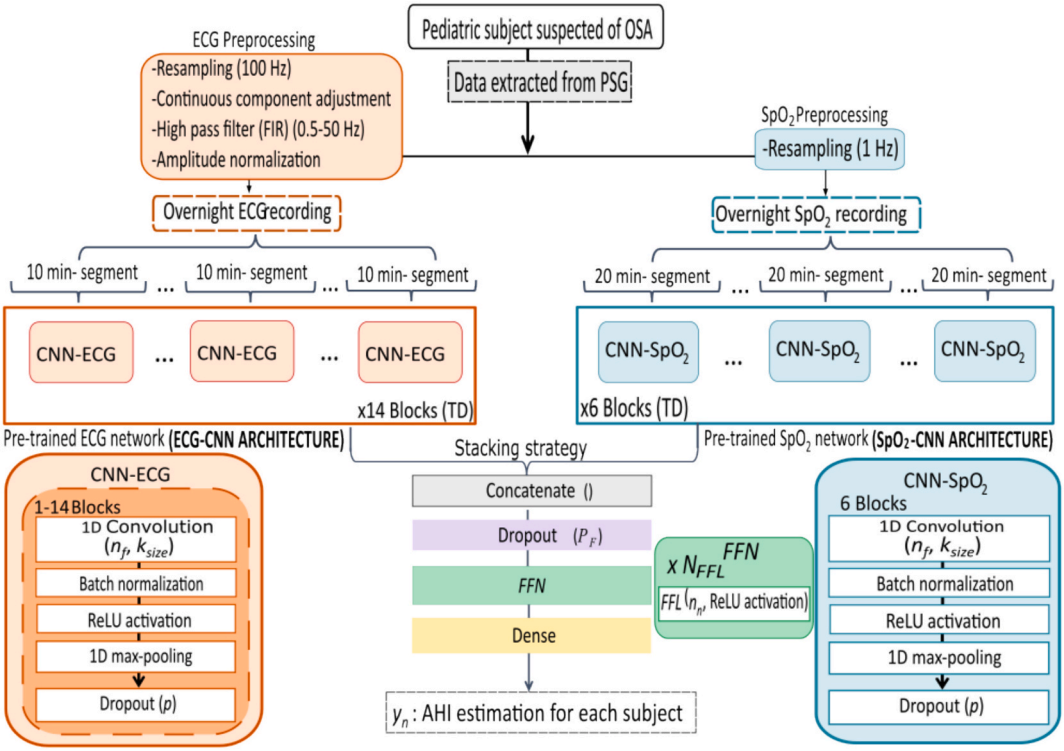

Fig. 2. Overall scheme of the proposed regression DL model based on a stacked ensemble strategy. The input data consists of each subject's nocturnal ECG and SpO_2 recordings. The output (y_n) corresponds to the AHI estimation per subject. CNN: convolutional neural network; ReLU: rectified linear unit activation; TD: time distributed; AHI: apnea hypopnea index; FFN: feed forward network; FFL: feed forward layer; n_f: number of filters; k_{size}: kernel size; P_F: stacking dropout layer probability; n_n: FFL neuron counts; N_{FFL}: Number of FFL; p: CNN dropout layer probability.

integration of the signals at the final stage of processing.

3.2. Stacking-based DL architecture

Fig. 2 illustrates the proposed DL architecture developed for estimating the AHI per subject from ECG and SpO_2 signals. The proposal is based on a stacking strategy, also known as stacked generalization, a machine-learning ensemble technique combining multiple models to enhance overall performance [39]. The core concept of stacking involves using the predictions from several base models as inputs to a higher-level model, referred to as the *meta*-model or blender, which aggregates them to generate the final prediction [39,50]. Following this strategy, our model processes nocturnal ECG and SpO_2 signals independently before integrating the extracted features to identify patterns indicative of pediatric OSA. Specifically, the architecture consists of two parallel base CNN models, each designed to process distinct physiological signals (ECG-CNN and SpO_2-CNN architectures, respectively, in Fig. 2). After independent feature extraction, the outputs from ECG-CNN and SpO_2-CNN architectures are concatenated following the stacking strategy. Subsequently, these combined features are passed through a fusion module for further processing, ultimately generating the final AHI estimation.

ECG-CNN architecture was implemented using a cluster of TD layers, incorporating the previously presented CNN architecture (CNN-ECG in Fig. 2) from a hybrid convolutional and recurrent neural network trained using 8-hour ECG signals [11]. ECG-CNN module, composed of 14 convolutional blocks, was optimized to extract relevant temporal and spatial patterns. Each of the blocks was encapsulated in TD layers, preserving the sequence format of the data as they were processed in the CNN-ECG layers. Within each CNN-ECG, a one-dimensional (1D) convolutional layer was applied with a specific number of filters (n_f) and a defined kernel size (k_{size}). This was followed by batch normalization, a rectified linear unit (ReLU) activation function, and a 1D max-pooling layer. Finally, a dropout layer with probability p was incorporated [17]. In parallel, the SpO_2-CNN architecture was implemented using a clustering of TD layers, which consist of the previously presented CNN layers trained on 20-minute SpO_2 signals (CNN-SpO_2 in Fig. 2) [21]. This architecture, composed of 6 convolutional blocks, followed a similar structure to those in the related CNN-ECG module, but optimized for lower-resolution data. Each block was also encapsulated in TD layers, which preserved the sequential format of the data as it was processed through the CNN-SpO_2 layers. Once the feature extraction processes were completed, the outputs from both the ECG-CNN and SpO_2-CNN architectures were concatenated to combine information and create a unified feature representation, following a stacking strategy. This combined feature vector was processed through a fusion module to refine the learned patterns and identify relationships between the two physiological signals. The fusion module began with an initial dropout layer with probability P_F to reduce overfitting, followed by several feed-forward layers (FFLs), where the number of layers is denoted as N_{FFL} and each layer contains n_n neurons. Each FFL was activated with a ReLU function to enhance feature abstraction [17]. Finally, a dense output layer with a linear activation function was implemented to estimate the AHI for each subject.

3.3. Training, optimization, and evaluation process

The proposed DL model was trained on an NVIDIA GeForce RTX 4090 GPU and Keras 2.10.0 framework with TensorFlow-gpu 2.10.1 backend. The He-normal method was utilized for weight initialization, and the adaptive moment estimation (Adam) optimizer was used with an initial learning rate of $1x10^{-4}$ [17]. The training process was conducted over 200 epochs with a batch size of 150 samples. Consistent with previous studies [17,20], the Huber loss function with a delta parameter ($\delta = 1.5$) was used in the Adam optimization, chosen for its established robustness in regression tasks with large outliers. Finally, to prevent overfitting, early stopping was implemented based on validation loss monitoring.

The hyperparameters of each of the convolutional branches were optimized in previous studies, obtaining models with high performance [11,21]. Regarding the CNN-ECG module, the convolutional layers were structured as follows: $blocks_{1\text{-}4}$ consisted of $n_f = 16$ with $k_{size} = 33$; $blocks_{5\text{-}8}$ involved $n_f = 32$ with $k_{size} = 17$, $blocks_{9\text{-}12}$ comprised $n_f = 64$ with $k_{size} = 7$, and $blocks_{13\text{-}14}$ consisted of $n_f = 64$, but with $k_{size} = 3$. A dropout layer was applied at the end of each module, with $p = 0.1$ for $blocks_{1\text{-}12}$, and $p = 0.4$ for $blocks_{13\text{-}14}$ [11]. In the CNN-SpO_2 module, all convolutional layers were composed of $n_f = 64$ with $k_{size} = 5$, and dropout layers comprised $p = 0.1$ [21]. At this stage, we tuned a set of hyperparameters from scratch. Following the application of the stacking strategy, the fusion module began with a dropout layer, where P_f was varied within the range {0.0, 0.1, 0.2, 0.3, 0.4, 0.5} to determine the optimal setting. Afterward, a total of N_{FFL} *FFLs*, each with a n_n neurons, were incorporated and trained from scratch. The value of N_{FFL} was varied in the range {0, 1, 2, 3}, while n_n in each *FFL* was explored within the range {8, 16, 32, 64, 128} to determine the optimal performance configuration. We implemented a comprehensive fitting approach using the grid search technique to optimize the remaining hyperparameters in the fusion module after stacked generalization. This process involved evaluating every possible combination of hyperparameters within the specified search space. Finally, the performance of the algorithm was assessed to identify the optimal hyperparameter configuration. To this end, the four-class Cohen's kappa coefficient (k) was used [51]. This metric was calculated for subject-wise classification of OSA severity within the validation set from CHAT and PATS. The architecture with the highest k was then selected as the optimal model.

3.4. Model interpretability using SHAP

To enhance the interpretability of our DL-based model, we applied XAI using SHAP, a post-hoc interpretability method founded on game theory and local explanations [23]. This method is based on the integration of various explanation techniques, providing a unified approach to model interpretability [23,52]. SHAP assigns importance values (termed SHAP values) to each model input, turning its decision-making process more transparent. In our case, it assigns Shapley values for each sample of a signal [53]. These values measure the impact of individual signal samples while considering all possible interactions, providing a comprehensive understanding of their influence. The SHAP method builds an additive attribution model, where the sum of the Shapley values approximates the output of the model [23]. This approach ensures that the model identifies relevant features during training and relies on appropriate inputs for inference.

In this study, we applied SHAP to interpret the decisions of the model based on ECG and SpO_2 input signals. It evaluated both the joint and individual contributions of these signals for pediatric OSA estimation [53]. For this purpose, we utilized the DeepExplainer method to apply SHAP, as it offers visualization maps and is compatible with our data [53]. It is based on the Deep Learning Important Features (DeepLIFT) algorithm, which assigns attribution values to individual nodes in a neural network [54]. DeepExplainer enhances this approximation by aggregating per-node attributions across multiple background samples, ensuring that the sum of Shapley values accurately captures the difference in model outputs between the background signals and the input being analyzed [52,53]. Specifically, we computed Shapley values for each ECG and SpO_2 signal sample to gain a deeper understanding of the mechanisms of the model in recognizing respiratory event-related information and identifying cardiorespiratory patterns associated with pediatric OSA [55]. Since our model processed ECG and SpO_2 independently, we calculated Shapley values separately for each input. Particularly, Shapley values were assigned to each ECG and SpO_2 sample. In our architecture, the ECG and SpO_2 signals were first processed independently and then fused through concatenation before the final

layers. This fusion step allowed the model to learn complex, non-linear relationships between both modalities. Therefore, any interdependence between ECG and SpO_2 features was implicitly reflected in the Shapley values of the fused representations. Thus, this analysis allowed us to evaluate the joint and distinct contributions of ECG and SpO_2 to the overall model performance, that is, how each signal influenced the AHI estimations. Finally, after computing the Shapley values for each ECG and SpO_2 signal sample concerning the AHI estimation, we aggregated them at the subject level. Then, we summed the Shapley values of each signal separately and categorized them by OSA severities (no OSA, mild OSA, moderate OSA, and severe OSA).

3.5. Statistical analysis and diagnostic ability

To assess the performance of the proposed algorithm in diagnosing pediatric OSA, subjects were classified into one of the four severity categories based on their estimated AHI values. After categorizing the subjects, we calculated the confusion matrix, followed by the four-class accuracy (Acc_4), and the k coefficient [51]. Finally, we evaluated the diagnostic effectiveness of the model by calculating several performance metrics, including specificity (Sp), sensitivity (Se), negative and positive predictive values (NPV and PPV), negative and positive likelihood ratios (LR^- and LR^+), as well as overall accuracy (Acc) for the different OSA severity thresholds, considering AHI values of 1, 5 and 10 e/h.

4. Results

4.1. Optimal model configuration and ablation tests

The optimal model configuration was determined through an extensive search of all possible hyperparameter combinations. The chosen hyperparameters for the fusion model layers, based on the stacking strategy, included a dropout rate of $P_f = 0.1$. Fig. 3 illustrates the search space of the fusion module, along with k values obtained on the validation set for each specific hyperparameter. The fusion module consisted of 3 FFLs, with specific configurations: $n_1 = 16$ for $N_{FFL1} = 1$, $n_2 = 32$ for $N_{FFL2} = 2$, and $n_3 = 64$ for $N_{FFL3} = 3$. This configuration yielded the highest k value ($k = 0.501$) on the validation set (CHAT and PATS). As a result, it was selected for model evaluation on the test sets from CHAT, PATS, and UofC.

Finally, several ablation tests were conducted to evaluate the impact of different architectural components on the performance of the model. Initially, the ECG-CNN and SpO_2-CNN architectures were tested separately, with their outputs concatenated for direct AHI prediction, without the fusion model after the stacking strategy. This configuration resulted in a $k = 0.4835$ on the validation set, which was lower compared to the $k = 0.5011$ achieved by the proposed approach. Next, a single FFL was introduced after the ECG-CNN and SpO_2-CNN architectures were concatenated, but this led to a slight decrease in performance, with a k of 0.4653. A subsequent variant, which included 2 FFLs following the concatenation, showed slight improvement, yielding a k of 0.4991. Finally, the optimal proposed model, incorporating 3 FFLs after the stacking strategy, was trained and validated using the original training and validation sets from CHAT and PATS, as well as data from UofC ($n = 587$ for training and $n = 197$ for validation). This methodology enabled the evaluation of the influence of UofC data on the performance of the model, resulting in a $k = 0.4940$ on the validation set from CHAT, PATS, and UofC. Despite testing various configurations in the ablation studies, none of them outperformed the optimal value of $k = 0.5011$, which was achieved with the optimal architecture. This underscores the significance of the selected architectural components. Simplifying the architecture by removing the FFLs did not result in any performance improvement, further reinforcing the robustness of the proposed approach.

4.2. Diagnostic performance of the DL-based proposal

Fig. 4 displays the confusion matrices obtained in the test sets after classifying OSA severity for each subject based on their estimated AHI. Additionally, the four-class classification metrics derived from the confusion matrices were $k = 0.549$ and $Acc_4 = 70.23$ % in the CHAT test set, $k = 0.457$ and $Acc_4 = 64.05$ % in the PATS test set, and $k = 0.378$ and $Acc_4 = 56.43$ % in the UofC test set. The diagnostic performance of pediatric OSA severity in CHAT, PATS, and UofC test sets is shown in Table 2 based on the standard AHI severity thresholds of 1, 5, and 10 e/h. The model demonstrated high Acc across all datasets, with the highest values obtained for the most severely affected children (AHI = 10 e/h), achieving $Acc = 94.98$ % in CHAT, $Acc = 98.08$ % in PATS, and $Acc = 88.98$ % in UofC. This result is particularly significant, as these children benefit the most from timely and accurate diagnosis. Se values remained consistently high across all thresholds and datasets, ranging from 80 % to 96 %, ensuring reliable detection of OSA at all severity levels. Sp values remained notably strong for moderate and severe OSA, with values ranging between 78 % and 98 % (96 % and 98 % for PATS and CHAT, respectively) for moderate OSA and between 90 % to 99 % for severe OSA. These results highlight the strong capability of the model in

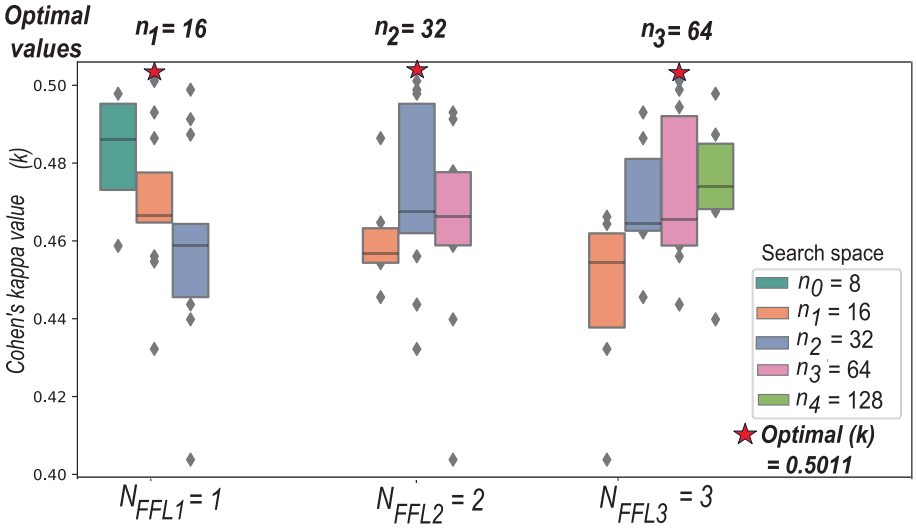

Fig. 3. Search space of the fusion module after the stacking strategy, as well as k values obtained in the validation set for each specific hyperparameter. The optimal values for the hyperparameters are indicated by red stars. FFL: feed forward layer; n_n: FFL neuron counts; N_{FFL}: Number of FFL.

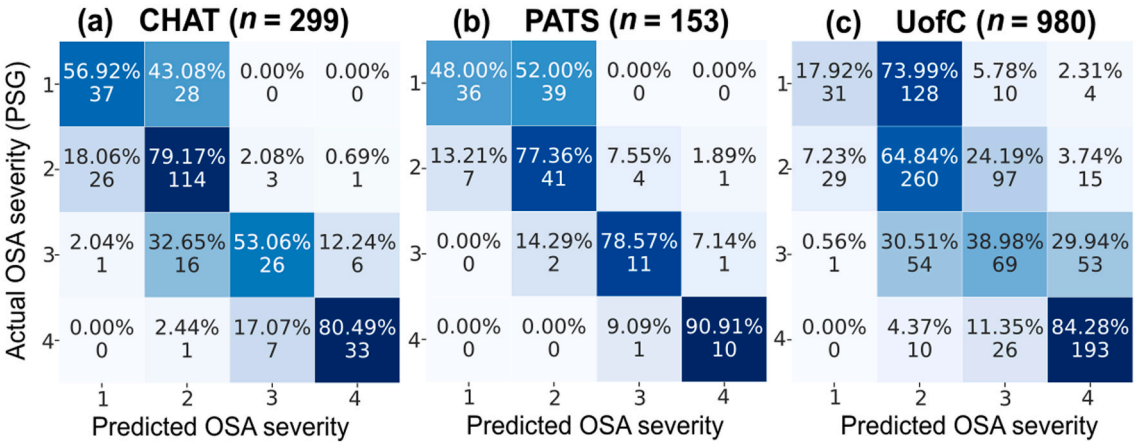

Fig. 4. Confusion matrices of the DL-based model in CHAT, PATS, and UofC test sets. *1: no OSA (AHI < 1 e/h); 2: mild OSA (1 ≤ AHI < 5 e/h); 3: moderate OSA (5 ≤ AHI < 10 e/h); 4: severe OSA (AHI ≥ 10 e/h).

Table 2
Diagnostic performance of the combined DL-based approach in the CHAT, PATS, and UofC test sets.

AHI threshold	Test set	*Sp* (%)	*Se* (%)	*NPV* (%)	*PPV* (%)	LR^-	LR^+	*Acc* (%)
1 e/h	CHAT	56.92	88.46	57.81	88.09	0.20	2.05	81.61
	PATS	48.00	91.03	83.72	64.55	0.19	1.75	69.93
	UofC	17.92	96.28	50.82	84.55	0.21	1.17	82.45
5 e/h	CHAT	98.09	80.00	91.93	94.74	0.20	41.80	92.64
	PATS	96.09	92.00	98.40	82.14	0.08	23.55	95.42
	UofC	78.05	83.99	87.33	73.02	0.21	3.82	80.51
10 e/h	CHAT	97.29	80.49	96.91	82.50	0.20	29.67	94.98
	PATS	98.59	90.91	99.29	83.33	0.09	64.55	98.04
	UofC	90.41	84.28	94.97	72.83	0.17	8.79	88.98

Sp (specificity); *Se* (sensitivity); *NPV* and *PPV* (negative and positive predictive value); LR^- and LR^+ (negative and positive likelihood ratio); AHI (apnea-hypopnea index); e/h (events per hour); CHAT (Childhood Adenotonsillectomy Trial); PATS (Pediatric Adenotonsillectomy Trial for Snoring); UofC (University of Chicago).

diagnosing clinically moderate-to-severe OSA. Significantly, the LR^+ values for 10 e/h were 29.67 in CHAT, 64.55 in PATS, and 8.79 in UofC, further revealing significant diagnostic ability for severe OSA, especially in CHAT and PATS datasets.

4.3. Identification of ECG and SpO_2 patterns using SHAP

Fig. 5 presents boxplots showing the sum of the Shapley values assigned to ECG signals (left) and SpO_2 signals (right) for each subject across the four pediatric OSA severity groups. The results indicate that, for both ECG and SpO_2 signals, the total sum of Shapley values increases with higher severity levels. This reveals that the model attributes higher importance to signal patterns in subjects with more severe disease, highlighting a relationship between OSA severity and the contribution of these signals to AHI estimation. Additionally, boxplots show that the overall sum of Shapley values is higher for SpO_2 signals compared to ECG signals in moderate and severe OSA. However, for no OSA and mild OSA, the Shapley values are higher for ECG signals. This suggests that ECG contributes more to AHI estimation in these lower-severity cases, whereas SpO_2 becomes more relevant as OSA severity increases.

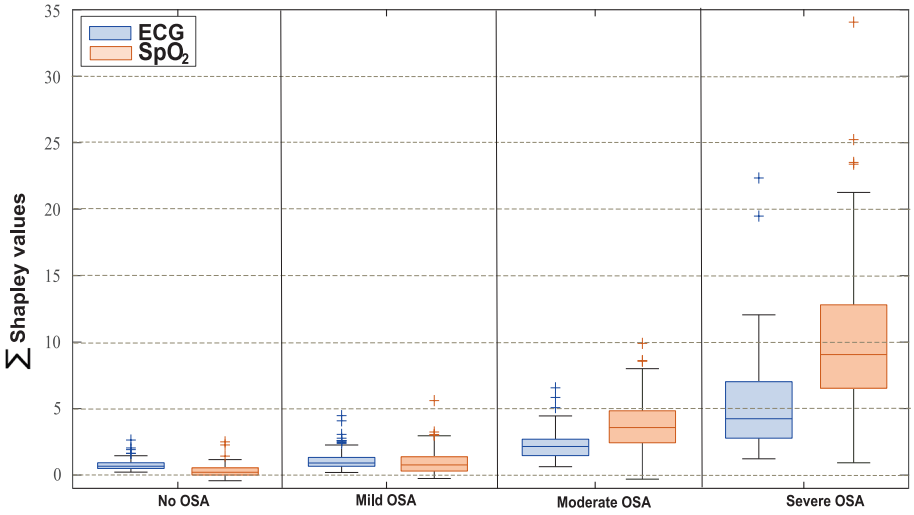

Fig. 5. Boxplots showing the sum of the Shapley values assigned to the ECG (left) and SpO_2 (right) signals for each subject, grouped by pediatric OSA severity groups.

Figs. 6-8 show examples of SHAP maps for SpO_2 and ECG signals from different PSGs, with red-colored regions highlighting key patterns that contribute to accurate AHI estimation and blue-colored regions indicating patterns that reduce the AHI model prediction. Fig. 6 (a), Fig. 7 (a, b), and Fig. 8 (a) display zoomed-in views of key regions extracted from the SpO_2 signals, while Fig. 6 (b, c), Fig. 7 (c, d, e), and Fig. 8 (b, c) present a zoom of relevant areas from the ECG signals. The annotations indicating the presence or absence of respiratory events, as obtained from the PSG, are shown in brown.

Specifically, in Fig. 6 (a), the model has shown fixation in regions with clusters of apneic events, where successive $>$ 3 % desaturations in the SpO_2 signal are observed, as well as SpO_2 drops of 2 %. Fig. 6 (b) shows the same time region for the ECG, highlighting in green the area shown in Fig. 6 (c), which also coincides with the time interval linked to D3 (SpO_2 desaturations $>$ 3 %) in Fig. 6 (a). Finally, in Fig. 6 (c), it is observed how SHAP visualization identifies the QRS complexes of different beats encompassing long PQ, and areas comprising long TP segments. Particularly, blue-colored QRS complexes (negative Shapley values) suggested morphologically normal cardiac patterns that drove the model toward lower AHI estimates, whereas red-colored QRS complexes (positive Shapley values) represented abnormal waveform patterns associated with more severe OSA, leading the model to predict higher AHI values. Fig. 7 (a) and Fig. 7 (c) show a region of SpO_2 and ECG signals, respectively, referring to a subject with the presence of apneic events. Fig. 7 (a) shows how the model focuses its attention mainly on the region marked in green, containing a cluster of higher than 3 % SpO_2 desaturations. Fig. 7 (b) corresponds to a zoom of the region marked in green in Fig. 7 (a), referring to SpO_2. In this case, it can be seen how SHAP maps highlight SpO_2 drops higher than 3 % (D3) and SpO_2 drops of 2 % (D2). Fig. 7 (d) corresponds to the first zoomed-in region marked in green in Fig. 7 (c), referring to ECG. In this case, the SHAP visualization emphasizes the delayed response of the regions when these events occur, as well as the areas between events. Fig. 7 (e) corresponds to the second ECG zoomed-in region marked in green in Fig. 7 (c). Here, the SHAP map highlights regions associated with HR variations, revealing a bradycardia-tachycardia pattern. Additionally, the P and T waves are distinctly identified in red color. Fig. 8 (a) and Fig. 8 (b) present SpO_2 and ECG regions, respectively, corresponding to a subject experiencing apneic events. In Fig. 8 (a), the model focuses on a cluster of higher than 3 % SpO_2 desaturations and offers a detailed visualization of how SHAP emphasizes SpO_2 drops over 3 % (D). Fig. 8 (c) corresponds to the ECG zoomed-in region marked in green in Fig. 8

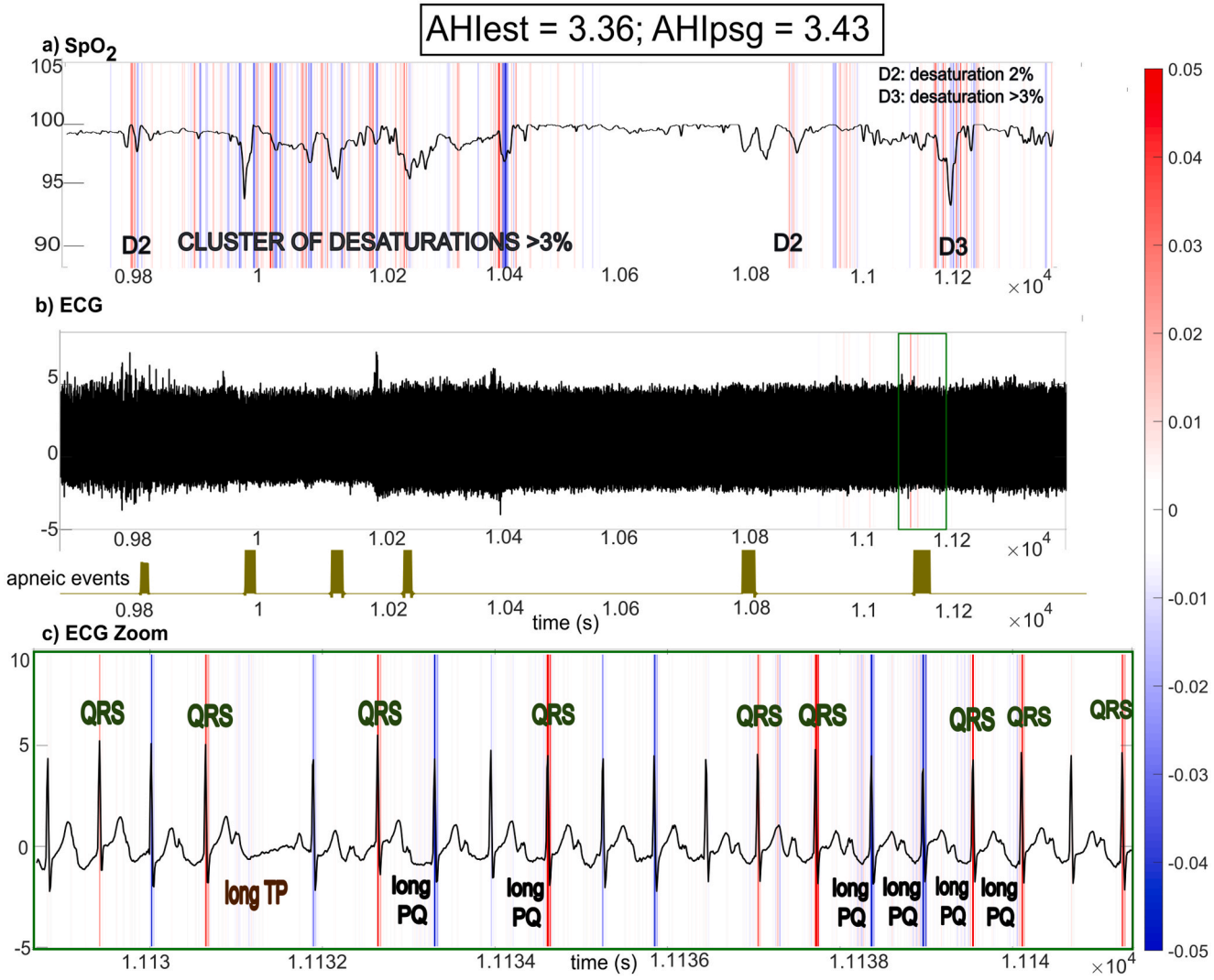

Fig. 6. SHAP visualizations of some representative findings in SpO_2 and ECG signals from the CHAT dataset corresponding to a child with mild OSA. Fig. 6 (a) and Fig. 6 (b) show the map of the same ECG and SpO_2 regions. Fig. 6 (c) is a zoom of the ECG area marked in green in Fig. 6 (b). Areas with SpO_2 desaturations over 3% are marked with D3, while areas with SpO_2 desaturation of 2% are marked with D2. QRS: ECG QRS complex. TP and PQ indicate intervals between T-P and P-Q waves of the ECG. The color bar indicates the Shapley values linked to each ECG and SpO_2 sample. Blue-colored areas show patterns that lower the AHI estimation, while red-colored regions indicate patterns that contribute to accurate AHI estimation.

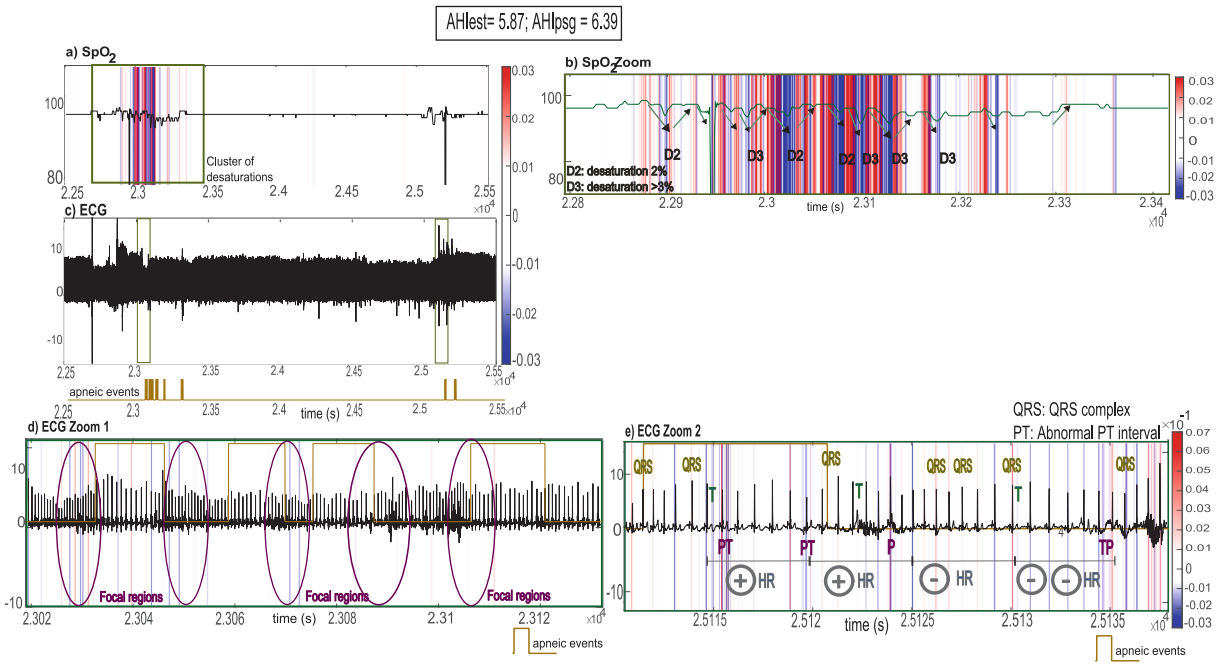

Fig. 7. SHAP visualizations of some representative findings in SpO_2 and ECG signals from the CHAT dataset corresponding to a child with moderate OSA. Fig. 7 (a) and Fig. 7 (c) show the map of the same ECG and SpO_2 regions. Fig. 7 (b) is a SpO_2 zoom of the area marked in green in Fig. 7 (a) referring to a cluster of SpO_2 desaturations. Areas with SpO_2 desaturation of over 3% are marked with D3, while areas with oxygen desaturation of 2% are marked with D2. Fig. 7 (d) and Fig. 7 (e) are ECG zooms of the areas marked in blue in Fig. 7 (c). P and T: P and T waves of the ECG. Zones of changes in heart rate (HR) increase and decrease (↑HR and ↓HR) are indicated in green. The color bar indicates the Shapley values linked to each ECG and SpO_2 sample. Blue-colored areas show patterns that lower the AHI estimation, while red-colored regions indicate patterns that contribute to accurate AHI estimation.

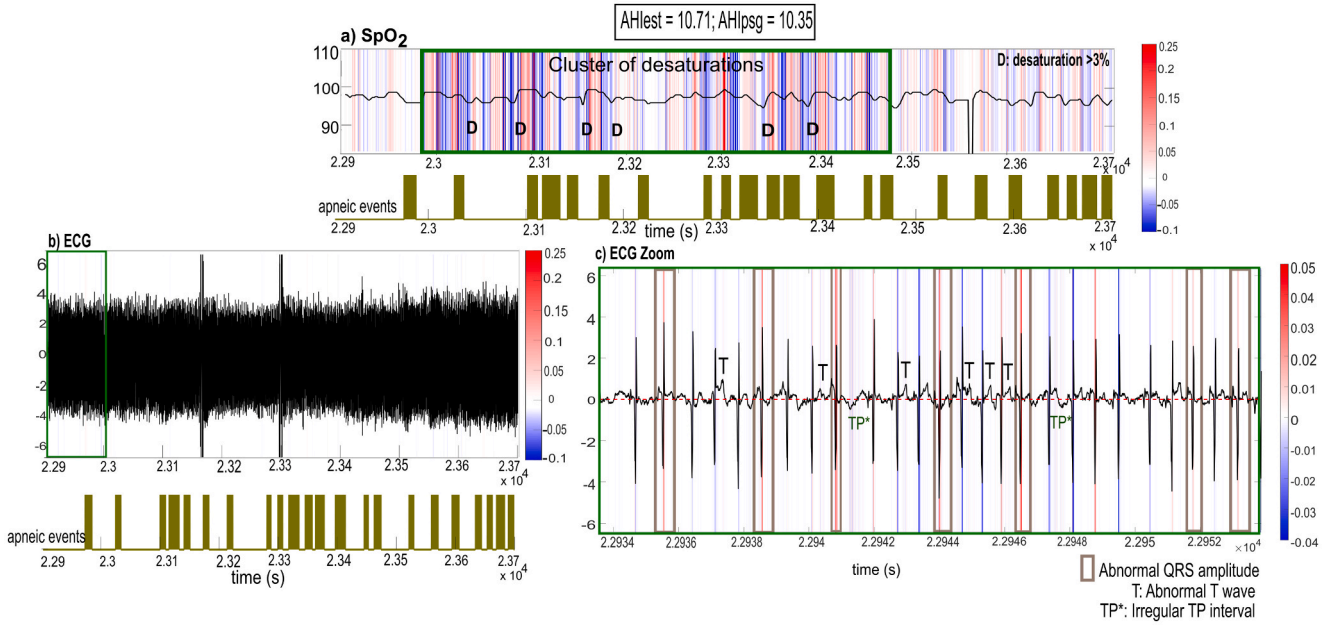

Fig. 8. SHAP maps of some representative findings in SpO_2 and ECG signals from the CHAT dataset corresponding to a child with severe OSA. Fig. 8 (a) and Fig. 8 (b) show the map of the same ECG and SpO_2 regions. Fig. 8 (a) is a SpO_2 region indicating an area with a cluster of SpO_2 desaturations. Areas with SpO_2 desaturation of over 3% are marked with D. Fig. 8 (c) is a zoom of the ECG area marked in green in Fig. 8 (b). T: ECG T wave. TP indicates intervals between T-P waves of the ECG. The color bar indicates the Shapley values linked to each ECG and SpO_2 sample. Blue-colored areas show patterns that lower the AHI estimation, while red-colored regions indicate patterns that contribute to accurate AHI estimation.

(b), highlighting areas where the model identifies T waves and abnormal TP intervals, despite the absence of annotations indicating respiratory events. The SHAP map in this region emphasizes QRS complexes across different beats, showing variations in their amplitude and duration.

Additional examples with visualization details are provided in Section 2 of the supplementary material (Figs. S1–S3). Fig. S1 (a) illustrates that the model does not focus on isolated regions with SpO_2 desaturations over 3 %, where respiratory events are assumed not to occur. Similarly, Fig. S1 (b) demonstrates how, in the ECG signal, the model does not highlight the region corresponding to the SpO_2

desaturation event in Fig. S1 (a), even though it is not associated with an apneic event. Figs. S2 (a, b) present SHAP visualizations related to SpO_2 and ECG signals, respectively, in cases where the model made incorrect predictions, with specific regions highlighted in green. Due to the overlap between signal segments associated with apneic events and the presence of artifacts, the model cannot identify clear patterns that distinguish event from non-event zones. Instead, it primarily detects noise, which prevents it from accurately recognizing event occurrences. As a result, the model does not find key patterns in certain regions, ultimately leading to an underestimation of the AHI. In Fig. S3, the SHAP method detects regions in both SpO_2 and ECG signals that exhibit significant patterns, even though no respiratory events are assumed to occur. The identification of relevant features in non-event regions causes the model to overestimate the AHI. Specifically, in Fig. S3 (b), the model recognizes areas of SpO_2 desaturation. Similarly, in Fig. S3 (d), SHAP detects ECG variations in the amplitude of the QRS complex as well as changes in TP segment durations.

5. Discussion

This study presents an innovative and explainable DL approach that integrates ECG and SpO_2 data to directly estimate the AHI for each patient, enabling the assessment of pediatric OSA severity. Notably, this is the first study to develop an interpretable DL-based model combining ECG and SpO_2 to clarify how predictions are made and to identify key SpO_2 and ECG patterns contributing to the decisions of the model. Using SHAP analysis, we have uncovered significant patterns related to both respiratory and cardiac activity by using full overnight SpO_2 and ECG signals, offering valuable insights into their role in pediatric OSA severity estimations. This approach enables the direct use of ECG and SpO_2 signals from nocturnal PSG recordings to estimate AHI and OSA severity, effectively reducing both the time and cost associated with traditional diagnostic methods. Furthermore, the combined DL model processes these signals through ECG-CNN and SpO_2-CNN architectures and the subsequent fusion module after the stacking strategy, while also capturing the temporal distribution of respiratory events throughout the night.

5.1. Configuration of the combined DL approach

To date, the architecture proposed in the present study has not been explored in either adult or pediatric OSA populations. Regarding the architecture selection, it is noteworthy that two separate CNNs are used for each signal (ECG-CNN and SpO_2-CNN architectures), ensuring that each signal is processed in a specific way. Thus, the proposed stacking-based multimodal architecture was designed to capture the distinct temporal characteristics of each physiological signal while ensuring computational efficiency [17]. It is important to note that by processing each signal separately, the model preserves the original sampling rates from the ECG and SpO_2, as well as their physiological integrity, avoiding distortions that could arise from resampling or signal alignment. The independent feature extraction enables each branch to learn modality-specific temporal patterns, such as QRS complexes in ECG or desaturation events in SpO_2, while the subsequent feature-level fusion allows the model to integrate complementary information and model inter-signal relationships without increasing architectural and computational complexity. Additional blocks of CNNs could be useful after the stacking strategy if the concatenated features required further hierarchical refinement [17]. However, in this case, the concatenated representations already contain well-processed features, which are effectively leveraged by a fusion module to model interactions between ECG and SpO_2 signals. Compared to including more complex architectures like transformers or RNNs, the stacking-based module of CNNs combined with a fusion module offers a trade-off between performance and computational efficiency [56,57]. While RNNs excel at capturing long-term temporal dependencies, these dependencies are already learned within the independent ECG-CNN and SpO_2-CNN architectures [17,57]. Moreover, the stacking strategy, which concatenates CNN output features, inherently removes the temporal relationships, making the inclusion of RNNs unnecessary [17,50]. Finally, this DL approach is highly and easily adaptable and scalable. If additional biomedical signals or clinical variables were to be introduced in future studies, independent CNNs could be incorporated seamlessly, concatenating their outputs for further processing.

5.2. SHAP explainability and ECG-SpO_2 interpretation

This study introduces an innovative approach by integrating, for the first time, a DL-based model with an XAI method for the analysis of nocturnal SpO_2 and ECG signals to detect OSA. To the best of our knowledge, only four previous studies have explored the use of XAI techniques for pediatric OSA diagnosis based on cardiorespiratory signals [11,22,27,58]. One of these studies used a DL model with an attention mechanism using SpO_2 signals [22], while another study applied a DL approach with Gradient-weighted Class Activation Mapping (Grad-CAM) on SpO_2 and AF signals [27]. Consistent with this study, their models primarily focused on regions exhibiting over 3 % SpO_2 desaturation. In another study, authors combined FE techniques along with SHAP using categorical data, including desaturation index, HR-derived variables, and demographic information [58]. Finally, our previous study proposed a DL-based approach using Grad-CAM with ECG signals, aligning with this study in its focus on regions associated with cardiac rhythm changes, as well as its specific attention to P waves, T waves, and QRS complexes [11].

Fig. 5 shows how Shapley values increase with OSA severity for both ECG and SpO_2, indicating higher model importance for signal patterns in severe cases. ECG dominates at lower severity levels, while SpO_2 becomes more influential as severity increases, reflecting meaningful physiological interactions between respiratory and cardiac systems. This supports the notion that SpO_2 desaturations are accompanied by cardiac responses, consistent with prior evidence [13,14,59].

Another advantage of SHAP lies in its ability to provide quantitative, interpretable insights into model decisions. As shown in Table 2, model *Acc* improves with OSA severity, indicating that SHAP contributions are more reliable for severe cases, as these children would benefit most from accurate and objective diagnosis and treatment. Importantly, unlike methods such as Grad-CAM [23,40], SHAP provides both qualitative and quantitative interpretability, enhancing confidence in the reasoning of the model.

By focusing on specific aspects of OSA cases, the model effectively uses desaturation and hypoxemia patterns in the SpO_2 signal for AHI prediction, which reflects the physiological response to apneic events, where repeated drops in oxygen levels serve as key indicators [1,60]. As illustrated in Fig. 8 (a), severe cases exhibit higher Shapley values and denser red regions, consistent with recurrent oxygen desaturations. Fig. 7 (a) demonstrates how the model distinguishes OSA-related SpO_2 drops from artifacts, prioritizing pathophysiological ranges. In this sense, cyclical SpO_2 desaturations are closely linked to respiratory obstruction events and their severity, especially in children with severe OSA [12]. In children with mild and moderate OSA (Figs. 6 and 7), desaturation patterns remain relevant but less pronounced. Variations in amplitude and duration of these desaturations are shown in Fig. 6 (a), Fig. 7 (b), and Fig. 8 (a). Those phenomena provide insight into oxygen reduction and recovery times, which may reflect pulmonary functional reserve and the extent of obstruction [59]. Additionally, signal fragmentation and fast oscillations may be associated with ventilatory instability and cardiovascular strain [59], while prolonged recovery after hypoxemia may suggest impaired compensatory capacity. Moreover, severe and recurrent SpO_2 desaturations, measured through hypoxic burden, have been linked to metabolic disturbances and an increased risk of cardiovascular complications in pediatric OSA [59] and increased mortality in adults [61].

In addition, the model also identifies synchronized patterns between SpO_2 and ECG signals. As shown in Fig. 7 (c) and Fig. 7 (d), it focuses on ECG regions that coincide with clusters of apneic events, capturing their temporal dynamics rather than isolated features. Furthermore, attention to P and T waves and their intervals during respiratory events, as illustrated in Fig. 6 (c), Fig. 7 (e), and Fig. 8 (c), indicates its ability to recognize changes in atrial and ventricular repolarization, which could be linked to ventilation and respiratory effort [62]. These findings are consistent with evidence of P-wave and QT interval dispersion in pediatric OSA, especially in severe cases [63]. A prolongation in P wave duration could point to a delay in atrial conduction, a mechanism linked to atrial fibrillation in adult OSA [63]. Likewise, alterations in the T wave or ST segment, together with increased QT interval dispersion, reflect repolarization abnormalities associated with ventricular arrhythmias and elevated cardiovascular risk. These alterations have also been reported in severe pediatric OSA [59,62]. Alterations in QRS complexes duration and/or amplitude, as seen in Fig. 6 (c) and Fig. 8 (d), may reflect ventricular alterations that may lead to ventricular hypertrophy or even changes in ventricular geometry, both conditions of cardiovascular risk in pediatric OSA [3,64]. Moreover, the model's focus on such regions, as shown in Fig. 7 (e), suggests that it captures physiologically meaningful bradycardia–tachycardia patterns typical of apneic events [3].

Other interesting findings are shown in Figs. S1-S3 in the supplementary material. Fig. S1 (a) shows that the model disregards isolated desaturations unrelated to apneic events, distinguishing sleep-phase changes from pathological patterns [65]. This could explain why ECG contributes more than SpO_2 in distinguishing mild and no OSA, as subtle cardiovascular variations may precede desaturations. Conversely, Fig. S3 (b) shows SHAP highlighting SpO_2 desaturations that may not have been annotated by specialists or that could be linked to other pathologies, such as chronic obstructive pulmonary disease (COPD) [66]. Fig. S1 (b) and Fig. S3 (d) highlight ECG areas without annotated apneic events but with distinctive patterns, suggesting that the model could be using additional cardiac information to improve OSA detection. Overall, these results demonstrate that SHAP captures well-known cardiorespiratory patterns and associations between ECG and SpO_2. By learning such representations, the model may contribute to identifying children at elevated cardiovascular risk, expanding its potential clinical utility [1].

Finally, to evaluate the reliability of SHAP, we conducted a quantitative analysis comparing ECG and SpO_2 Shapley values during the presence and absence of annotated apneic events across OSA severities (see Fig. S4 and Fig. S5 of the supplementary material). For both ECG and SpO_2 signals, Shapley values were consistently higher in regions containing apneic events and near zero in non-event regions, indicating higher model feature relevance during physiological disturbances. Statistical analysis using the Mann–Whitney *U* test confirmed statistically significant differences between event and non-event occurrences across all severities in both signals (p-value < 0.01). These findings suggest that SHAP captures physiologically meaningful patterns rather than random fluctuations, reinforcing their validity and potential clinical relevance.

5.3. Diagnostic ability and comparison with previous studies

To the best of our knowledge, no prior studies have directly used SpO_2 and ECG data to estimate OSA presence and severity in children. Therefore, the present approach emphasizes comparing the results obtained in this study concerning previous research that evaluated SpO_2 and ECG independently.

Considering the diagnostic ability, as illustrated in the confusion matrices (see Fig. 4) and diagnostic metrics (see Table 2), the model achieved higher performance in the CHAT and PATS datasets compared to UofC. Nevertheless, the results in UofC remain noteworthy considering that the optimal stacking-based CNN model was trained and optimized exclusively using CHAT and PATS data. Moreover, substantial variability exists in PSG scoring among sleep technologists, which could have influenced the external evaluation of our DL approach across the UofC independent dataset.

Beyond this factor, other specific differences may help explain the observed variation in diagnostic performance. The mean AHI values were 5.16 e/h for CHAT, 3.32 e/h for PATS, and 9.30 e/h for UofC. Likewise, interquartile ranges differed, being 2.32 [5.11] in CHAT, 1.0 [2.60] in PATS, and 3.8 [7.76] in UofC. Thus, information reflects heterogeneity in disease severity distributions. Participant age also varied across datasets. CHAT included children aged 5–10 years, PATS included subjects aged 3–12 years, and UofC encompassed a broader range from 0 to 13 years. Additionally, the sampling frequencies of ECG and SpO_2 signals differed considerably. In CHAT, ECG ranged from 50–512 Hz and SpO_2 from 1–512 Hz. In PATS, ECG was 128 Hz and SpO_2 ranged from 10–200 Hz. In UofC, ECG ranged from 200–500 Hz and SpO_2 from 25–500 Hz. These variations are consistent with prior research reporting differences in diagnostic performance when sleep datasets differ in their clinical and technical characteristics [67].

To facilitate comparison with previous research, Table 3 provides an exhaustive comparative overview of prior studies that focused on SpO_2 and ECG for estimating the presence and severity of pediatric OSA, alongside the findings of the current study [11,15,16,20–22,36,58,68–74]. By comparing the studies based on DL techniques, Vaquerizo-Villar *et al.* developed a CNN using SpO_2 to estimate apneic events within 20-minute segments [21]. The present study analyzed 3,320 children, while Vaquerizo-Villar *et al.* used a slightly smaller dataset comprising 3,196 subjects. Notably, their datasets were used for both model validation and testing, whereas this study used the UofC database exclusively for external validation, enhancing robustness and generalizability. The present approach achieved a higher k_4 value (0.549 vs. 0.515) in CHAT. When comparing the results with the UofC database, the k_4 value was slightly lower (0.378 vs. 0.422). However, this study evaluated a different test dataset ($n = 980$ vs. $n = 392$), implementing a more extensive validation process. Finally, the previous study lacked interpretability.

In another study, Mortazavi *et al.* developed a CNN-RNN model using SpO_2 signals to estimate also apneic events in 20-minute segments [22]. Their study analyzed 844 PSGs from the CHAT database, whereas this study used 3,320 PSGs from three different databases. Moreover, their use of CHAT for training, validation, and testing limited the generalizability of their findings. Additionally, Mortazavi *et al.* incorporated an attention mechanism into their model to enhance interpretability. While attention-based methods can highlight relevant features within the input data, the present approach, integrated with SHAP, offers a more comprehensive interpretation of the decisions made by the model. SHAP not only identifies the most influential features contributing to a given prediction but also quantifies their impact, allowing for a deeper understanding of how SpO_2 and ECG signals in our study contribute to the AHI estimation. This level of interpretability enhances clinical applicability by providing more transparent insight.

In the context of ECG analysis, various studies have explored both conventional and advanced FE methods to assess pediatric OSA severity using cardiac signals beyond ECG. However, the most relevant comparison is with studies utilizing DL approaches, as previous research has demonstrated the superiority of DL over traditional FE techniques when using ECG for pediatric OSA diagnosis [11,20]. In an earlier study, we developed a CNN model to assess pediatric OSA severity using ECG signals from the same CHAT database [20]. The present approach obtained a higher k_4 value (0.549 vs. 0.373). Notably, the previous study lacked interpretability. In a more recent study [11], we used a CNN-RNN model to assess pediatric OSA severity using ECG signals. This approach outperformed the previous model, achieving a k_4 of 0.549 vs. 0.410 in CHAT and a k_4 of 0.378 vs. 0.335 in UofC. This study also improves robustness and generalizability by incorporating the public PATS database. Regarding model interpretability, the previous study implemented Grad-CAM to provide a localized and visual representation of the ECG

Table 3
State-of-the-art studies on using cardiorespiratory signals to establish pediatric OSA severity.

Study	Signal	ML approach/Validation/Model/ XAI	#Total children/ #Test set	AHI (events/ hour)	*Se* (%)	*Sp (%)*	LR^+	*Acc* (%)
Hornero et al. (2017) [36]	SpO_2	FE/-/MLP/-	4191/3602	1	84.0	53.2	1.79	75.2
				5	68.2	87.2	5.32	81.7
				10	68.7	94.1	11.64	90.2
Calderón *et al.* (2020) [72]	SpO_2 (ODI)	FE / 15-fold-cv / LR, AdaBoost / –	453/453 CHAT	5	62.0	96.0	–	79.0
Xu *et al.* (2019) [68]	SpO_2 (ODI)	FE/External validation/ MLP regression / – /-	432	1	95.3	19.1	1.18	79.6
				5	77.8	80.5	3.99	79.4
				10	73.5	92.7	10.07	88.2
Vaquerizo-Villar et al. (2021) [21]	SpO_2	DL/Holdout/CNN/-	3,196/312 CHAT	1	71.2	81.8	3.92	77.6
				5	83.7	100.0	N.D	97.4
				10	83.9	99.3	117.8	97.8
			3,196/392 UofC	1	90.8	36.4	1.43	80.1
				5	76.0	88.6	6.68	83.9
				10	79.5	95.8	18.90	92.3
			3,196/231 BUH	1	88.8	53.2	1.90	79.2
				5	61.1	93.7	9.72	83.5
				10	65.0	96.9	20.69	91.3
Mortazavi et al. (2024) [22]*	SpO_2	DL/3-fold-cv /CNN-RNN/Attention	844/253 CHAT	1	96.3	61.3	2.7	86.5
				5	77.8	97.2	29.9	93.3
				10	76.6	98.8	65.0	96.2
Shouldice et al. (2004) [15]	RR[a]	FE/Loo cv /QDA /-	50/25	1	85.7	81.80	4.7	84.0
Gil et al. (2010) [74]	PPG [a]	FE/-/QDA / –	21/21	>18 OSA <5 No OSA	87.5	71.40	3.1	80.0
Lázaro et al. (2014) [73]	PPG [a]	FE/loo cv/LDA/-	21/21	>18 OSA <5 No OSA	100.0	71.40	3.5	86.7
Garde et al. (2019) [71]	SpO_2 (ODI) + PRV (Spectral)	FE/loo-cv/LR (binary classification for each threshold)/-	207	1 5	80.0	65.0	75.0	N.D
				10	85.0	79.0	82.0	N.D
					82.0	91.0	89.0	N.D
Martín-Montero et al. (2021) [70]	HRV [a]	FE/-/MLP/ –	1738/757 CHAT, UofC	1	76.3	38.30	1.2	63.4
				5	62.5	84.20	4.0	81.0
				10	66.7	91.60	7.9	89.3
Martín-Montero et al. (2021) [69]	HRV [a]	FE/-/LDA/ –	1738/757 CHAT, UofC	1	85.5	35.38	1.3	74.6
				5	64.4	93.78	10.4	85.0
				10	53.7	97.67	23.1	91.6
Ye et al. (2023) [58]	SpO_2 (ODI) + HR	FE/Holdout/XGBoost/SHAP	3,139/628	1	90.3	100.0	N.D	90.4
				5	82.1	93.8	N.D	85.7
				10	84.8	92.1	N.D	89.8
Martín-Montero et al. (2023) [16]	HRV [a]	FE/Holdout/LSBoost/ LIME	1610/296 CHAT	1	90.8	23.40	1.2	80.1
				5	66.7	61.17	1.7	63.2
				10	40.0	92.03	5.0	84.1
García-Vicente et al. (2023) [20]	ECG	DL/Holdout/CNN/-	1610/299 CHAT	1	84.2	46.15	1.6	75.9
				5	76.7	91.39	8.9	87.0
				10	53.7	98.06	27.7	92.0
García-Vicente et al. (2025) [11]	ECG	DL/Holdout/CNN-RNN/Grad-CAM	2,655/299 CHAT	1	89.7	46.2	1.7	80.3
				5	72.2	93.8	11.6	87.3
				10	58.8	97.7	25.2	92.3
			2,655/64 CFS (external validation)	1	83.3	42.5	1.45	57.8
				5	66.7	100.0	N.D	98.4
				10	66.7	100.0	N.D	98.4
			2,655/981 UofC (external validation)	1	88.5	33.0	1.3	78.7
				5	66.3	89.2	6.1	79.7
				10	63.8	95.9	15.5	88.4
This model	**ECG + SpO_2**	**DL/Holdout/Stacked ensemble CNNs/SHAP**	**3,320/299 CHAT**	**1**	**88.5**	**56.9**	**2.1**	**81.6**
				5	**80.0**	**98.1**	**41.8**	**92.6**
				10	**80.5**	**97.3**	**29.7**	**95.0**
			3,320/153 PATS	**1**	**91.0**	**48.0**	**1.8**	**69.9**
				5	**92.0**	**96.1**	**23.6**	**95.4**
				10	**90.9**	**98.6**	**64.6**	**98.0**
			3,320/980 UofC (external validation)	**1**	**96.3**	**17.9**	**1.2**	**82.5**
				5	**84.0**	**78.1**	**3.8**	**80.5**
				10	**84.3**	**90.4**	**8.8**	**89.0**

RR: the period between two R peaks; PPG: photoplethysmography; HRV and PRV: heart and pulse rate variability; ECG: electrocardiogram; AHI: apnea-hypopnea index; OSA: obstructive sleep apnea; *Se*: sensitivity; LR^+: positive likelihood ratio. N.D: Not defined; FE: Feature engineering; MLP: Multilayer perceptron; cv: cross validation: Loo: leave one out; QDA and LDA: quadratic and linear discriminative analysis; LR: linear regression; ODI: Oxygen Desaturation Index; AdaBoost: Adaptive Boosting; XGBoost: Extreme Gradient Boosting. [a] Derived features from cardiac signals. *The results of Mortazavi et al. (2024) are obtained as the mean value of the metrics obtained in the different folds.

regions influencing model decisions. However, relying solely on Grad-CAM and ECG signals restricts the depth of interpretability. In contrast, this study incorporates SHAP, which quantifies the precise contribution of both ECG and SpO_2 signals to the predictions. This approach enables the explanation of individual predictions and offers a comprehensive analysis of overall model behavior, leveraging the complementary information provided by both physiological signals.

5.4. Limitations and future work

This study has several limitations that should be noted. First, we used three databases to develop and evaluate our model. Particularly, we used the CHAT and PATS datasets for model development and internal validation, and the UofC dataset for external validation. In this sense, the use of UofC to externally evaluate our proposal and the differences in signal sampling rate values, age range, and AHI distribution may have resulted in a lower diagnostic performance in this dataset. Thus, alternative strategies could be implemented to enhance the generalizability of our approach. Nonetheless, the DL model showed a higher diagnostic ability than the models previously proposed, in all the datasets. Moreover, the databases were annotated by different specialists, which may represent a limitation in terms of consistent model learning and its generalizability, particularly in the UofC database. This likely contributed to reduced *Acc* in the UofC external validation. However, this diversity also enhances objectivity by minimizing potential bias from any single annotator. In this sense, while these datasets provided valuable insights regarding pediatric OSA, prospective testing of the model in a broader range of databases, as well as real-world home-based studies, would be advantageous to evaluate its performance across various contexts and populations. Additionally, further validation in specific pediatric populations, such as individuals with Down syndrome or those with complex medical conditions who are also at high risk for OSA, could provide more targeted insights and improve the clinical applicability of the model.

With respect to the design of the model, future works may explore unifying both signals into a single input representation processed through 2D convolutional architectures, thereby enabling the model to better capture long-range and inter-segment temporal dependencies. However, these extensions should be carefully evaluated, considering their computational cost and scalability to ensure that the model remains efficient and practical for large-scale clinical datasets. Regarding the stacking-based DL approach, it could serve as a basis for future research. In this context, additional models could be integrated alongside databases containing other variables representing cardiorespiratory risk factors, such as hypertension, obesity, and genetic predisposition, to estimate both OSA and related comorbidities. Furthermore, while SHAP proved useful for interpreting the model and identifying ECG and SpO_2 patterns, future work should explore complementary XAI methods to improve the explainability, reliability, and generalizability in complex physiological contexts.

Our proposal has been validated in a laboratory setting, and the next phase will focus on testing the system with prospectively collected home-based data and in real clinical environments to evaluate its performance, integration into sleep unit workflows, and comparison with the PSG gold standard. A key step toward clinical deployment will be creating a user-friendly interface that presents SHAP outputs in a format aligned with medical workflows. This interface will display the automatic diagnosis (estimated AHI), protocol-based recommendations, and model annotations over the ECG and SpO_2 signals. Integrating these elements into a desktop application, with the option to export clinical reports, will help ensure that model interpretability supports practical diagnostic use and strengthens clinician confidence in AI-assisted pediatric OSA screening. In addition to technical and clinical validation, the future deployment of the proposed system will also require careful attention to ethical aspects related to data collection and patient protection. For potential clinical implementation, signal acquisition procedures would be designed to ensure complete safety for participants. Informed consent would be obtained from all newly recruited subjects for the use of their clinical, pulse oximetry, and ECG data. Recruitment would follow the ethical principles of the Declaration of Helsinki and the Council of Europe's Resolution on Human Rights and Biomedicine (CETS No. 195, 2005). All collected data would be anonymized, and researchers would not have access to identifiable patient information to guarantee privacy. Data processing would comply with the European General Data Protection Regulation (GDPR, EU 2016/679). These future steps would be guided by a data management plan based on the FAIR principles (Findable, Accessible, Interoperable, and Reusable), ensuring secure storage, controlled access, full traceability, and ethically compliant data reuse within the scientific community.

6. Conclusion

As far as we know, this is the first study to explore an interpretable model that uses a stacking strategy combining CNNs with SpO_2 and ECG input signals to directly predict AHI and assess OSA severity in children. Our approach has demonstrated improved diagnostic ability than previous studies, particularly in severe OSA. This is crucial as that population is closely linked to increased cardiovascular risk and other complications, such as impaired cognitive function. Furthermore, XAI results offered both visual and quantitative insights, identifying well-known respiratory and cardiac patterns related to OSA. Our quantitative XAI findings emphasized the distinct contributions of SpO_2 and ECG signals in diagnosing OSA. ECG appears to be more critical for identifying healthy populations and mild OSA, whereas SpO_2 assumes a primary role in detecting moderate and severe OSA. These findings suggest that the ECG may identify subtle cardiovascular variations that SpO_2 alone does not detect, especially in children with fewer respiratory events. Moreover, this distinction highlights the complementary nature of these signals, ultimately enhancing the model performance across varying OSA severity levels. In conclusion, an interpretable DL tool combining SpO_2 and ECG data emerges as a promising alternative to PSG, offering a fast, reliable, and objective diagnosis of OSA in children. Furthermore, the incorporation of XAI techniques increases model trust and supports clinical adoption by providing both quantitative and visual explanations of the model decisions.

CRediT authorship contribution statement

Clara García-Vicente: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Gonzalo C. Gutiérrez-Tobal:** Writing – review & editing, Writing – original draft, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Fernando Vaquerizo-Villar:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis. **Adrián Martín-Montero:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis. **David Gozal:** Writing – review & editing, Writing – original draft, Funding acquisition, Conceptualization. **Roberto Hornero:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

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.

Acknowledgements

This research is part of the project PID2023-148895OB-I00, funded by MICIU/AEI/10.13039/501100011033 and FSE+, and part of the project CPP2022-009735, funded by MICIU/AEI/10.13039/501100011033 and the European Union NextGenerationEU/PRTR. This research was also supported by the project "0043_NET4SLEEP_2_E", cofunded by the European Union through the Interreg VI-A Spain-Portugal Program (POCTEP) 2021-2027; and by "CIBER-Consorcio Centro de Investigación Biomédica en Red" (CB19/01/00012) through "Instituto de Salud Carlos III (ISCIII)", co-funded with European Regional Development Fund. 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). The Pediatric Adenotonsillectomy Trial for Snoring (PATS) study was supported by the U. S. National Institutes of Health, National Heart, Lung, and Blood Institute (1U01HL125307, 1U01HL125295). The National Sleep Research Resource was supported by the U.S. National Institutes of Health, National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

C. García-Vicente was supported by 'Ayudas para contratos predoctorales para la Formación de Doctores' grant (PRE2021-100792) from the "Ministerio de Ciencia, Innovación y Universidades". David Gozal was supported in part by NIH grants HL166617 and HL169266.

Ethical approval

The research adhered to the principles outlined in the Declaration of Helsinki. The original CHAT and PATS databases clinical trials are identified as NCT00560859 and NCT02562040, respectively. Written consent was obtained from all parents under the research protocol, which can be found in the supplementary material of Marcus *et al.* [33] for CHAT and in the supplementary material of Redline *et al.* [32] for PATS. Moreover, assent was given by children aged 7 or 8 and older in both studies. Regarding the UofC database, the research protocol was approved by the UofC Ethics Committee of the Comer Children's Hospital (#11-0268-AM017, #09-115-B-AM031, and #IRB14-1241), and informed consent was obtained from the legal guardians of all participants.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.measurement.2025.120259.

Data availability

The CHAT data is publicly available upon request at https://sleepdata.org/datasets/chat. The PATS data is publicly available upon request at https://sleepdata.org/datasets/pats.

References

[1] C.L. Marcus, L.J. Brooks, S.D. Ward, et al., Diagnosis and management of childhood obstructive sleep apnea syndrome, Pediatrics 130 (2012) e714–e755, https://doi.org/10.1542/peds.2012-1672.
[2] H.L. Tan, D. Gozal, L. Kheirandish-Gozal, Obstructive sleep apnea in children: a critical update, Nat. Sci. Sleep 5 (2013) 109–123, https://doi.org/10.2147/NSS.S51907.
[3] C. Guilleminault, R. Winkle, S. Connolly, et al., Cyclical variation of the heart rate in sleep apnoea syndrome. Mechanisms, and usefulness of 24 h electrocardiography as a screening technique, Lancet 323 (1984) 126–131, https://doi.org/10.1016/S0140-6736(84)90062-X.
[4] R. Tauman, D. Gozal, Obstructive sleep apnea syndrome in children, Expert Rev. Respir. Med. 5 (2011) 425–440, https://doi.org/10.1586/ers.11.7.
[5] R. Bhattacharjee, L. Kheirandish-Gozal, G. Pillar, D. Gozal, Cardiovascular complications of obstructive sleep apnea syndrome: evidence from children, Prog. Cardiovasc. Dis. 51 (2009) 416–433, https://doi.org/10.1016/j.pcad.2008.03.002.
[6] L. Kheirandish-Gozal, What is "abnormal" in pediatric sleep? Respir. Care 55 (2010) 1366–1374.
[7] H.L. Tan, M.L. Alonso Alvarez, M. Tsaoussoglou, et al., When and why to treat the child who snores? Pediatr. Pulmonol. 52 (2017) 399–412, https://doi.org/10.1002/ppul.23658.
[8] R.B. Berry, R. Brooks, C.E. Gamaldo, et al., The AASM manual for the scoring of sleep and associated events, American Academy of Sleep Medicine, Darien, IL, 2013.
[9] 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.
[10] G.C. Gutiérrez-Tobal, D. Álvarez, L. Kheirandish-Gozal, et al., 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.
[11] C. García-Vicente, G.C. Gutierrez-Tobal, F. Vaquerizo-Villar, et al., SleepECG-Net: explainable deep learning approach with ECG for pediatric sleep apnea diagnosis, IEEE J. Biomed. Heal Inform. 29 (2025) 1021–1034, https://doi.org/10.1109/JBHI.2024.3495975.
[12] O. Vitelli, M. Del Pozzo, G. Baccari, et al., Autonomic imbalance during apneic episodes in pediatric obstructive sleep apnea, Clin. Neurophysiol. 127 (2016) 551–555, https://doi.org/10.1016/j.clinph.2015.05.025.
[13] T. Penzel, J.W. Kantelhardt, R.P. Bartsch, et al., Modulations of heart rate, ECG, and cardio-respiratory coupling observed in polysomnography, Front. Physiol. 7 (2016), https://doi.org/10.3389/fphys.2016.00460.
[14] G. Aljadeff, D. Gozal, V.L. Schechtman, et al., Heart rate variability in children with obstructive sleep apnea, Sleep 20 (1997) 151–157, https://doi.org/10.1093/sleep/20.2.151.
[15] R.B. Shouldice, L.M. O'Brien, C. O'Brien, et al., Detection of obstructive sleep apnea in pediatric subjects using surface lead electrocardiogram features, Sleep 27 (2004) 784–792, https://doi.org/10.1093/sleep/27.4.784.
[16] A. Martín-Montero, P. Armañac-Julián, E. Gil, et al., Pediatric sleep apnea: characterization of apneic events and sleep stages using heart rate variability, Comput. Biol. Med. 154 (2023) 106549, https://doi.org/10.1016/j.compbiomed.2023.106549.
[17] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT Press, 2016.
[18] K. Cao, X. Lv, Multi-task feature fusion network for Obstructive sleep apnea detection using single-lead ECG signal, Meas. J. Int. Meas. Confed. 202 (2022) 111787, https://doi.org/10.1016/j.measurement.2022.111787.
[19] B. Ganguly, D. Dey, An improved time-frequency representation aided deep learning framework for automated diagnosis of sleep apnea from ECG signals, Meas. J. Int. Meas. Confed. 242 (2025) 116170, https://doi.org/10.1016/j.measurement.2024.116170.
[20] C. García-Vicente, G.C. Gutiérrez-Tobal, J. Jiménez-García, et al., ECG-based convolutional neural network in pediatric obstructive sleep apnea diagnosis, Comput. Biol. Med. 167 (2023) 107628, https://doi.org/10.1016/j.compbiomed.2023.107628.
[21] F. Vaquerizo-Villar, D. Alvarez, L. Kheirandish-Gozal, et al., A convolutional neural network architecture to enhance oximetry ability to diagnose pediatric obstructive sleep apnea, IEEE J. Biomed. Heal Inform. 25 (2021) 2906–2916, https://doi.org/10.1109/JBHI.2020.3048901.
[22] E. Mortazavi, B. Tarvirdizadeh, K. Alipour, M. Ghamari, Deep learning approaches for assessing pediatric sleep apnea severity through SpO2 signals, Sci. Rep. 14 (2024) 22696, https://doi.org/10.1038/s41598-024-67729-9.
[23] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, et al., 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.
[24] A. Chaddad, J. Peng, J. Xu, A. Bouridane, Survey of explainable AI techniques in healthcare, Sensors 23 (2023) 634, https://doi.org/10.3390/s23020634.
[25] F. Di Martino, F. Delmastro, Explainable AI for clinical and remote health applications: a survey on tabular and time series data, Springer, Netherlands, 2023.
[26] K. Shkileva, N. Zolotykh, Explainable artificial intelligence techniques in medical signal processing, Procedia Comput. Sci. 212 (2022) 474–484, https://doi.org/10.1016/j.procs.2022.11.031.
[27] J. Jiménez-García, M. García, G.C. Gutiérrez-Tobal, et al., An explainable deep-learning architecture for pediatric sleep apnea identification from overnight airflow and oximetry signals, Biomed. Signal Process. Control 87 (2024), https://doi.org/10.1016/j.bspc.2023.105490.
[28] Á.S. Alarcón, N.M. Madrid, R. Seepold, J.A. Ortega, Obstructive sleep apnea event detection using explainable deep learning models for a portable monitor, Front. Neurosci. 17 (2023) 1–19, https://doi.org/10.3389/fnins.2023.1155900.
[29] F. Vaquerizo-Villar, G.C. Gutiérrez-Tobal, E. Calvo, et al., An explainable deep-learning model to stage sleep states in children and propose novel EEG-related patterns in sleep apnea, Comput. Biol. Med. 165 (2023) 107419, https://doi.org/10.1016/j.compbiomed.2023.107419.
[30] Y. Shi, Y. Zhang, Z. Cao, et al., Application and interpretation of machine learning models in predicting the risk of severe obstructive sleep apnea in adults, BMC Med. Inform. Decis. Mak. 23 (2023) 1–15, https://doi.org/10.1186/s12911-023-02331-z.
[31] S. Redline, R. Amin, D. Beebe, et al., 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.
[32] S. Redline, K. Cook, R.D. Chervin, et al., Adenotonsillectomy for snoring and mild sleep apnea in children: a randomized clinical trial, JAMA 330 (2023) 2084–2095, https://doi.org/10.1001/jama.2023.22114.
[33] C.L. Marcus, R.H. Moore, C.L. Rosen, et al., A randomized trial of adenotonsillectomy for childhood sleep apnea, N. Engl. J. Med. 368 (2013) 2366–2376, https://doi.org/10.1056/NEJMoa1215881.
[34] G.Q. Zhang, L. Cui, R. Mueller, et al., The national sleep research resource: towards a sleep data commons, J. Am. Med. Inform. Assoc. 25 (2018) 1351–1358, https://doi.org/10.1093/jamia/ocy064.
[35] R. Wang, J.P. Bakker, R.D. Chervin, et al., Pediatric adenotonsillectomy trial for snoring (PATS): protocol for a randomised controlled trial to evaluate the effect of adenotonsillectomy in treating mild obstructive sleep-disordered breathing, BMJ Open 10 (2020) e033889, https://doi.org/10.1136/bmjopen-2019-033889.
[36] R. Hornero, L. Kheirandish-Gozal, G.C. Gutiérrez-Tobal, et al., 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.

[37] R.B. Berry, R. Budhiraja, D.J. Gottlieb, et al., 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. 08 (2012) 597–619, https://doi.org/10.5664/jcsm.2172.
[38] C. Iber, S. Ancoli-Israel, A. Chesson, S. Quan, The AASM manual for scoring of sleep and associated events: rules terminology and technical specification, American Academy of Sleep Medicine, Westchester, IL, 2007.
[39] A.I. Naimi, L.B. Balzer, Stacked generalization: an introduction to super learning, Eur. J. Epidemiol. 33 (2018) 459–464, https://doi.org/10.1007/s10654-018-0390-z.
[40] S. Ali, T. Abuhmed, S. El-Sappagh, et al., Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence, Inf. Fusion 99 (2023) 101805, https://doi.org/10.1016/j.inffus.2023.101805.
[41] D. Rothman, Hands-On Explainable AI (XAI) with Python: Interpret, visualize, explain, and integrate reliable AI for fair, secure, and trustworthy AI apps, (2020).
[42] F.R. Mashrur, M.S. Islam, D.K. Saha, et al., 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.
[43] A. Sheta, H. Turabieh, T. Thaher, et al., Diagnosis of obstructive sleep apnea from ECG signals using machine learning and deep learning classifiers, Appl. Sci. 11 (2021) 6622, https://doi.org/10.3390/app11146622.
[44] J. Jiménez-García, M. García, G.C. Gutiérrez-Tobal, et al., A 2D convolutional neural network to detect sleep apnea in children using airflow and oximetry, Comput. Biol. Med. 147 (2022), https://doi.org/10.1016/j.compbiomed.2022.105784.
[45] L. Sörnmo, P. Laguna, The electrocardiogram—a brief background, in: L. Sörnmo, Laguna PBT-BSP in C and NA (Eds.), Processing in Cardiac and Neurological Applications, Elsevier, Burlington, 2005, pp. 411–452.
[46] J.-W. Chen, S.-T. Lin, C.-Y. Wang, et al., A signal segmentation-free model for electrocardiogram-based obstructive sleep apnea severity classification, Adv. Intell. Syst. 5 (2023) 1–10, https://doi.org/10.1002/aisy.202200275.
[47] J.W. Chen, C.Y. Wang, C.C. Lin, et al., Predicting apnea-hypopnea index in patients with obstructive sleep apnea using unsegmented ECG-signal-based algorithms, IEEJ Trans. Electr. Electron. Eng. 18 (2023) 1550–1552, https://doi.org/10.1002/tee.23868.
[48] J.W. Chen, C.M. Liu, C.Y. Wang, et al., A deep neural network-based model for OSA severity classification using unsegmented peripheral oxygen saturation signals, Eng. Appl. Artif. Intell. 122 (2023) 106161, https://doi.org/10.1016/j.engappai.2023.106161.
[49] R.T. Brouillette, A. Morielli, A. Leimanis, et al., 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.
[50] D. Wolpert, Stacked generalization (stacking), Neural Netw. 5 (1992) 241–259.
[51] J. Cohen, A coefficient of agreement for nominal scales. Educ. Psychol. Meas., 20 (1960) 37–46. https://doi.org/Doi: 10.1177/001316446002000104.
[52] K. Zhang, Y. Zhang, M. Wang, A unified approach to interpreting model predictions scott, Nips 16 (2012) 426–430.
[53] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions. Adv. Neural Inf. Process Syst., (2017) 4766–4775. https://doi.org/Doi:10.48550/arXiv.1705.07874.
[54] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences. 34th Int Conf Mach Learn, 2017.
[55] R.R. Selvaraju, M. Cogswell, A. Das, et al., Grad-CAM: visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2020) 336–359, https://doi.org/10.1007/s11263-019-01228-7.
[56] Z. Ebrahimi, M. Loni, M. Daneshtalab, A. Gharehbaghi, A review on deep learning methods for ECG arrhythmia classification, Expert Syst. with Appl. X 7 (2020) 100033, https://doi.org/10.1016/j.eswax.2020.100033.
[57] H. Ismail Fawaz, G. Forestier, J. Weber, et al., Deep learning for time series classification: a review, Data Min. Knowl. Discov. 33 (2019) 917–963.
[58] P. Ye, H. Qin, X. Zhan, et al., Diagnosis of obstructive sleep apnea in children based on the XGBoost algorithm using nocturnal heart rate and blood oxygen feature, Am. J. Otolaryngol. 44 (2023) 103714, https://doi.org/10.1016/j.amjoto.2022.103714.
[59] S. Ebrahimian, S. Sillanmäki, S. Hietakoste, et al., Beat-to-beat cardiac repolarization lability increases during hypoxemia and arousals in obstructive sleep apnea patients, Am. J. Physiol. Circ. Physiol. (2024), https://doi.org/10.1152/ajpheart.00760.2023.
[60] R. Hornero, L. Kheirandish-Gozal, G.C. Gutiérrez-Tobal, et al., 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.
[61] A. Azarbarzin, S.A. Sands, K.L. Stone, et al., The hypoxic burden of sleep apnoea predicts cardiovascular disease-related mortality: the osteoporotic fractures in men study and the sleep heart health study, Eur. Heart J. 40 (2019) 1149–1157, https://doi.org/10.1093/eurheartj/ehy624.
[62] S. Solhjoo, M.C. Haigney, N.M. Punjabi, Sleep-disordered breathing destabilizes ventricular repolarization: cross-sectional, longitudinal, and experimental evidence, Hear Rhythm (2024), https://doi.org/10.1016/j.hrthm.2024.08.054.
[63] C. Kraikriangsri, A. Khositseth, T. Kuptanon, P-wave dispersion as a simple tool for screening childhood obstructive sleep apnea syndrome, Sleep Med. 54 (2019) 159–163, https://doi.org/10.1016/j.sleep.2018.09.032.
[64] R.S. Amin, T.R. Kimball, J.A. Bean, et al., Left ventricular hypertrophy and abnormal ventricular geometry in children and adolescents with obstructive sleep apnea, Am. J. Respir. Crit. Care Med. 165 (2002) 1395–1399, https://doi.org/10.1164/rccm.2105118.
[65] F. Vaquerizo-Villar, D. Álvarez, G.C. Gutiérrez-Tobal, et al., Accurate and interpretable deep learning model for sleep staging in children with sleep apnea from pulse oximetry. In: IFMBE Proceedings, 2024. pp 38–47.
[66] A.M. Andrés-Blanco, D. Álvarez, A. Crespo, et al., Assessment of automated analysis of portable oximetry as a screening test for moderate-to-severe sleep apnea in patients with chronic obstructive pulmonary disease, PLoS One 12 (2017) e0188094, https://doi.org/10.1371/journal.pone.0188094.
[67] H. Korkalainen, T. Leppanen, J. Aakko, et al., Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea, IEEE J. Biomed. Heal Inform. 24 (2019) 1, https://doi.org/10.1109/JBHI.2019.2951346.
[68] Z. Xu, G.C. Gutiérrez-Tobal, Y. Wu, et al., 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.
[69] A. Martín-Montero, G.C. Gutiérrez-Tobal, D. Gozal, et al., Bispectral analysis of heart rate variability to characterize and help diagnose pediatric sleep apnea, Entropy 23 (2021) 1016, https://doi.org/10.3390/e23081016.
[70] A. Martín-Montero, G.C. Gutiérrez-Tobal, L. Kheirandish-Gozal, et al., Heart rate variability spectrum characteristics in children with sleep apnea, Pediatr. Res. 89 (2021) 1771–1779, https://doi.org/10.1038/s41390-020-01138-2.
[71] A. Garde, X. Hoppenbrouwer, P. Dehkordi, et al., 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.
[72] J.M. Calderón, J. Álvarez-Pitti, I. Cuenca, et al., 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.
[73] J. Lazaro, E. Gil, J.M. Vergara, P. Laguna, Pulse rate variability analysis for discrimination of sleep-apnea-related decreases in the amplitude fluctuations of pulse photoplethysmographic signal in children, IEEE J. Biomed. Heal Inform. 18 (2014) 240–246, https://doi.org/10.1109/JBHI.2013.2267096.
[74] E. Gil, R. Bailon, J.M. Vergara, P. Laguna, PTT variability for discrimination of sleep apnea related decreases in the amplitude fluctuations of PPG signal in children, I.E.E.E. Trans. Biomed. Eng. 57 (2010) 1079–1088, https://doi.org/10.1109/TBME.2009.2037734.