SleepECG-Net: explainable deep learning approach with ECG for pediatric sleep apnea diagnosis

Clara García-Vicente, Gonzalo C. Gutiérrez-Tobal, *Member, IEEE*, Fernando Vaquerizo-Villar,
Adrián Martín-Montero, David Gozal, and Roberto Hornero, *Senior Member, IEEE*

***Abstract*— Obstructive sleep apnea (OSA) in children is a prevalent and serious respiratory condition linked to cardiovascular morbidity. Polysomnography, the standard diagnostic approach, faces challenges in accessibility and complexity, leading to underdiagnosis. To simplify OSA diagnosis, deep learning (DL) algorithms have been developed using cardiac signals, but they often lack interpretability. Our study introduces a novel interpretable DL approach (SleepECG-Net) for directly estimating OSA severity in at-risk children. A combination of convolutional and recurrent neural networks (CNN-RNN) was trained on overnight electrocardiogram (ECG) signals. Gradient-weighted Class Activation Mapping (Grad-CAM), an eXplainable Artificial Intelligence (XAI) algorithm, was applied to explain model decisions and extract ECG patterns relevant to pediatric OSA. Accordingly, ECG signals from the semi-public Childhood Adenotonsillectomy Trial (CHAT, n=1610) and Cleveland Family Study (CFS, n=64), and the private University of Chicago (UofC, n=981) databases were used. OSA diagnostic performance reached 4-class Cohen's Kappa of 0.410, 0.335, and 0.249 in CHAT, UofC, and CFS, respectively. The proposal demonstrated improved performance with increased severity along with heightened cardiovascular risk. XAI findings highlighted the detection of established ECG features linked to OSA, such as bradycardia-tachycardia events and delayed ECG patterns during apnea/hypopnea occurrences, focusing on clusters of events. Furthermore, Grad-CAM heatmaps identified potential ECG patterns indicating cardiovascular risk, such as P, T, and U waves, QT intervals, and QRS complex variations. Hence, SleepECG-Net approach may improve pediatric OSA diagnosis by also offering cardiac risk factor information, thereby increasing clinician confidence in automated systems, and promoting their effective adoption in clinical practice.**

***Index Terms*— Pediatric obstructive sleep apnea (OSA), deep learning (DL), eXplainable Artificial Intelligence (XAI), electrocardiogram (ECG), Gradient-weighted Class Activation Mapping (Grad-CAM).**

I. INTRODUCTION

PEDIATRIC obstructive sleep apnea (OSA) is a frequent respiratory disorder with a prevalence ranging from 1% to 5% [1]. It occurs when the upper airway is collapsible during sleep, causing repeated episodes of apnea (no airflow) and hypopnea (reduced airflow) [1], [2]. When untreated, pediatric OSA can lead to a multitude of end-organ morbidities, including long-term neurocognitive and behavioral deficits and decreased metabolic function [3]. It also induces impaired autonomic nervous system regulation, ventricular hypertrophy, and high blood pressure in both the pulmonary and systemic circulation. These conditions increase the risk of long-lasting cardiovascular disease that may manifest in both childhood and adulthood, while adversely impacting the overall health and quality of life of affected children [1], [3], [4].

OSA is typically diagnosed by an overnight polysomnography (PSG) test, performed in a specialized laboratory while the patient sleeps [1]. During the test, different biomedical signals, such as electrocardiogram (ECG), electroencephalogram (EEG), electromyogram (EMG), blood oxygen saturation (SpO_2), airflow (AF), and end-tidal capnography are recorded synchronously to determine the apnea-hypopnea index (AHI). This index is the rate of apneic and hypopneic events per hour of sleep (e/h) [5]. According to the guidelines of the American Academy of Sleep Medicine (AASM), apneas in children are defined as a decrease of ≥90% in AF signal for at least two respiratory cycles. In comparison, hypopneas are defined as a decrease of ≥30% in AF accompanied by at least a 3% reduction in SpO_2 or electroencephalographic arousal [6]. Following these criteria, once the AHI is established, it is used to measure the presence

---

[1] This work is part of the projects PID2020-115468RB-I00 and CPP2022-009735, funded by MCIN/AEI/10.13039/501100011033 and the European Union “NextGenerationEU”/PRTR. This research was also co-funded by the European Union through the Interreg VI-A Spain-Portugal Program (POCTEP) 2021-2027 (0043_NET4SLEEP_2_E), and by “CIBER-Consorcio Centro de Investigación Biomédica en Red” (CB19/01/00012) through “Instituto de Salud Carlos III”, co-funded with European Regional Development Fund, as well as under the project TinyHeart from 2022 Early Stage call. The Childhood Adenotonsillectomy Trial (CHAT) was supported by the National Institutes of Health (HL083075, HL083129, UL1-RR-024134, UL1 RR024989). The Cleveland Family Study (CFS) was supported by grants from the National Institutes of Health (HL46380, M01 RR00080-39, T32-HL07567, RO1-46380). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002). *(Corresponding author: A. Martín-Montero). First author: C. García-Vicente.*

C. García-Vicente was supported by ‘Ayudas para contratos predoctorales para la Formación de Doctores’ grant from the ‘Ministerio de Ciencia, Innovación y Universidades’ (PRE2021-100792). GC. Gutiérrez-Tobal was supported by a post-doctoral grant from the University of Valladolid. F. Vaquerizo-Villar was supported by a “Sara Borrell” grant (CD23/00031) from the ISCIII cofounded by the ‘Fondo Social Europeo Plus (FSE+)’. D. Gozal is supported in part by NIH grants HL166617 and AG061824.

C. García-Vicente, G. C. Gutiérrez-Tobal, F. Vaquerizo-Villar, A. Martín-Montero, and R. Hornero are with the Biomedical Engineering Group, Universidad de Valladolid, Av. Ramón y Cajal 7, 47003, Valladolid, Spain (phone: +34 983 423000, ext. 4713) (e-mail: clara.garciav@uva.es).

G. C. Gutiérrez-Tobal, F. Vaquerizo-Villar, A. Martín-Montero, and R. Hornero are with the Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina, (CIBER-BBN), Spain.

F. Vaquerizo-Villar is with the Department of Anaesthesiology, Hospital Clínico Universitario of Valladolid, Spain.

D. Gozal is with the Office of The Dean and the Department of Pediatrics, Joan C. Edwards School of Medicine, Marshall University, 1600 Medical Center Dr, Huntington, WV 25701 (email: gozal@marshall.edu).

and severity of OSA. Despite its effectiveness, PSG requires monitoring multiple biomedical signals, specialized facilities, and qualified personnel [1], [2]. This makes PSG an expensive, complex, and uncomfortable method with limited availability, particularly in pediatric facilities. As a result, the disease is often underdiagnosed or diagnosed late [1].

Over the last decade, researchers have been intensively exploring the feasibility of alternative automated methods to simplify the diagnosis of pediatric OSA [7]. Most studies have focused solely on signals such as SpO_2 and AF, thus limiting their attention to respiratory information associated with the disease [8]. However, it is essential to consider the ECG signal because of the strong interdependencies between the cardiovascular and respiratory systems during apneic events. These episodes have been linked to bradycardia-tachycardia patterns, highlighting the importance of incorporating the ECG signal while studying OSA [3], [9]. Additionally, OSA has been associated with an increased likelihood of developing cardiovascular morbidity that may continue and progress into adulthood, particularly when the condition remains untreated [3], [4]. In past studies, several proposals have investigated cardiac signals using feature-engineering (FE) techniques such as heart rate variability (HRV) or photoplethysmography (PPG) [10], [11].

Unlike traditional FE methods, deep learning (DL) approaches can handle the complexities of signals directly without the need for a prior characterization process [12]. We previously combined the use of nocturnal ECG data along with DL techniques based on a convolutional neural network (CNN) to estimate pediatric OSA [13]. In adults with OSA, CNNs and recurrent neural networks (RNNs) have been effectively proposed to analyze various cardiorespiratory signals derived from sleep studies, including ECG signals [14], [15]. On the one hand, CNNs allow the automatic identification of complex patterns in data [12], while RNNs were designed to capture temporal interdependences [16]. Combining both computational architectures, CNNs allow for automatically extracting features of apneic events, and RNNs allow capturing temporal and sequential relationships that are crucial to correctly identify the recurrence of apneic events and cardiac patterns throughout the nocturnal recording [12], [16]. In this regard, CNN-RNN architectures are a feasible option to directly address AHI estimation and OSA detection, as they can simulate the recurrence of cardiac patterns induced by apneic events in the ECG signal [17].

Even though advanced DL methods have shown promise in predicting OSA, their main drawback lies in the inherent requirement for more explainability of their outcomes [18]. This shortcoming is crucial, especially in healthcare, where professionals must understand the process and reasoning behind automated decisions to fully trust them. In this context, eXplainable Artificial Intelligence (XAI) techniques are essential to provide transparency and interpretability to complex computational models [18]. Specifically, we consider the XAI application to be particularly relevant when analyzing ECG in pediatric OSA context. The ability to discern the patterns present in the signal, those on which automatic algorithms focus, could offer important insights on how to determine the severity of OSA in children. Furthermore, it could generate new pathophysiological cardiovascular risk-related features linked to the ECG signal in pediatric OSA. One of the most popular XAI techniques used to analyze biomedical signals is Gradient-weighted Class Activation Mapping (Grad-CAM), which uses gradient information in the convolutional layers to identify regions of the highest importance in the input data, thereby informing the predictions of a CNN model [19]. Grad-CAM has already proven its usefulness in identifying physiological features related to apneic events and sleep stages in pediatric OSA, although no previous study has applied it to ECG data [20], [21].

Accordingly, our study presents two noteworthy novelties. Firstly, we aimed to develop a novel CNN-RNN based regression approach using the transfer learning technique to directly estimate the AHI with overnight single-lead ECG signals, thus enabling an automated and interpretable diagnosis of pediatric OSA presence and severity. One of the main reasons for selecting regression is the delay in cardiac manifestations concerning apneic events [3]. Secondly, we implemented the use of Grad-CAM to interpret a CNN-RNN approach fed with ECG signals from the pediatric population. This novelty facilitates the extraction of relevant ECG patterns, allowing for the establishment of the assumed link between the cardiovascular and autonomic systems and pediatric OSA. This aspect also serves as an intriguing research gateway aiming to assess the potential cardiovascular risk in pediatric OSA.

Thus, we hypothesized that our integrated SleepECG-Net model, along with the Grad-CAM technique, would simplify diagnosis and enhance interpretability. Consequently, the objectives of this study were twofold: 1) To evaluate a CNN-RNN regression approach using full-night ECG recordings for estimating AHI and establishing the severity of pediatric OSA, and 2) to apply Grad-CAM XAI algorithm as a method for interpreting the decisions taken by the model and identify relevant ECG patterns related to pediatric OSA.

II. Subjects And Signals

Three databases were used to conduct the present study. First, we used the upon-request public database Childhood Adenotonsillectomy Trial (CHAT), which contains 1610 valid ECG recordings from PSG studies performed in pediatric patients aged 5 to 9.9 years [22]. CHAT constitutes a randomized, multicenter, and single-masked design study that complied with the Declaration of Helsinki (clinical trial: NCT00560859) [22], [23]. Study recordings were partitioned into a training set (60%) to train the model, a validation set (20%) to adjust the optimal configuration, and a test set (20%) to evaluate model performance and interpret the results. This partitioning was conducted so that each subject was exclusively assigned to one of the sets, avoiding duplication. AHI values were used as labels for the input data in the algorithm.

The private database obtained from the Pediatric Sleep Unit of the Medicine Comer Children's Hospital from University of Chicago (UofC), USA, was also included [24]. This database included 981 sleep studies of children aged between 0 and 13 years, who were referred to the pediatric sleep laboratory following symptoms and complaints commensurate with clinically suspected OSA. The UofC Ethics Committee approved the research protocol (#11-0268-AM017, #09-115-B-AM031 and #IRB14-1241). Informed consent was obtained from the legal caretakers of all children. This de-identified database was only used for external validation of the model developed with CHAT training dataset. Therefore, the 981 ECG recordings in the UofC sample served as a test set.

Finally, the public database obtained from the Cleveland Family Study (CFS) Coordinating Center was used [25]. Detailed information can be obtained from the National Sleep Research Resource website (https://sleepdata.org/datasets/cfs). The CFS is the world's most extensive family study dedicated to the research of sleep apnea. The study covers 2284 individuals from 361 families, followed for up to 16 years over five visits (1990-2006). Visit 5 involved complete overnight PSGs used to extract ECG recordings. Study participants were divided into 9 categories according to their age range. Because our model is designed for pediatric OSA, ECG recordings corresponding to the first age category, encompassing ages 5 to 14 years, were selected (n=64). Parents were asked to complete a questionnaire for their children. This database was only used for external validation of the model.

PSG recordings from the databases were visually assessed by medical sleep specialists, who manually scored apneic and hypopneic events and established AHI following the scoring rules according to the AASM [5], [26]. Consequently, the AHI was used for diagnosing and determining the severity of OSA in children. For our study, we defined four OSA severity groups based on three common AHI thresholds (AHI = 1, 5, and 10 e/h), in agreement with previous pediatric OSA studies [24], [27]. Thus, the severity groups in this study included: no OSA (AHI<1 e/h), mild OSA (1≤AHI<5 e/h), moderate OSA (5≤AHI<10 e/h), and severe OSA (AHI≥10 e/h). In CHAT and CFS, the same scorers analyzed all recordings. Conversely, the UofC dataset replicated real-world circumstances by including many scorers as needed for the clinical operations of the sleep center and by having different scoring standards throughout the databases. Table I presents demographic and clinical data from all children included in this study.

III. METHODS

Fig. 1 shows a general methodological workflow. Our methodology involved implementing and evaluating an interpretable CNN-RNN architecture using single-lead ECG recordings ($S_1, \ldots, S_N$) to estimate the AHI per subject ($\hat{y}_1, \ldots, \hat{y}_N$). The model was fed with minimally preprocessed ECG signals. Each signal was initially processed into CNN blocks to obtain feature maps. Next, the CNN-derived feature sequences were used to feed the RNN layers to obtain the AHI estimation. Subsequently, an evaluation of the diagnostic ability of the proposed algorithm was conducted. CHAT dataset was used for SleepECG-Net training, hyperparameter tuning, and optimal architecture selection, as well as to evaluate the diagnostic capability of the model (training, validation, and test sets, respectively). UofC and CFS databases were used to externally validate the model developed and tested with CHAT. Finally, the Grad-CAM method was implemented to identify and interpret the regions where the model was fixed to make the AHI estimation.

Table I
Demographic and clinical information of the children under study.

	CHAT			UofC	CFS
	Training	**Validation**	**Test**	**Test**	**Test**
Subjects (n)	988 (61.37%)	323 (20.06%)	299 (18.57%)	981 (100%)	64 (100%)
Age (years)	7.00 [2.00]	7.00 [2.00]	6.90 [2.00]	6.0 [6.0]	11.34 [3.25]
Females (n)	477 (48.28%)	164 (50.77%)	161 (53.85%)	379 (38.63%)	31 (48.44%)
BMI (kg/m^2)	17.31 [5.92]	17.12 [6.25]	17.43 [6.04]	18.02 [5.86]	21.59 [8.56]
AHI (e/h)	2.64 [4.77]	2.45 [4.77]	2.32 [5.11]	3.8 [7.76]	0.57 [1.18]
No OSA(1) (n)	212 (21.46%)	67 (20.74%)	65 (21.74%)	173 (17.64%)	40 (62.5%)
Mild OSA(2) (n)	488 (49.39%)	167 (51.70%)	144 (48.16%)	401 (40.88%)	21 (32.81%)
Moderate OSA(3) (n)	159 (16.09%)	44 (13.62%)	49 (16.39%)	178 (18.14%)	0 (0%)
Severe OSA(4) (n)	129 (13.06%)	45 (13.93%)	41 (13.71%)	229 (23.34%)	3 (4.69%)

Data are presented as number (percentage) or median [interquartile range].
BMI: body mass index; AHI: apnea-hypopnea index; e/h: events/hour.
(1): AHI<1 (e/h); (2): 1<AHI<5 (e/h); (3): 5<AHI<10 (e/h); (4): AHI>10 (e/h).

A. Signal preprocessing

Following the guidelines established by the AASM, the ECG-II lead was obtained in obtained in CHAT, UofC, and CFS [5]. All databases were homogeneously preprocessed. The raw signals were resampled at 100 Hz, as previously reported [28], [29]. Subsequently, the continuous component was corrected by removing the signal mean in 30-second windows. A high-pass filter with a passband between 0.5 and 50 Hz was then applied to reduce noise and avoid the loss of essential frequency components, such as those related to the QRS complexes [30].

All ECG recordings were empirically adjusted to 8 hours, as this value yielded the highest performance in the validation set. Recordings with fewer samples were padded including zeros at the beginning of the signal, while those with more samples were reduced by removing samples from the beginning, as was implemented in previous OSA studies using unsegmented cardiorespiratory signals [31], [32]. Then, ECG signals were

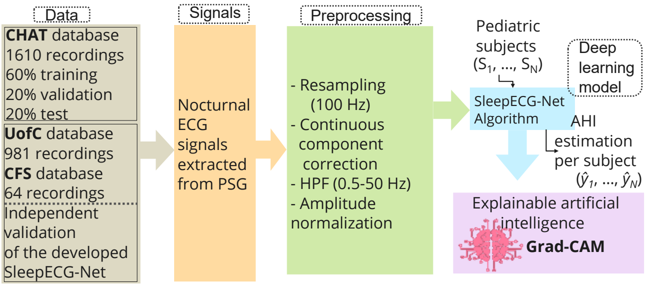
Data
CHAT database
1610 recordings
60% training
20% validation
20% test
UofC database
981 recordings
CFS database
64 recordings
Independent validation of the developed SleepECG-Net
Signals
Nocturnal ECG signals extracted from PSG
Preprocessing
- Resampling (100 Hz)
- Continuous component correction
- HPF (0.5-50 Hz)
- Amplitude normalization
Pediatric subjects ($S_1, \ldots, S_N$)
Deep learning model
SleepECG-Net Algorithm
AHI estimation per subject ($\hat{y}_1, \ldots, \hat{y}_N$)
Explainable artificial intelligence
Grad-CAM

Fig. 1: Proposed workflow for developing, validating, and explaining SleepECG-Net enabling prediction and interpretation of pediatric OSA severity using ECG signal recordings. HPF: High pass filter; S_N: subject N; $\hat{y}_N$: estimation of AHI in subject N.

resized to have the input dimensions allowed by the SleepECG-Net. This resulted in arrays of 48 segments, each lasting 10-min (48x60,000x1 = 2,880,000 samples). This dimension is suitable for block processing of CNNs encapsulated in time-distributed (TD) layers. The 10-min segments were the optimal length obtained to train the previously presented CNN model [13], being appropriate to detect clusters of apneic events that have a minimum duration of 10-min [33]. In addition, its dimension allowed the transfer of the optimal architecture, including its associated weights and layers from the CNN model presented previously to the CNN-related blocks used in this study. Finally, each segment was standardized [12]. As a data augmentation method, we randomized 3 times the 48 segments comprising ECG recordings in the CHAT training and validation sets [34].

B. SleepECG-Net architecture

Fig. 2 shows the main components of the architecture developed in this study. Below, the model process is outlined in detail, showing the flow of data through each stage:

Input data: Overnight ECG recordings.

#Step 1: CNN component (conv blocks automatically discern the intricate patterns within the ECG signal associated with apneic events). The conv part was implemented using a clustering of TD layers consisting of the previously presented CNN layers trained with 10-min ECG signals [13].

1. Initialize CNN parameters: number of filters (N_F), kernel size (K_{SIZE}), dropout probability (P_{CNN})
2. Load pre-trained weights from previous study for transfer learning.
3. For each 10-min ECG segment:

Pass through 14 conv blocks, each composed of:

   a. 1D-conv layer comprised of a set of filters (N_F) with a kernel size (K_{SIZE}) and zero padding.
   b. Batch normalization
   c. Rectified linear unit (ReLU) activation
   d. Max-pooling
   e. Spatial dropout with probability P_{CNN} [12].

4. Repeat this for a total of 14 conv blocks (last 2 blocks added and trained from scratch)
5. Flatten feature maps to 1D data for RNN processing.

#Step 2: RNN component (analyze the temporal distribution of apneic events in the nocturnal sequence by identifying characteristic ECG patterns).

6. Initialize bidirectional long short-term memory (BiLSTM) parameters: number BiLSTM units (U_{LSTM}), dropout probability (P_{LSTM})
7. Process each feature map from CNN output:
   - Pass through 2 BiLSTM layers:
   - Evaluate temporal dependencies in both forward and backward directions.
   - Output: refined sequence with temporal patterns detected by LSTM [12].
8. Apply dropout with probability P_{LSTM} after BiLSTM.
9. Output: temporal features from the BiLSTM layers

#Step 3: Pass BiLSTM output to a fully connected (FC) layer with a linear activation

Final output: AHI estimation per subject

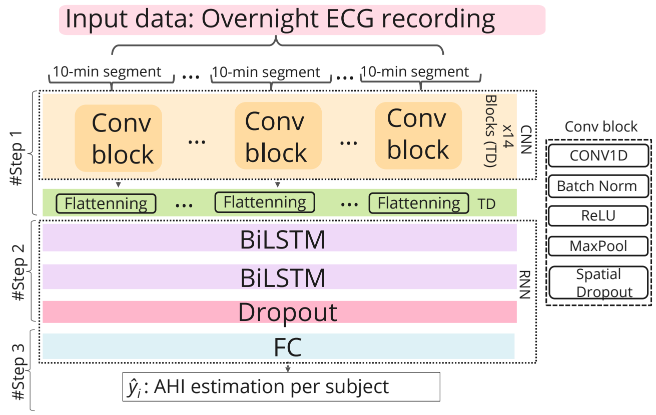

Fig. 2: Overall scheme of the regression model based on a CNN-RNN architecture proposed in the study. The input data to the CNN model consists of the complete nocturnal ECG recordings of each subject. RELU = rectified linear unit activation; TD = time distributed.

Specifically, an LSTM network is preferable over a basic RNN since it can maintain relevant data in the sequence and preserve it for several instances. Therefore, it maintains short-term memory capabilities similar to conventional RNNs while incorporating long-term memory functions. LSTM networks comprise several units (U_{LSTM}) that determine the output dimensionality and a dropout ratio with a certain probability (P_{LSTM}). LSTM was chosen instead of the gated recurrent unit (GRU) network because, in preliminary tests, higher performance was observed.

Notably, the proposal is an improved approach to the CNN-based model developed and validated in a previous study, where the number of events per 10-min ECG segment was estimated [13]. The optimal hyperparameters and weights of the CNN from that study were transferred to the current model, training and validating the SleepECG-Net as a transfer learning process. CNN allows the model to learn complex patterns of the ECG signal and RNN holds their temporal distribution throughout the night. This enables the model to retain relevant patterns across the time sequence while filtering out irrelevant ones.

C. SleepECG-Net optimization and algorithm assessment

To achieve optimal algorithm performance, we tuned a set of hyperparameters to minimize the generalization error of the SleepECG-Net model and find the optimal-performing configuration. Finally, an evaluation of the performance of the algorithm was conducted to determine the optimal hyperparameter configuration. For this purpose, 4-class Cohen’s kappa coefficient (k_4) was considered [35]. Within the validation set, we calculated the k_4 for subject-wise classification of OSA severity. The selection process relied on identifying the architecture with the highest k_4.

D. SleepECG-Net interpretability using Grad-CAM

As a subsequent step in implementing SleepECG-Net, XAI was applied using the Grad-CAM method. The purpose was to analyze and comprehend the internal mechanisms of the model in recognizing respiratory event-related information and discerning cardiac patterns linked to pediatric OSA [36]. The class activation mapping (CAM) technique emerged as an XAI method to identify critical areas in the input that significantly impact the output predicted by CNN used in image classification [36]. Effective implementation of CAM requires a specific architecture incorporating a Global Average Pooling layer to the final feature maps, followed by a fully connected final layer that generates the predictions [18]. Grad-CAM emerged in the context of improving the approach provided by CAM. This approach extends the CAM method by using gradients derived from convolutional layers, thus enabling the identification of relevant areas in the input that impact the final prediction. This method utilizes the gradient information flowing into a specific convolutional layer by providing gradient-based heatmaps. It offers detailed insight into discriminative regions significantly influencing SleepECG-Net decision-making [18], [36]. Heatmaps can be obtained for each of the convolutional layers by following the steps below. First, the gradients of the model output are determined concerning the feature maps of the s-th convolutional layer. The resulting gradients are averaged over all these feature maps according to the following expression [36]:

$$a_s = \frac{1}{T}\sum_k \frac{\partial \hat{y}}{\partial M_s^k} \quad (1)$$

where T is the number of filters, corresponding to 64 features maps in the s-th layer, $\hat{y}$ represents the output of the class of interest and M_s^k denotes the k-th feature map in the s-th layer. Subsequently, heatmaps are obtained by a gradient-weighted combination of the feature maps, which have been subjected to ReLU activation. This process is expressed as follows [36]:

$$L_{GradCAM} = ReLU\left(\sum_s a_s * M_s^k\right) \quad (2)$$

A heatmap of the exact dimensions as the corresponding convolutional layer feature maps is generated as a final step. Then, this heatmap is normalized and resized to support a joint display with the ECG signal. In this study, Grad-CAM was calculated using the gradients obtained from each convolutional layer to calculate a layer-specific heatmap. The final heatmap was acquired by averaging all normalized and resized heatmaps generated in each convolutional layer, as has been done in previous studies [20], [21]. It allowed obtaining more detailed representations with the contribution of all convolutional layers. This approach improves the identification of specific ECG patterns linked to pediatric OSA.

E. Statistical analysis and diagnostic performance

Test sets from CHAT, UofC, and CFS were used to estimate the AHI. UofC was used to validate the SleepECG-Net model independently from the database used to develop the model. To assess the concordance between the estimated AHI and the actual AHI, we computed the intraclass correlation coefficient (*ICC*) [37]. To evaluate the efficacy of the proposed algorithm in diagnosing pediatric OSA, subjects were assigned to one out of the four OSA severities based on their estimated AHI. Following the establishment of these categories, we computed the confusion matrix and 4-class accuracy (Acc_4). Moreover, k_4 was computed [35]. Finally, we computed the accuracy (*Acc*), sensitivity (*Se*), specificity (*Sp*), positive and negative predictive values (*PPV* and *NPV*), and positive and negative likelihood ratios (LR^+ and LR^-) for OSA thresholds (AHI = 1, 5, and 10 e/h).

IV. RESULTS

A. Optimal SleepECG-Net configuration

SleepECG-Net was trained on an NVIDIA GeForce RTX 4090 GPU. Training configuration utilized the He-normal technique [12]. We optimized weight updates using the adaptative moment estimation (Adam) method with an initial learning rate of 10^{-4} [12]. Training data was fed in batches of 64 samples over 400 epochs [12]. The mean squared logarithmic error was heuristically selected as the loss function to minimize the Adam algorithm in the validation set [12]. Additionally, early stopping was implemented to prevent overfitting.

Determining the optimal configuration for the model involved exhaustive training with all combinations of hyperparameter values, as outlined in Table II. SleepECG-Net training concluded with an early stop at epoch 57. A learning rate value of 2.5 x 10^{-4} was reached to find the desirable configuration of the model.

The convolutional layers in blocks$_{1\text{-}4}$ were composed of N_F = 16, K_{SIZE} = 33; blocks$_{5\text{-}8}$ were composed of N_F = 32, K_{SIZE} = 17, and blocks$_{9\text{-}12}$ were composed of N_F = 64, K_{SIZE} = 7. A dropout layer with a probability value P_{CNN} of 0.1 was applied in the last layer of all blocks. To optimize the CNN architecture, an exhaustive analysis was carried out. A detailed explanation of this process can be found in a previous study [13]. Then, additional convolutional blocks trained from scratch were included. We varied the number of convolutional blocks (N_{CNN}) in the range {1,2,3} to determine the optimal value. Experimentation showed that an N_{CNN} of 2 achieved the highest performance in the validation set (k_4 = 0.316, k_4 = 0.350, and k_4 = 0.344, adding 1, 2, and 3 layers, respectively). Accordingly, increasing the value of N_{CNN} would not lead to an improvement in capturing relevant patterns in the ECG. We also explored training the CNN from scratch for AHI estimation (k_4 = 0.328). However, higher performance was obtained using the transfer learning process (k_4 = 0.350). Moreover, to find the optimal number of LSTM layers (N_{LSTM}), we varied in the range {1,2,3}, obtaining the highest value of k_4 in the validation set using 2 layers (k_4 = 0.343, k_4 = 0.344, and k_4 = 0.325, adding 1, 2, and 3 layers, respectively). Thus, increasing the complexity of the model would not lead to better processing of the time domain for identifying ECG regions linked to respiratory events. Once the value of N_{CNN} and N_{LSTM} was selected, we implemented an exhaustive fitting strategy using the grid search method to optimize the remaining hyperparameters in the N_{CNN} and N_{LSTM} added layers. This involved testing all possible combinations of the hyperparameters within the defined search space. Specifically, the value N_F was varied in the range {64, 128, 256} for the additional convolutional blocks. Then, a search for K_{SIZE} in range {1:2:7} and P_{CNN} in range {0.0:0.1:0.5} was performed. Regarding BiLSTM layers, we varied U_{LSTM} in range {2:2:12} and P_{LSTM} in range {0.0:0.1:0.5}. Lastly, P_{DROP} was explored within the range {0.0:0.1:0.5} regarding the final dropout layer. The optimal hyperparameters of the 2 additional convolutional blocks were K_{SIZE} = 3, and P_{CNN} = 0.4. Concerning LSTM layers, the optimal hyperparameters were U_{LSTM} = 10, P_{LSTM} = 0.2, and P_{DROP} = 0.3. This configuration achieved the highest k_4 value (k_4 = 0.350) in the validation set

Table II
Search space of the SleepECG-Net hyperparameters, optimal configuration, and highest *4-class kappa* values obtained in the validation subset for each model with each specific hyperparameter.

Model	Hyperparameter	Search space	4-class kappa*	Optimal value
Additional CNN blocks	N_{CNN}	1, 2, 3	0.316,**0.350**,0.344	2
	N_F	64,128,256	**0.350**,0.322,0.323	64
	K_{SIZE}	1,3,5,7	0.349,0.350,0.343,0.344	3
	p_{CNN}	0.1,0.2,0.3,0.4,0.5	0.342, 0.346, 0.334,**0.350**, 0.338	0.4
LSTM	N_{LSTM}	1,2,3	0.343, **0.344**, 0.325	2
	U_{LSTM}	2,4,6,8,10,12	0.349,0.345,0.349,0.346,**0.350**,0.344	10
	P_{LSTM}	0.1,0.2,0.3,0.4,0.5	0.349,**0.350**,0.349,0.345,0.349	0.2
	P_{DROP}	0.1,0.2,0.3,0.4,0.5	0.343,0.349,**0.350**,0.349,0.343	0.3
4-class kappa			**0.350**	

N_{CNN}: number of conv blocks; K_{SIZE}: kernel size of conv layers; P_{CNN}: the probability of dropout layers in CNN blocks; N_{LSTM}: number of LSTM layers; U_{LSTM}: units in LSTM layers; P_{LSTM}: dropout probability of LSTM layers; P_F: the probability of the last dropout layer.
*Highest *kappa* value achieved by the model concerning each specific hyperparameter.

from CHAT. Consequently, it was chosen to evaluate SleepECG-Net in the test sets from CHAT, UofC, and CFS.

In preliminary strategies, several tests were carried out to obtain an optimal model and achieve the highest performance of the algorithm. Initially, a model based on the pre-trained CNN from the previous study, followed by a flattening layer encapsulated in a TD layer combined with an RNN network, was implemented. Different variants of RNN, including bidirectional gated recurrent unit and BiLSTM networks, were evaluated, and higher performance was achieved with a BiLSTM in the validation set (k_4 = 0.310 vs. k_4 = 0.261). These findings suggest that GRU intrinsic simplifications are suboptimal for the complexity of our data. Once the RNN architecture was optimized, preliminary tests were conducted to reduce overfitting during training and facilitate optimal model performance without applying data augmentation (k_4 = 0.310), by doubling (k_4 = 0.317) and tripling (k_4 = 0.343) the signals by randomizing the ECG segments of the input array. Further increases in data generation resulted in unapproachable computational costs. The highest value of k_4 = 0.343 in the validation set was obtained with the model trained by tripling the source data. Thus, applying data augmentation suggests the improvement of the performance of SleepECG-Net and generalization capability, as well as its ability to avoid overfitting and improve robustness.

To corroborate the suitability of the regression approach for our problem, we also implemented the SleepECG-Net architecture for a binary (presence or absence of OSA) and quaternary classification (presence and severity of OSA) task. Comparing these results with those obtained in our proposed regression approach, we found that, in the case of CHAT, in general terms the results with our approach were superior (Acc_4 = 61.54% vs. 56.52% and k_4 = 0.410 vs. 0.335). For UofC, the results with our approach were similar (Acc_4 = 53.82% vs. 53.21% and k_4 = 0.335 vs. 0.333). In the case of CFS, we obtained slightly higher values with this approach (Acc_4 = 56.25% vs. 57.81% and k_4 = 0.249 vs. 0.259). Regarding the binary classification results, if we compare the 2-class kappa (k_2) obtained with this model and those obtained with our proposal for the 5 e/h threshold, we found that in CHAT we obtained a higher performance (k_2 = 0.684 vs. 0.669), in UofC database the performance was slightly lower (k_2 = 0.571 vs. 0.582), and the performance was the same in CFS (k_2 = 0.792). Thus, AHI estimation is considered more reliable than direct classification methods that require fixed AHI thresholds. Relying solely on the classification agreement for model evaluation and optimization could lead to disadvantages if the AHI criteria change, which could result in less accurate AHI estimates and less information for clinicians.

Finally, several ablation tests were conducted to assess the contribution of different architectural components to the performance of the model. When the RNN was replaced with a feedforward network (FFN) while retaining the pretrained CNN and the N_{CNN} additional layers, the k_4 value on the validation set dropped to 0.295, compared to the proposed approach, which achieved a k_4 of 0.350. Further, when both the RNN and the last additional N_{CNN} were removed, the k_4 value decreased even more, reaching 0.283. Moreover, removing both the RNN and the two additional N_{CNN} led to a slight improvement, with the k_4 = 0.296. Finally, removing the last layers of the pre-trained CNN resulted in a reduced k_4 value of 0.286. These findings indicate that the additional N_{CNN}, when used in isolation, does not significantly enhance the model's performance. Instead, the features extracted by these blocks must be processed by the RNN module to unlock their full potential.

B. Diagnostic ability of the CNN-RNN approach

The agreement between the estimated and actual AHI was indicated with *ICC* of 0.76 in the CFS test set, being higher than in CHAT (*ICC*=0.73) and UofC (*ICC*=0.66). The *ICC* results suggest that SleepECG-Net has a moderate to acceptable predictive ability in all three databases. Furthermore, in global computation, the *ICC* ranges between 0.6 and 0.7 in the three databases, suggesting that the model generalizes well and is robust to different data samples [37]. Fig. 3 presents the confusion matrices obtained after classifying the severity of OSA for each subject based on their estimated AHI. The 4-class metrics obtained were Acc_4 = 61.54% and k_4 = 0.410 in the CHAT test set, Acc_4 = 53.82% and k_4 = 0.335 in the UofC test set, and Acc_4 = 56.25% and k_4 = 0.249 in the CFS test set. Analyzing the severity of subjects correctly classified, it is observed that in the three databases, SleepECG-Net presents the optimal performance for mild OSA, followed by severe OSA, no OSA, and moderate OSA. Table III reveals the diagnostic performance of pediatric OSA severity according to the conventional AHI cutoffs (1, 5, and 10 e/h) in CHAT, UofC, and CFS test sets. High to very high *Acc* are reached for 5 and

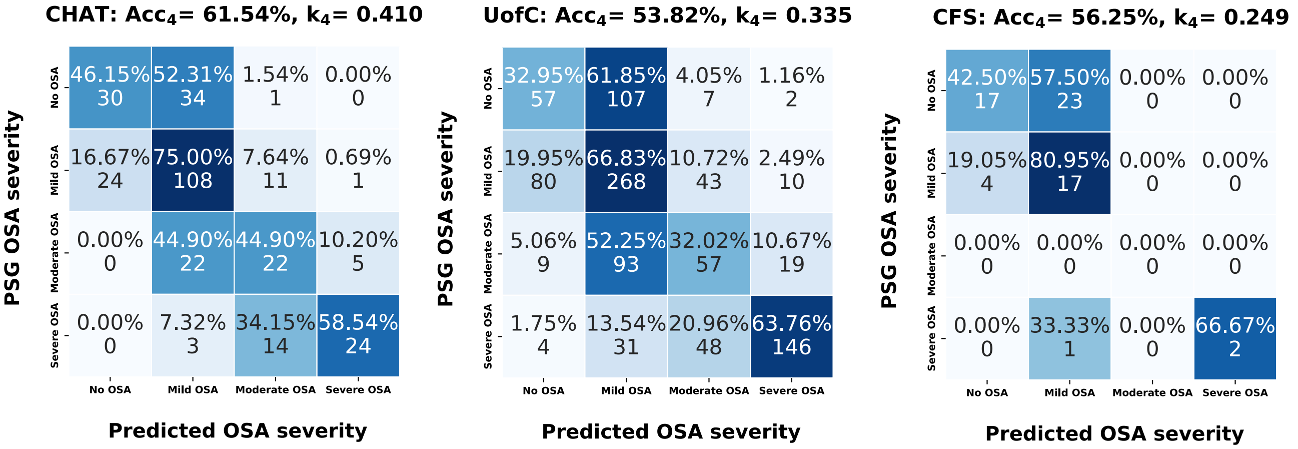

Fig. 3: Confusion matrices of the model in the CHAT, UofC, and CFS test sets. no OSA: AHI<1 (e/h); mild OSA: 1≤AHI<5 (e/h); moderate OSA: 5≤AHI<10 (e/h); severe OSA: AHI≥10 (e/h).

Table III
Diagnostic performance of the SleepECG-Net model in the test set.

AHI cutoff	**Test set**	***Se* (%)**	***Sp* (%)**	***PPV* (%)**	***NPV* (%)**	LR^+	***Acc* (%)**	***2class kappa***
1 e/h	CHAT	89.7	46.2	85.7	55.6	1.67	80.26	0.38
	UofC	88.5	33.0	86.0	38.0	1.32	78.70	0.23
	CFS	83.3	42.5	46.5	81.0	1.45	57.81	0.22
5 e/h	CHAT	72.2	93.8	83.3	88.7	11.61	87.29	0.69
	UofC	66.3	89.2	81.3	78.9	6.14	79.71	0.57
	CFS	66.7	100	100	98.4	N. D	98.44	0.79
10 e/h	CHAT	58.5	97.7	80.0	93.7	25.17	92.31	0.63
	UofC	63.8	95.9	82.5	89.7	15.47	88.38	0.65
	CFS	66.7	100	100	98.4	N. D	98.44	0.79

Se (sensitivity); *Sp* (specificity); *PPV* and *NPV* (positive and negative predictive value); LR^+ and LR^- (positive and negative likelihood ratio); N.D (not defined).

10 e/h cutoffs. Moreover, increasing performance with OSA severity is observed, with SleepECG-Net reaching 92.31%, 88.38%, and 98.44% *Acc* for 10 e/h in CHAT, UofC, and CFS respectively. Based on the established criteria for interpreting *k* values [38], we can conclude that for 5 and 10 e/h cutoffs we have moderate to substantial agreement in the three databases.

C. *Identification of ECG patterns using Grad-CAM*

The Grad-CAM method was implemented after conducting the diagnostic evaluation of SleepECG-Net. Fig. 4, 5, and 6 illustrate the heatmaps obtained using Grad-CAM on various examples of ECG signals, showing relevant patterns for accurate AHI estimation. On the one hand, Fig. 4 (a), Fig. 5 (a), and Fig. 6 (a.1, b.1) correspond to heatmaps and examples of full-night ECG signals. On the other hand, Fig. 4 (b, c), Fig. 5 (b, c, d), and Fig. 6 (a.2, b.2) show a zoom of relevant regions extracted from the ECG signals. In all heatmaps, the annotations of the presence and absence of respiratory events obtained from the PSG are highlighted in red with overlapping dotted lines.

In Fig. 4 (a), Grad-CAM highlights the region containing a cluster of apneic events. Fig. 4 (b) identifies areas characterized by heart rate (HR) variations evidencing bradycardia-tachycardia pattern. Furthermore, in Fig. 4 (c), it is observed how Grad-CAM identifies regions in event transition zones encompassing PQ and QT segments, and areas comprising TP segments. Regarding Fig. 5 (a), Grad-CAM highlights regions where clusters of respiratory events exist. Fig. 5 (b) shows a delay of regions when such events are manifest. Fig. 5 (c) and Fig. 5 (d) identify with dark color the PQ, QT, and TP segments illustrated in Fig. 4. Moreover, different patterns are also noted. Fig. 5 (c) shows relevant areas where the presence of the U wave is discernable. Finally, the heatmap emphasizes the QRS complexes of different beats in Fig. 5 (d). Fig. 6 (a.1, b.1) highlights areas where the model identifies relevant patterns, although in those regions, no annotations indicating the presence of respiratory events were provided. Fig. 6 (a.2) shows regions with changes in the amplitude of QRS complexes. In Fig. 6 (b.2), changes in HR can be seen, although they are not associated with the presence of annotated events. Fig. 7 (a, b) shows heatmaps associated with ECG signals in which SleepECG-Net made a wrong prediction, along with a zoom of relevant regions. In Fig. 7 (a), the method finds regions of decreasing QRS complex amplitude and changes in HR in regions where respiratory events are assumed not to occur. Identifying relevant patterns in regions without events leads SleepECG-Net to overestimate AHI. In Fig. 7 (b), due to the dense occurrence of adjacent events, the model does not find distinctive patterns to discriminate between event and non-event zones. This fact leads the model to incorrectly identify regions where events occur, resulting in underestimation of the AHI.

V. DISCUSSION

This study introduces a novel approach to evaluate an interpretable CNN-RNN model using nocturnal ECG signals to directly estimate the AHI per subject and thus determine pediatric OSA severity. This is the first study using an interpretable DL model focused on explaining the decision of the model and interpreting relevant ECG patterns. It is noteworthy that using single-channel ECG signals allows the use of nocturnal recordings from PSG to estimate OSA severity, reducing the time and cost of diagnosis. Moreover, SleepECG-Net allows the extraction of intricate ECG patterns through CNN while determining the temporal distribution of respiratory episodes in the nocturnal sequence using RNN. The findings underscore the potential of using a DL approach with ECG signals for accurately establishing the severity of pediatric OSA. Grad-CAM facilitates the identification of cardiac

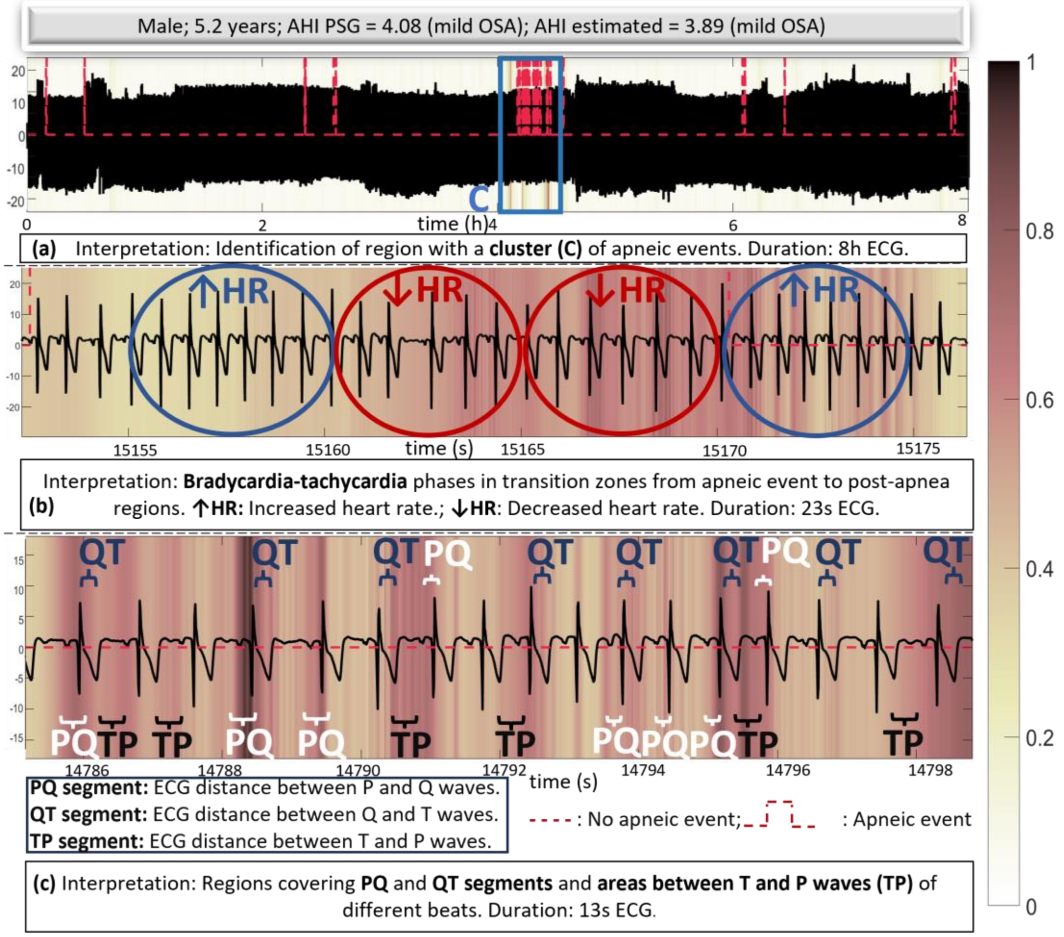

Fig. 4: Grad-CAM visualizations of some representative findings in an ECG signal of the CHAT dataset. Fig. 4 (a) shows the heatmap of the nocturnal ECG signal. Fig. 4 (b) and 4 (c) are zooms of the cluster (C) marked in Fig.4 (a). Zones of change in heart rate (HR) increase and decrease (↑HR and ↓HR) are indicated in circles. The color bar indicates at 0 (yellow) the zones of lower relevance and at 1 (brown) the zones of higher relevance.

patterns linked to pediatric OSA. This approach could serve as a valuable starting point to assess risk of cardiovascular co-morbidities, a clinically relevant issue in children with OSA [1].

A. *Configuration of SleepECG-Net approach*

This study implemented a novel regression-based model using a combination of CNN and RNN to directly estimate the AHI and subsequent pediatric OSA severity by analyzing nocturnal ECG recordings. This type of architecture has been previously evaluated in adults [14], [15]. However, in the pediatric OSA context, a CNN-RNN approach had only been used once and applied to AF and SpO_2 signals, but not to ECG tracings [20]. Our approach demonstrates a moderate to high level of accuracy in predicting pediatric OSA, thus revealing its diagnostic usefulness.

Nocturnal ECG recordings with dimensions of 48 x 60000 are suitable for block processing of CNNs encapsulated in TD layers, as demonstrated in a prior study [13]. This dimensional choice enabled the transfer of the optimal architecture, along with its associated weights and layers, from the model used in the previous study to the CNN-related blocks used in this research. The transfer learning process using a fine-tuning approach was performed by training the model, leveraging previously acquired knowledge while reducing model training time and complexity.

Finally, although training the SleepECG-Net model is computationally expensive, the training process is performed offline**,** allowing the trained model to be efficiently tested in real-time (online) with reduced computational cost ($6\text{x}10^{-2}$ seconds per subject).

B. *Diagnostic performance*

Looking at the confusion matrices, it can be seen that 98.5% (CHAT), 94.8% (UofC), and 100% (CFS) of no OSA patients have an estimated AHI (AHI_{est}) <5 e/h (no OSA or mild OSA). In addition, of the subjects with actual AHI (AHI_{PSG}) <5 e/h, 93.8% (CHAT), 89.2% (UofC), and 100% (CFS) were estimated as AHI_{est}<5 e/h. Additionally, 99.5% (CHAT), 97.9% (UofC), and 100% (CFS) of subjects belonging to the no OSA or mild OSA were estimated with an AHI_{est}<10 e/h. Finally, 100% (CFS), 93.2% (UofC), and 96.6% (CHAT) predicted as severe OSA ($AHI_{est} \geq 10$ e/h) are at least moderate OSA patients. Hence, a possible screening protocol can be derived to show the clinical usefulness of our proposal as follows: *i)* If AHI_{est}<1 e/h, discard the presence of OSA because most of these patients (91.3% in UofC, 100% in CFS, and 100% in CHAT) will have an AHI_{PSG}<5 e/h. If symptoms persist, these children may be eventually referred to PSG [39]; *ii)* if $1 \leq AHI_{est}$<5 e/h, suggest PSG since doubts arise about the actual diagnosis of the patients; *iii)* if $5 \leq AHI_{est}$<10 e/h, consider treatment, since most probably (95.5% in UofC, and 97.9% in CHAT) these subjects

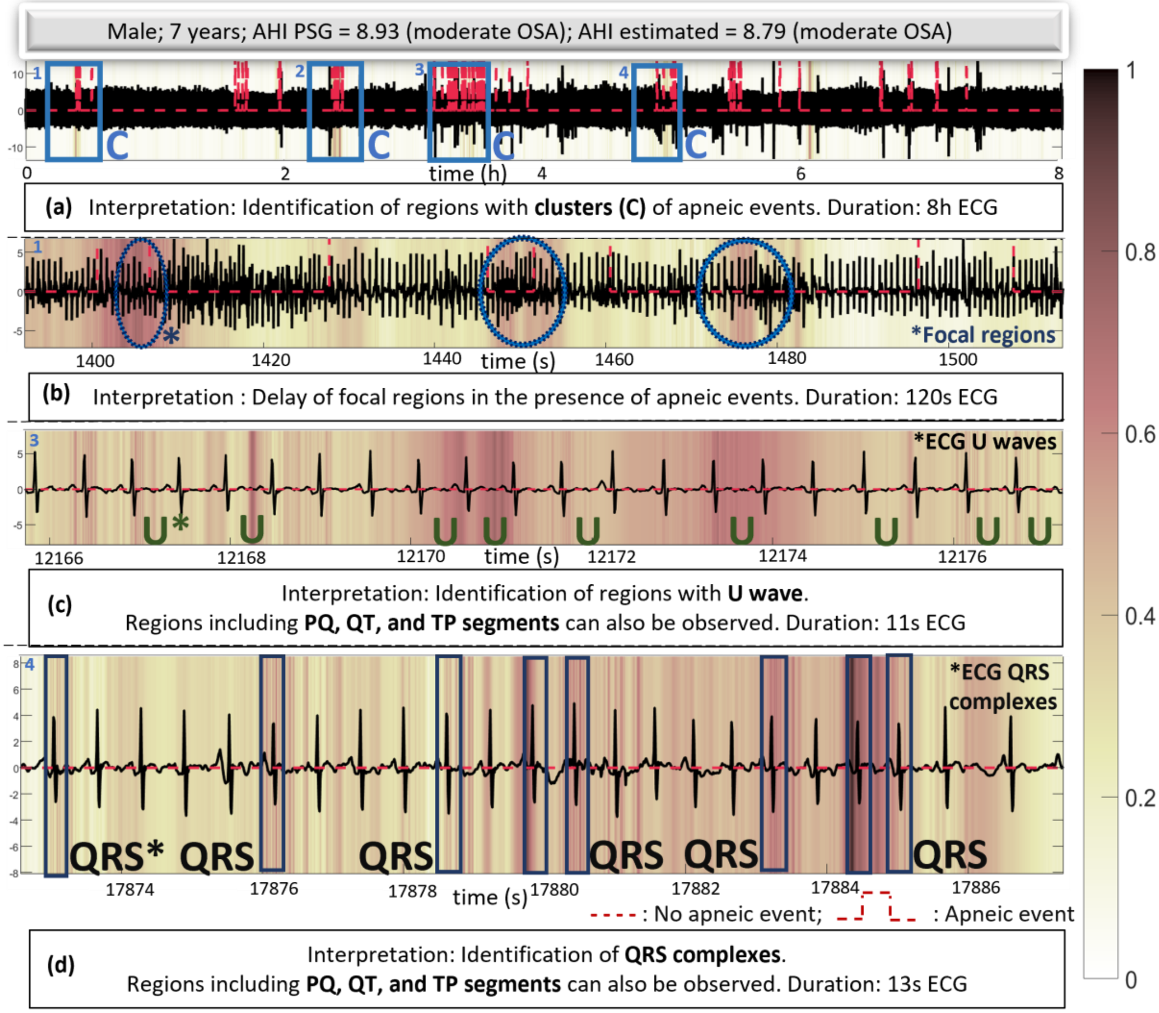

Fig. 5: Grad-CAM visualizations of some representative findings in different regions of one single ECG signal of the CHAT test set illustrating correct predictions. Fig. 5 (a) shows the heatmap of the overnight ECG signal. Fig. 5 (b), 5 (c), and 5 (d) are zooms of the clusters (C) 1, 3, and 4, as noted in Fig. 5 (a). U: ECG U wave; QRS: ECG QRS complex. The color bar indicates at 0 (yellow) the zones of lower relevance and at 1 (brown) the zones of higher relevance.

have at least a mild OSA. This threshold does not apply to the CFS database, because there are no patients with moderate OSA; *iv)* if $AHI_{est} \geq 10$ e/h, suggest treatment since most of these children (93.2% in UofC, 96.7% in CHAT, and 100% in CFS) have an $AHI_{PSG} \geq 5$ e/h. Additionally, it should be considered a further observation of these patients since they are likely to have residual OSA after treatment [1]. This screening protocol would present an innovation in the context of pediatric OSA diagnosis. In this respect, this protocol would avoid the need for 49.1% (UofC), 44.1% (CHAT), and 35.9% (CFS) of complete PSGs. Moreover, only 1.5% (CHAT), 5.2% (UofC), and 0% (CFS) of children with an $AHI_{PSG} < 1$ e/h would be indicated for treatment and 0% (CHAT), 3.2% (UofC), and 0% (CFS) of children with an $AHI_{PSG} > 5$ e/h would not be referred to PSG/treatment in the first visit to the specialist. This solution helps reduce waiting lists and medical costs for diagnosing OSA, while also offering a more appropriate diagnostic procedure for children.

C. *SleepECG-Net explicability using Grad-CAM*

This study introduces a novel approach by presenting, for the first time, the combination of a DL model and the application of an XAI method with nocturnal ECG signals to detect pediatric OSA. Analyzing the XAI results, it seems obvious that SleepECG-Net uses the bradycardia-tachycardia patterns (Fig. 4 (b) and Fig. 6 (b.2)) as these are well-known physiological responses of the heart to respiratory events, particularly when such events are obstructive in nature [3]. Similarly, the use of information from QT segments, T, and P waves (Fig. 4 (c) and Fig. 5 (d)) coincides with evidence on P wave and QT interval dispersion in pediatric OSA, mainly in the most severely affected cases [40]. Prolonged P wave duration could indicate a delay in atrial conduction related to the pathophysiological mechanisms contributing to the development of atrial fibrillation in adult OSA [40], [41]. Concerning the T-wave, this presents a contractile property associated with increased HR. However, abnormalities in this wave and changes in the ST segment could suggest possible cardiac alterations [30], [41], [42]. Likewise, QT interval dispersion may be associated with an increased risk of ventricular arrhythmia and be linked to a higher probability of sudden death [30], [43], [44]. These facts have also been evidenced in severe pediatric OSA [45]. Moreover, Grad-CAM focus on the U wave (Fig. 5 (c)) is supported by evidence indicating its dependence on HR [46]. The prominent presence of this wave can be correlated with bradycardia and long QT syndrome, both clinical conditions documented in pediatric OSA [3], [40]. In addition, Grad-CAM

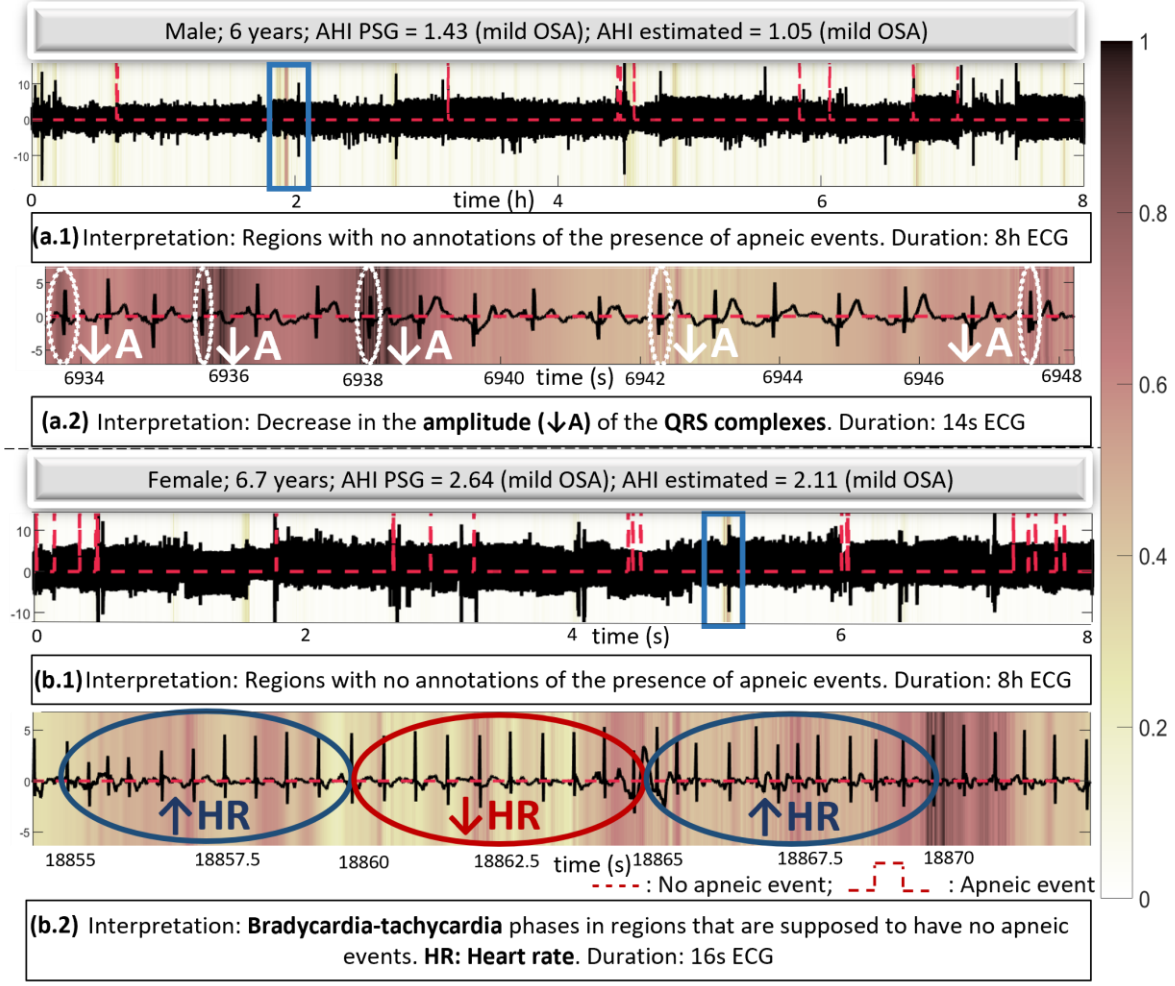

Fig. 6: Grad-CAM visualizations of some representative findings in an ECG signal of the CHAT test set illustrating accurate predictions. Fig. 6 (a.1) and 6 (b.1) show the heatmaps of two different nocturnal ECG signals. Fig. 6 (a.2) and 6 (b.2) are zooms of the areas marked in blue rectangles. Zones of change in HR increase and decrease (↑HR and ↓HR) are indicated in circles. Colorbar indicates at 0 (yellow) the zones of lower relevance and at 1 (brown) the zones of higher relevance. ↓A: Decrease in the amplitude of QRS complexes. D: ECG segment duration; h: hours; s: seconds.

focus on QRS duration and amplitude (Fig. 5 (d) and Fig. 6 (a.2)) responds to evidence of the occurrence of cardiac arrhythmias related to pediatric OSA due to changes in HR in the form of bradycardia-tachycardia patterns [3], [45]. Alterations in the QRS complexes are likely related to ventricular hypertrophy or even altered ventricular geometry [47], [48]. This condition increases the risk of cardiovascular abnormalities in pediatric OSA, preferentially in the most severe cases [3], [4].

Taken together, Grad-CAM results show that SleepECG-Net is focusing on not only well-known cardiac OSA-pediatric patterns but also on ECG patterns coherent with the risk of cardiovascular disease, thus paving the way towards the identification of pediatric OSA instances in which a higher cardiac risk may be present.

D. Comparison with previous studies

Focusing on pediatric OSA, several studies have evaluated conventional and advanced FE methods to estimate pediatric OSA severity by analyzing cardiac signals other than ECG. *Gutiérrez-Tobal et al.* [8] conducted a systematic review which noted that most ML research on pediatric OSA diagnosis has centered around SpO_2 signals. Their review, which did not include any studies based on ECG signals, conducted a meta-analysis to compile *Se* and *Sp* metrics from 19 studies. While comparisons between different biological signals should be approached cautiously, a comparison with the performance metrics in our proposal is still valid. This study reported diagnostic performance at 1, 5, and 10 e/h, with *Se* values of 84.9%, 71.4%, and 65.2%, respectively, and *Sp* values of 49.9%, 83.2%, and 93.1%. When compared to our model, SleepECG-Net achieved a higher *Se* at the 1 e/h cutoff in CHAT and UofC. At the 5 and 10 e/h cutoffs, the *Sp* values of our model were higher in CHAT, UofC, and CFS than those in the meta-analysis, as well as the *Se* in CHAT and CFS at 5 e/h and 10e/h, respectively. A comparison between the current study and previous studies is presented in Table IV. Our DL approach exhibited higher *Se* for CHAT and UofC when compared to Shouldice et al. [9] (88.5-89.7% vs. 85.7%) for 1 e/h, and higher LR^+ in CHAT and UofC for 10 e/h (LR^+ = 15.5-25.2 vs. LR^+ = 3-3.5) [11], [49], [50]. It should be noted that our study incorporated 2,655 subjects, providing statistically more robust and generalizable results compared to previous studies that included relatively restricted data samples consisting of 21 to 50 subjects/cohort. In terms of *Acc*, the current DL approach achieved higher values at all thresholds in CHAT compared to

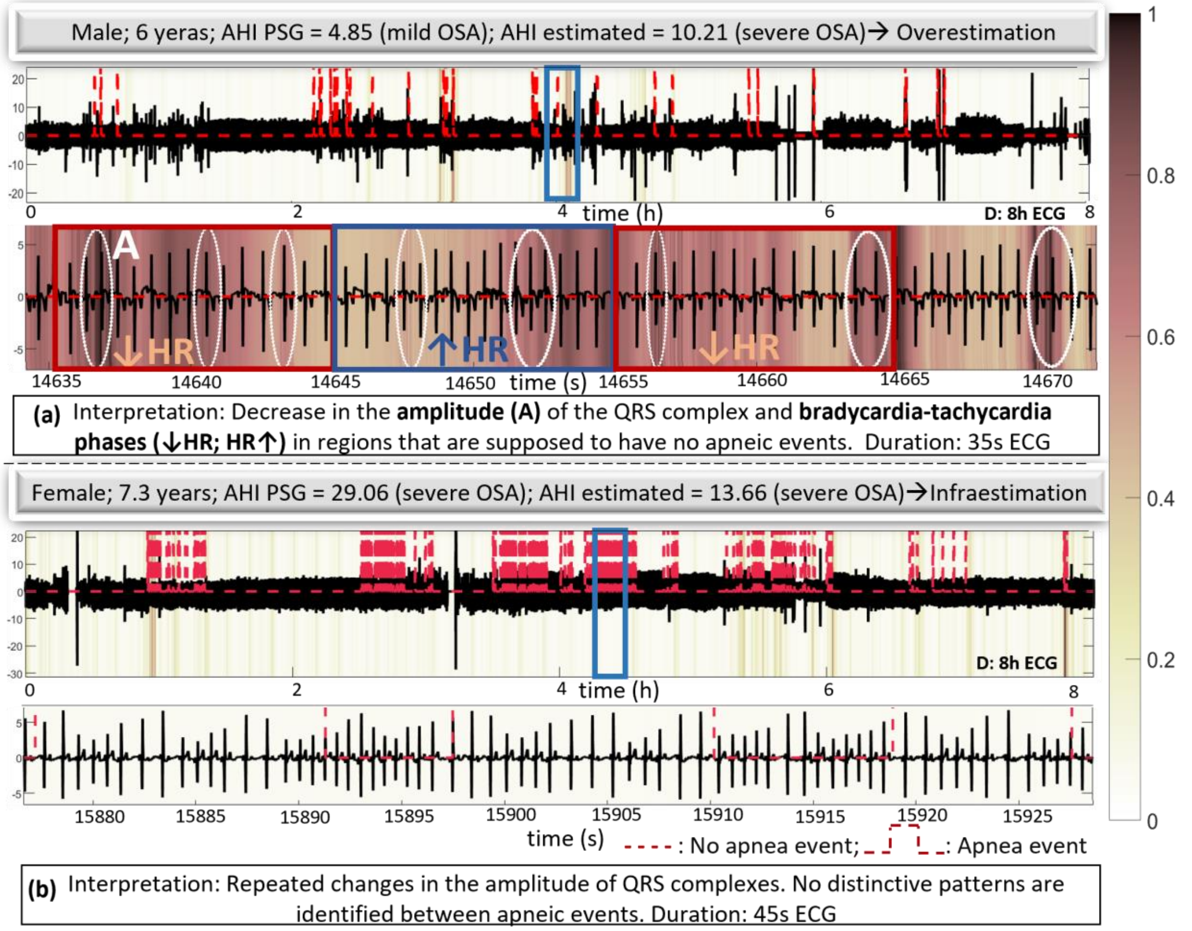

Fig. 7: Grad-CAM visualizations for some representative findings in ECG signals of the CHAT test set incorrectly predicted. Fig. 7 (a) depicts the heatmap of a nocturnal ECG signal and a zoom region associated with model overestimation. Fig. 7 (b) shows the heatmap of a nocturnal ECG signal and a zoom region related to underestimating the model. The colorbar indicates at 0 (yellow) the zones of lower relevance and at 1 (brown) the zones of higher relevance.

those reported by Martín-Montero et al. [51], [52]. This study obtained *Acc*<80% in 1 e/h, *Acc*<87% in 5 e/h, and *Acc*<92% in 10 e/h. Of note, Martín-Montero et al. [10] used the same CHAT database as in the present study. Overall, our algorithm presented higher values of k_4 (0.410 vs. 0.166) and Acc_4 (61.54% vs. 41.89%), highlighting the superiority of DL over traditional FE methods.

In a previous study [13], we applied a CNN-based model to estimate apneic events and pediatric OSA severity using single-lead ECG signals and the same CHAT database. The approach we propose here is an improved version of that earlier model. We obtained higher overall performance by evaluating the current proposal using this database, reaching higher k_4 (0.410 vs. 0.373) and Acc_4 (61.54% vs. 57.86%), and higher *Acc* at 1, 5, and 10 e/h. Furthermore, we obtained higher values in all metrics for 1 e/h apart from *Se*. This finding demonstrates the validity of our proposal to discern between pediatric OSA and unaffected children. Additionally, some LR^+ values obtained from 5 e/h onwards are remarkable, and SleepECG-Net presents high reliability in detecting moderate OSA ($LR^+ = 11.6$ vs. $LR^+ = 8.9$). Comparing k_2 results, we obtained the same value for 5 e/h ($k_2 = 0.69$) and higher values at 1 e/h ($k_2 = 0.38$ vs. $k_2 = 0.30$) and 10 e/h ($k_2 = 0.63$ vs. $k_2 = 0.60$). This difference in both thresholds is noteworthy because it indicates that this proposal can better discern at the more restrictive threshold and differentiate between the presence or absence of disease. Furthermore, a higher k_2 value at the 10 e/h threshold suggests that SleepECG-Net performs better in detecting the most severe subjects. This is very encouraging given that children with moderate to severe OSA are more likely to experience cardiovascular and neurocognitive morbidities [3]. This improvement in diagnostic performance is indicative that SleepECG-Net is better suited to the analysis of long ECG sequences with possible apneic clusters and signal patterns linked to the presence of the events. Accordingly, the proposal performs better at identifying OSA, providing more accurate predictions when using ECG signals. Finally, it is essential to highlight that the previous study could not justify the decisions made by the model, which limits its confidence in clinical practice settings. The application of XAI herein interprets the results obtained while opening the way to acquiring new knowledge in the field of OSA and discovering new clinically relevant ECG patterns.

Table IV also show previous studies that have used various traditional machine learning (ML) approaches, combining the use of SpO_2 signals with features derived from cardiac information (HR and pulse rate variability) and achieving encouraging performance [53]–[55]. Ye et *al.* [53] achieved an *Acc* = 90.4% at 1 e/h using the XGBoost method, but their reduced test set limited generalizability. Garde *et al.* [54]

reported *Acc* of 75%, 82%, and 89% at 1, 5, and 10 e/h, respectively, using a binary logistic regression model. Our study showed higher *Acc* across all thresholds. Dehkordi *et al.* [55] achieved an *Acc* = 71% and *Se* = 76% sensitivity at 5 e/h using PRV signal analysis, our model surpassing these metrics. Moreover, the limited accessibility of the data in these studies hinders the generalization of their results. Additionally, our study improves technical complexity and simplification by using only one signal instead of two.

Despite the most interesting discussion relying on comparing our results with those studies using cardiac information, other previous studies have implemented various traditional ML approaches [7], [56]. Moreover, some studies used the same CHAT and UofC databases along with DL alternatives for diagnosing pediatric OSA, demonstrating higher performance in pediatric OSA-related diagnosis than previous ML approaches [20], [57], [58]. All these studies mainly focused on the analysis of overnight SpO_2 and/or AF signals and reached promising results. However, by analyzing ECG signals, we can consider the possible OSA pathophysiological effects on the cardiac system, including those associated with cardiac morbidities. In addition, our approach enhanced diagnostic simplification compared to the methods proposed by Bertoni et al. [7], and Jiménez-García et al. [20], [57], which relied on two signals.

Finally, regarding the state of the art using XAI techniques for healthcare, most studies interpreting FE models with categorical input data used Shapley Additive Explanations (SHAP), while those implementing CNN-based methods with time series or images typically used Grad-CAM [59]. Our study involves overnight ECG recordings acting as time series data, where temporal relationships are crucial. The high computational load of SHAP, especially with DL models, and its additive nature may oversimplify the complex relationships in ECG data, potentially providing incomplete explanations [18]. Preliminary tests using SHAP revealed a computational cost 2.4 times higher than that of Grad-CAM. This point, combined with our objective of highlighting specific patterns in

Table IV

State-of-the-art studies on using cardiac signals to diagnose pediatric OSA.

Author	***Signal***	***#Total children/#Test***	***ML approach/Model/ Validation /XAI method***	***AHI (e/h)***	***Se (%)***	***Acc (%)***	***LR^+***
Shouldice et al. [9]	RR[a]	50/25	FE/QD /Loo cv /-	1	85.7	84.0	4.7
Gil et al. [49]	PPG+HRV[a]	21/21	FE/QD / - / -	>18 OSA <5 No OSA	87.5	80.0	3.1
Gil et al. [50]	PPG+PTTV[a]	21/21	FE/QD / - / -	>18 OSA <5 No OSA	75.0	80.0	3.1
Lázaro et al. [11]	PPG[a]	21/21	FE/LDA/ Loo cv /-	>18 OSA <5 No OSA	100.0	86.7	3.5
Martín-Montero et al.[52]	HRV[a]	1738/757 (CHAT, UofC)	FE/MLP / - / -	1 5 10	76.3 62.5 66.7	63.4 81.0 89.3	1.2 4.0 7.9
Martín-Montero et al. [51]	HRV[a]	1738/757 (CHAT, UofC)	FE/LDA / - / -	1 5 10	85.5 64.4 53.7	74.6 85.0 91.6	1.3 10.4 23.1
Ye et al. (2023). [53]	SpO_2 (ODI)+ HR[a]	3,139/628	FE/XGBoost/Holdout/SHAP	1 5 10	90.3 82.1 84.8	90.4 85.7 89.8	N.D N.D N.D
Garde et al. [54]	SpO_2 (ODI)+ PRV (Spectral)	207	FE/LR (for each threshold) / Holdout/-	1 5 10	80.0 85.0 82.0	75.0 82.0 89.0	N.D N.D N.D
Dehkordi et al. [55]	PRV	146/146	FE/LASSO/ -/-	5	76.0	71.0	N.D
Martín-Montero et al. [10]	HRV[a]	1610/296 (CHAT)	FE/LSboost/ Holdout / LIME	1 5 10	90.8 66.7 40.0	80.1 63.2 84.1	1.2 1.7 5.0
García-Vicente et al. [13]	ECG	1610/299 (CHAT)	DL/CNN / Holdout / -	1 5 10	84.2 76.7 53.7	75.9 87.0 92.0	1.6 8.9 27.7
Our proposal	**ECG**	**1610/299 (CHAT)**	**DL/CNN-BiLSTM / Holdout / Grad-CAM**	**1 5 10**	**89.7 72.2 58.8**	**80.3 87.3 92.3**	**1.7 11.6 25.2**
		981/981 (UofC)			**88.5 66.3 63.8**	**78.7 79.7 88.4**	**1.3 6.1 15.5**
		64/64 (CFS)			**83.3 66.7 66.7**	**57.8 98.4 98.4**	**1.45 N.D N.D**

RR = the period between two R peaks; PPG = photoplethysmography; HRV = heart rate variability; HR = heart rate; PTTV = pulse transit time variability; PRV = pulse rate variability, LASSO = least absolute shrinkage operating characteristic curves. [a] Features derived from these signals. N.D: Not defined.

the ECG signal that influence AHI estimation, led to the preference for using Grad-CAM. Grad-CAM effectively visualizes the areas of the ECG signal that contribute the most to the model's decisions, making it more suitable for our study than SHAP, which would be better for a detailed quantitative interpretation of sample importance. Therefore, although we recognize the existence of multiple techniques for explainability in DL, we chose Grad-CAM because of its balance between transparency, performance, computational burden, ease of use, and clinical applicability. In this sense, Grad-CAM has been notably applied to explain models using ECG signals to diagnose cardiac diseases [59]. However, limited studies have used XAI methods for diagnosing OSA, mainly focusing on FE models rather than direct DL on raw signals. In pediatric OSA diagnosis, only three studies used XAI, with methods including SHAP for demographic and heart rate variables and Grad-CAM for localization maps over AF and SpO_2 signals [10], [20], [53].

E. *Limitations and future work*

Several limitations should be mentioned. We used CHAT to develop and internally validate SleepECG-Net and UofC and CFS to externally validate SleepECG-Net. However, other strategies could be used to increase the generalizability of our approach. Moreover, the different databases were annotated by various specialists. This may be a limitation for SleepECG-Net to learn properly, but at the same time, it increases the objectivity of the method by not being biased to a single specialist. Thus, validating the algorithm on a more extensive set of databases and ambulatory overnight/daytime Holter ECG recordings would be beneficial to assess its performance in different circumstances and populations. Furthermore, despite the adequacy of the DL algorithm used, novel architectures, like other hybrid models and transformers, could be explored. Another future work could consist of implementing a multiclass regression model where the AHI and other variables related to cardiac risk factors are jointly estimated. Relative to this approach, exploring the use of ECG patterns to identify specific phenotypes OSA would allow for better classification and understanding of the disease, facilitating the development of more accurate and effective diagnostic and treatment strategies. Additionally, it would be interesting to develop DL algorithms for directly detecting abnormalities in ECG data to identify possible apneic events. In terms of model interpretability, although Grad-CAM is suitable for our proposal, other global XAI methods could also be evaluated in the future.

VI. CONCLUSIONS

To the best of our knowledge, this study is the first to evaluate an interpretable model based on a combination of CNN and RNN networks using overnight one-lead ECG signals to directly estimate the AHI and the OSA severity in pediatric patients. Our approach demonstrated higher diagnostic performance for pediatric OSA than in prior studies, especially in severe cases, which are closely linked to heightened heart cardiovascular risk. In relation, XAI results indicated recognition of both known ECG patterns associated with OSA and potential patterns related to cardiovascular morbidity. These findings pave the way for automated ECG analysis to identify cardiovascular end-organ dysfunction. In conclusion, implementing an interpretable DL approach using nocturnal ECG signals could offer an alternative tool to PSG with a high potential to facilitate timely, objective, and accurate diagnosis of the disease. Furthermore, integrating XAI techniques to demonstrate the decisions generated by the models contributes to strengthening confidence in such systems, promoting their effective adoption in clinical practice.

REFERENCES

[1] C. L. Marcus *et al.*, “Diagnosis and Management of Childhood Obstructive Sleep Apnea Syndrome,” *Pediatrics*, vol. 130, no. 3, pp. e714–e755, Sep. 2012.

[2] H. L. Tan, *et al.*, “Obstructive sleep apnea in children: A critical update,” *Nat. Sci. Sleep*, vol. 5, pp. 109–123, Sep. 2013.

[3] C. Guilleminault, *et al.*, “Cyclical variation of the heart rate in sleep apnoea syndrome. Mechanisms, and Usefulness of 24 h Electrocardiography as a Screening Technique,” *Lancet*, vol. 323, no. 8369, pp. 126–131, Jan. 1984.

[4] R. Bhattacharjee, *et al.*, “Cardiovascular Complications of Obstructive Sleep Apnea Syndrome: Evidence from Children,” *Prog. Cardiovasc. Dis.*, vol. 51, no. 5, pp. 416–433, Mar. 2009.

[5] R. B. Berry *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.*, vol. 8, no. 5, pp. 597–619, 2012.

[6] B. V. Berry, *et al.*, “The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications.,” *Am. Acad. Sleep Med.*, 2020.

[7] D. Bertoni and A. Isaiah, “Towards Patient-centered Diagnosis of Pediatric Obstructive Sleep Apnea—A Review of Biomedical Engineering Strategies,” *Expert Rev. Med. Devices*, vol. 16, no. 7, pp. 617–629, Jul. 2019.

[8] G. C. Gutiérrez-Tobal *et al.*, “Reliability of machine learning to diagnose pediatric obstructive sleep apnea: Systematic review and meta-analysis,” *Pediatr. Pulmonol.*, vol. 57, no. 8, pp. 1931–1943, Aug. 2022.

[9] R. B. Shouldice, *et al.*, “Detection of Obstructive Sleep Apnea in Pediatric Subjects using Surface Lead Electrocardiogram Features,” *Sleep*, vol. 27, no. 4, pp. 784–792, Jun. 2004.

[10] A. Martín-Montero *et al.*, “Pediatric sleep apnea: Characterization of apneic events and sleep stages using heart rate variability,” *Comput. Biol. Med.*, vol. 154, p. 106549, Mar. 2023.

[11] J. Lazaro, *et al.*, “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. Informatics*, vol. 18, no. 1, pp. 240–246, Jan. 2014.

[12] I. Goodfellow, *et al.*, *Deep learning*. MIT Press, 2016.

[13] C. García-Vicente, *et al.*, “ECG-based convolutional neural network in pediatric obstructive sleep apnea diagnosis,” *Comput. Biol. Med.*, vol. 167, no. September, p. 107628, Dec. 2023.

[14] P. Moridian *et al.*, “Automatic diagnosis of sleep apnea from biomedical signals using artificial intelligence techniques: Methods, challenges, and future works,” *WIREs Data Min. Knowl. Discov.*, vol. 12, no. 6, Nov. 2022.

[15] Mostafa, *et al.*, “A Systematic Review of Detecting Sleep Apnea Using Deep Learning,” *Sensors*, vol. 19, no. 22, p. 4934, Nov. 2019.

[16] H. Korkalainen *et al.*, “Accurate Deep Learning-Based Sleep Staging in a Clinical Population with Suspected Obstructive Sleep Apnea,” *IEEE J. Biomed. Heal. Informatics*, vol. 24, no. 7, pp. 1–1, 2019.

[17] A. Zarei, *et al.*, “Detection of sleep apnea using deep neural networks and single-lead ECG signals,” *Biomed. Signal Process. Control*, vol. 71, p. 103125, Jan. 2022.

[18] A. Barredo Arrieta *et al.*, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[19] H. W. Loh, *et al.*, “Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022),” *Comput. Methods Programs Biomed.*, vol. 226, p. 107161, 2022.

[20] J. Jiménez-García *et al.*, “An explainable deep-learning architecture for pediatric sleep apnea identification from overnight airflow and oximetry signals,” *Biomed. Signal Process. Control*, vol. 87, no.

September 2023, 2024.

[21] F. Vaquerizo-Villar *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.*, vol. 2, no. 2, p. 107419, Aug. 2023.

[22] S. Redline *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*, vol. 34, no. 11, pp. 1509–1517, Nov. 2011.

[23] C. L. Marcus *et al.*, “A Randomized Trial of Adenotonsillectomy for Childhood Sleep Apnea,” *N. Engl. J. Med.*, vol. 368, no. 25, pp. 2366–2376, Jun. 2013.

[24] R. Hornero *et al.*, “Nocturnal Oximetry–based Evaluation of Habitually Snoring Children,” *Am. J. Respir. Crit. Care Med.*, vol. 196, no. 12, pp. 1591–1598, Dec. 2017.

[25] S. Redline *et al.*, “The familial aggregation of obstructive sleep apnea,” *Am. J. Respir. Crit. Care Med.*, vol. 151, no. 3 I, pp. 682–687, 1995.

[26] C. Iber, *et al.*, “The-AASM-Manual-for-Scoring-of-Sleep-and-Associated-Events-2007-,” *AASM Manual for Scoring Sleep*. pp. 3–49, 2007.

[27] H. L. Tan, *et al.*, “Overnight Polysomnography versus Respiratory Polygraphy in the Diagnosis of Pediatric Obstructive Sleep Apnea,” *Sleep*, vol. 37, no. 2, pp. 255–260, Feb. 2014.

[28] J. Zhang *et al.*, “Automatic Detection of Obstructive Sleep Apnea Events Using a Deep CNN-LSTM Model,” *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–10, Mar. 2021.

[29] A. Sheta *et al.*, “Diagnosis of Obstructive Sleep Apnea from ECG Signals Using Machine Learning and Deep Learning Classifiers,” *Appl. Sci.*, vol. 11, no. 14, p. 6622, Jul. 2021.

[30] L. Sörnmo and P. Laguna, “The Electrocardiogram—A Brief Background,” in *Bioelectrical Signal Processing in Cardiac and Neurological Applications*, L. Sörnmo and P. B. T.-B. S. P. in C. and N. A. Laguna, Eds. Burlington: Elsevier, 2005, pp. 411–452.

[31] J. W. Chen, *et al.*, “Predicting Apnea-Hypopnea Index in Patients with Obstructive Sleep Apnea Using Unsegmented ECG-Signal-Based Algorithms,” *IEEJ Trans. Electr. Electron. Eng.*, vol. 18, no. 9, pp. 1550–1552, Sep. 2023.

[32] J.-W. Chen *et al.*, “A Signal Segmentation-Free Model for Electrocardiogram-Based Obstructive Sleep Apnea Severity Classification,” *Adv. Intell. Syst.*, p. 2200275, Jan. 2023.

[33] R. T. Brouillette, *et al.*, “Nocturnal pulse oximetry as an abbreviated testing modality for pediatric obstructive sleep apnea,” *Pediatrics*, vol. 105, no. 2, pp. 405–412, 2000.

[34] D. A. van Dyk and X.-L. Meng, “The Art of Data Augmentation,” *J. Comput. Graph. Stat.*, vol. 10, no. 1, pp. 1–50, Mar. 2001.

[35] J. Cohen, “A coefficient of agreement for nominal scales,” *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[36] R. R. Selvaraju, *et al.*, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[37] J. J. Bartko, “The Intraclass Correlation Coefficient as a Measure of Reliability,” *Psychol. Rep.*, vol. 19, no. 1, pp. 3–11, 1966.

[38] M. L. McHugh, “Lessons in biostatistics interrater reliability : the kappa statistic,” *Biochem. Medica*, vol. 22, no. 3, pp. 276–282, 2012.

[39] D. Álvarez *et al.*, “Assessment of feature selection and classification approaches to enhance information from overnight oximetry in the context of apnea diagnosis,” *Int. J. Neural Syst.*, vol. 23, no. 05, p. 1350020, Oct. 2013.

[40] C. Kraikriangsri, *et al.*, “P-wave dispersion as a simple tool for screening childhood obstructive sleep apnea syndrome,” *Sleep Med.*, vol. 54, pp. 159–163, Feb. 2019.

[41] S. Ebrahimian *et al.*, “Beat-to-beat cardiac repolarization lability increases during hypoxemia and arousals in obstructive sleep apnea patients,” *Am. J. Physiol. Circ. Physiol.*, Mar. 2024.

[42] A. Shamsuzzaman *et al.*, “Daytime cardiac repolarization in patients with obstructive sleep apnea,” *Sleep Breath.*, vol. 19, no. 4, pp. 1135–1140, Dec. 2015.

[43] M. M. Khan, *et al.*, “Management of recurrent ventricular tachyarrhythmias associated with Q-T prolongation,” *Am. J. Cardiol.*, vol. 47, no. 6, pp. 1301–1308, Jun. 1981.

[44] S. Solhjoo, *et al.*, “Sleep-Disordered Breathing Destabilizes Ventricular Repolarization,” *medRxiv*, Mar. 2023.

[45] C. L. MARCUS, “Sleep-disordered Breathing in Children,” *Am. J. Respir. Crit. Care Med.*, vol. 164, no. 1, pp. 16–30, Jul. 2001.

[46] L. Duque-González, *et al.*, “The U wave: an ignored wave filled with information,” *Cardiovasc. Metab. Sci.*, vol. 32, no. 4, pp. 197–205, 2021.

[47] K. A. Domany *et al.*, “Effect of Adenotonsillectomy on Cardiac Function in Children Age 5-13 Years With Obstructive Sleep Apnea,” *Am. J. Cardiol.*, vol. 141, pp. 120–126, Feb. 2021.

[48] R. S. Amin *et al.*, “Left Ventricular Hypertrophy and Abnormal Ventricular Geometry in Children and Adolescents with Obstructive Sleep Apnea,” *Am. J. Respir. Crit. Care Med.*, vol. 165, no. 10, pp. 1395–1399, May 2002.

[49] E. Gil, *et al.*, “Discrimination of Sleep-Apnea-Related Decreases in the Amplitude Fluctuations of PPG Signal in Children by HRV Analysis,” *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1005–1014, Apr. 2009.

[50] E. Gil, *et al.*, “PTT Variability for Discrimination of Sleep Apnea Related Decreases in the Amplitude Fluctuations of PPG Signal in Children,” *IEEE Trans. Biomed. Eng.*, vol. 57, no. 5, pp. 1079–1088, May 2010.

[51] A. Martín-Montero *et al.*, “Heart rate variability spectrum characteristics in children with sleep apnea,” *Pediatr. Res.*, vol. 89, no. 7, pp. 1771–1779, May 2021.

[52] A. Martín-Montero *et al.*, “Bispectral Analysis of Heart Rate Variability to Characterize and Help Diagnose Pediatric Sleep Apnea,” *Entropy*, vol. 23, no. 8, p. 1016, Aug. 2021.

[53] P. Ye *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. - Head Neck Med. Surg.*, vol. 44, no. 2, p. 103714, Mar. 2023.

[54] A. Garde *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.*, vol. 60, pp. 45–52, Aug. 2019.

[55] P. Dehkordi *et al.*, “Evaluation of cardiac modulation in children in response to apnea / hypopnea using the Phone Evaluation of cardiac modulation in children in response to apnea / hypopnea using the Phone Oximeter TM”.

[56] Z. Xu *et al.*, “Cloud algorithm-driven oximetry-based diagnosis of obstructive sleep apnoea in symptomatic habitually snoring children,” *Eur. Respir. J.*, vol. 53, no. 2, p. 1801788, Feb. 2019.

[57] J. Jiménez-García *et al.*, “A 2D convolutional neural network to detect sleep apnea in children using airflow and oximetry,” *Comput. Biol. Med.*, vol. 147, no. April, 2022.

[58] F. Vaquerizo-Villar *et al.*, “A Convolutional Neural Network Architecture to Enhance Oximetry Ability to Diagnose Pediatric Obstructive Sleep Apnea,” *IEEE J. Biomed. Heal. Informatics*, vol. 25, no. 8, pp. 2906–2916, 2021.

[59] Y. M. Ayano, *et al.*, “Interpretable Machine Learning Techniques in ECG-Based Heart Disease Classification: A Systematic Review,” *Diagnostics*, vol. 13, no. 1, pp. 1–37, 2023.