Computers in Biology and Medicine 165 (2023) 107419

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

An explainable deep-learning model to stage sleep states in children and propose novel EEG-related patterns in sleep apnea

Fernando Vaquerizo-Villar [a,b,*], Gonzalo C. Gutiérrez-Tobal [a,b], Eva Calvo [a], Daniel Álvarez [a,b,c], Leila Kheirandish-Gozal [d], Félix del Campo [a,b,c], David Gozal [e], 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] Sleep-Ventilation Unit, Pneumology Department, Río Hortega University Hospital, Valladolid, Spain
[d] Departments of Neurology and Child Health and Child Health Research Institute, The University of Missouri School of Medicine, Columbia, MO, USA
[e] Office of The Dean, Joan C. Edwards School of Medicine, Marshall University, 1600 Medical Center Dr, Huntington, WV, 25701, USA

ARTICLE INFO

Keywords:
Deep learning
Electroencephalogram (EEG)
Explainable artificial intelligence (XAI)
Gradient-weighted class activation mapping (grad-CAM)
Pediatric obstructive sleep apnea (OSA)
Sleep staging

ABSTRACT

Automatic deep-learning models used for sleep scoring in children with obstructive sleep apnea (OSA) are perceived as black boxes, limiting their implementation in clinical settings. Accordingly, we aimed to develop an accurate and interpretable deep-learning model for sleep staging in children using single-channel electroencephalogram (EEG) recordings. We used EEG signals from the Childhood Adenotonsillectomy Trial (CHAT) dataset (n = 1637) and a clinical sleep database (n = 980). Three distinct deep-learning architectures were explored to automatically classify sleep stages from a single-channel EEG data. Gradient-weighted Class Activation Mapping (Grad-CAM), an explainable artificial intelligence (XAI) algorithm, was then applied to provide an interpretation of the singular EEG patterns contributing to each predicted sleep stage. Among the tested architectures, a standard convolutional neural network (CNN) demonstrated the highest performance for automated sleep stage detection in the CHAT test set (accuracy = 86.9% and five-class kappa = 0.827). Furthermore, the CNN-based estimation of total sleep time exhibited strong agreement in the clinical dataset (intra-class correlation coefficient = 0.772). Our XAI approach using Grad-CAM effectively highlighted the EEG features associated with each sleep stage, emphasizing their influence on the CNN's decision-making process in both datasets. Grad-CAM heatmaps also allowed to identify and analyze epochs within a recording with a highly likelihood to be misclassified, revealing mixed features from different sleep stages within these epochs. Finally, Grad-CAM heatmaps unveiled novel features contributing to sleep scoring using a single EEG channel. Consequently, integrating an explainable CNN-based deep-learning model in the clinical environment could enable automatic sleep staging in pediatric sleep apnea tests.

1. Introduction

Characterization of the sleep-macrostructure (i.e., sleep stages) is essential in the evaluation and diagnosis of numerous sleep disorders [1]. Overnight polysomnography (PSG) consists of the gold standard approach and is commonly coupled with analytical guidelines as stipulated by the American Academy of Sleep Medicine (AASM) [2]. PSG involves the recording of a large number neurophysiological and cardiorespiratory signals, including electroencephalogram (EEG), electrooculogram (EOG), and electromyogram (EMG) channels. Subsequently, after the completion of overnight recordings, technicians have to visually examine EEG, EOG, and submental EMG signals using strict criteria, i.e.., AASM rules, to classify each 30-s non-overlapping epoch of nearly 480–600 min of sleep recordings into one out of the five different stages: wake (W), three levels of non-Rapid Eye Movement (non-REM) sleep (N1, N2, and N3), and REM sleep [2]. The process of manual sleep staging is laborious and tedious and requires up to 2 h to complete [3]. Furthermore, a considerable inter-rater variability has been reported in manual sleep scoring [3]. Thus, automated sleep scoring from a minimum number of channels would be preferable to

* Corresponding author. Biomedical Engineering Group, Facultad de Medicina, Av. Ramón y Cajal, 7, 47003, Valladolid, Spain.
E-mail address: fernando.vaquerizo@gib.tel.uva.es (F. Vaquerizo-Villar).
URL: http://www.gib.tel.uva.es (F. Vaquerizo-Villar).

https://doi.org/10.1016/j.compbiomed.2023.107419
Received 6 June 2023; Received in revised form 26 July 2023; Accepted 28 August 2023
Available online 31 August 2023

enhance consistency, improve children's comfort, simplify the procedure, and reduce associated costs.

Multiple studies have proposed automated approaches for sleep staging [4–16]. Since a large proportion of the sleep studies currently conducted in sleep laboratories around the world aim at establishing the diagnosis of obstructive sleep apnea (OSA), a condition that affects nearly 1 billion people across all age groups [17], many studies aimed at automatic sleep scoring have focused on OSA patient cohorts [4–6,8–14,16]. OSA diagnosis is established based on the apnea-hypopnea index (AHI), which is computed as the number of apneas and hypopneas per hour of sleep. Therefore, the identification of sleep stages and the determination of the total sleep time (TST) are essential in this context. Unfortunately, inter-rater agreement on sleep stages is lower in OSA patients than in healthy subjects [18], which further highlights the need for the objectivity provided by automated sleep scoring models for OSA patients. Accordingly, most automated approaches have primarily targeted adults for development and validation [4,5,9–14], with only few studies focusing on children being evaluated for suspected OSA [6,8,16]. This discrepancy is not surprising, given that pediatric OSA presents distinguishing etiological, diagnostic, and treatment considerations when compared to adult subjects. Children present a reduced upper airway collapsibility [19], which results in less frequent respiratory events, and accordingly imposes more restrictive scoring rules for apneas and hypopneas, as well as in lower cut-off values of the AHI for diagnosis and severity grading than in adults [19,20]. Sleep architecture and electroencephalographic activity also present substantial developmental differences [2], even during different stages of childhood. Consequently, specific scoring rules for sleep stages are applied in the pediatric population. Due to these differences, there is a much higher level of uncertainty and variability across centers and sleep scorers when pediatric OSA is suspected since this diagnosis is exceedingly more challenging than in the adult population. This emphasizes the necessity for developing specific automatic sleep scoring models tailored to pediatric OSA patients.

In the last few years, deep-learning approaches have emerged as an overarching novel methodological approach with ability to improve automatic sleep scoring [3]. Particularly, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have demonstrated their effectiveness to automatically detect sleep stages from EEG recordings [3]. However, sleep professionals perceive these deep-learning methods as 'black boxes' [21], which limits their acceptance and application in clinical settings. A recent report from the European Union (EU) has emphasized the need to enhance the robustness, transparency, and explainability of artificial intelligence (AI)-based systems to ensure their responsible and informed deployment in society [22]. In this respect, explainable AI (XAI) techniques have recently gained increasing attention due to their capability to explain AI-based models (including deep-learning ones) *a posteriori* [21]. This is particularly relevant in sleep staging, given the substantial discrepancy observed among human experts [3]. By applying XAI analysis, it is possible not only to explain EEG patterns associated with each sleep stage but also to identify novel sleep stage-related patterns that could potentially improve the consistency of sleep scoring. One of these XAI techniques is Gradient-weighted Class Activation Mapping (Grad-CAM), which utilizes gradient information flowing into convolutional layers to identify the regions in the input data that have the highest importance in the predictions of a CNN-based network [23].

In the sleep context, there are some very recent studies proposing XAI approaches to explain the decisions made by deep-learning models [7,11,24–27]. On the one hand, Barnes et al. [27], Troncoso-García et al. [25], and Rossi et al. [24], used XAI to identify physiological features related to apnea/hypopnea events in adult OSA subjects. Conversely, Kuo et al. [26], Phan et al. [11] and Dutt et al. [7] applied XAI techniques to provide an interpretation of the time-frequency EEG patterns considered by their corresponding deep-learning models for predicting sleep stages in adult subjects. In contrast to these studies, which have focused on the application of XAI techniques in adult subjects, we propose a XAI-based methodology for sleep staging in children with clinical suspicion of OSA. As aforementioned, sleep scoring is more challenging in children than in the adult population.

Based on the aforementioned factors, the novelties of this study rely on the application of a XAI methodology based on Grad-CAM to obtain an interpretable deep-learning model aimed at accurately classify sleep stages in pediatric OSA patients while examining a single EEG channel, namely the C4-M1 derivation. Fig. 1 shows the general outline of the proposed methodology. We applied three deep-learning architectures and compared them in their performance to automatically score sleep stages in pediatric OSA patients. Subsequently, Grad-CAM was implemented such as to explain the EEG features associated with each predicted sleep stage within the pediatric cohort. We hypothesized that the combination of deep-learning and XAI algorithms can yield high-performance and interpretable models that are clinically applicable for the automated detection of sleep stages in children. Major contributions of the current study consist of: (*i*) the combination of deep-learning and XAI analysis for automated sleep stage detection in children with suspected OSA; (*ii*) the use of Grad-CAM to explain the EEG features related to each sleep stage predicted by the deep-learning models; (*iii*) the application of Grad-CAM to identify and analyze epochs that have a high probability of being misclassified; (*iv*) the use of Grad-CAM to propose novel hallmarks for sleep stage detection in single-channel EEG overnight recordings.

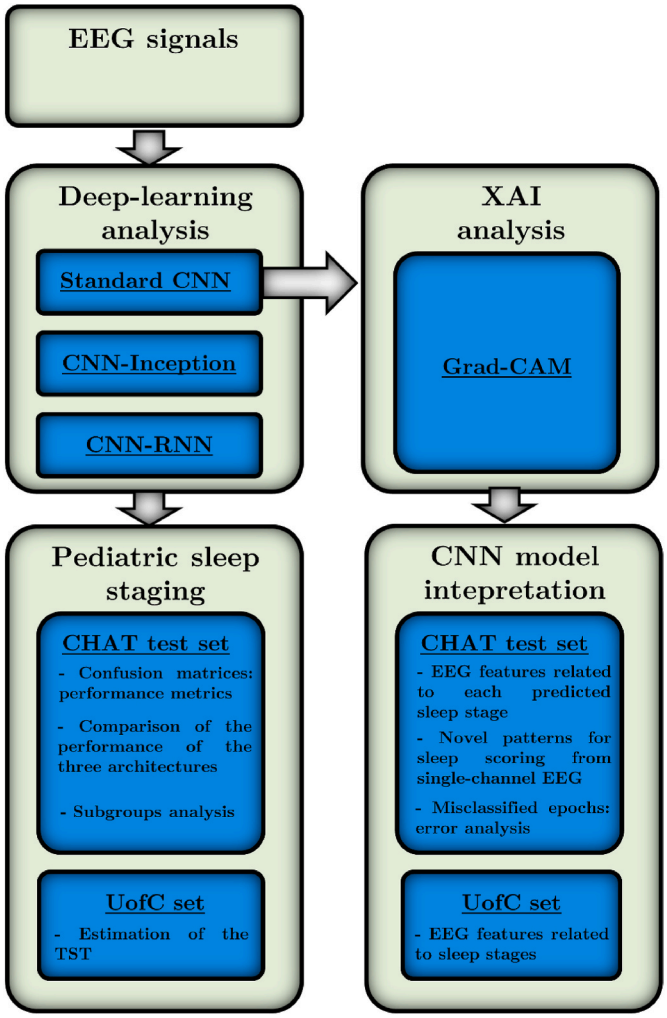

Fig. 1. Flowchart of the proposed methodology.

2. Materials and methods

The following subsections describe the databases and methods used in this study.

2.1. Subjects and signals under study

Two datasets were used in this work. The first one was the Childhood Adenotonsillectomy Trial (CHAT), which is a semi-public dataset consisting of 1639 sleep studies conducted on pediatric subjects aged 5–10 years old who were evaluated due to clinical suspicion of OSA [28]. Within this dataset, there were 1637 valid EEG recordings available. The subjects in the CHAT database were originally divided into three subgroups [28,29]: baseline (453 subjects), follow-up (406 subjects), and non-randomized (778 subjects). Each sleep study in the dataset included annotations of sleep stages and apnea/hypopnea events, which were initially determined by the participating centers according to the AASM rules [30]. These annotations were subsequently re-evaluated, scored and confirmed by a centralized scoring facility [28]. The data from the three subgroups were randomly divided into three sets: training (50%), used for training the deep-learning models, validation set (25%), used for adjusting regularization and monitor the convergence of the models, and test set (25%), used for performance assessment and interpretation of the deep-learning models. To avoid any bias resulting from including EEG recordings from the same pediatric subject in multiple sets, the same division of subjects into training/validation/test sets as performed in the baseline subgroup was also applied to the follow-up subgroup. In this study, we used the C4-M1 EEG derivation, which is one of the EEG channels recommended by the AASM for sleep staging [2]. The original data were acquired at sampling rates (fs) ranging from 200 to 500 Hz. To standardize the frequency and reduce computational costs, the data were resampled to a common sampling rate of 125 Hz. Table 1 presents the clinical and demographic information of the population under study.

The second dataset utilized in this study consisted of a proprietary database comprising 980 valid EEG recordings obtained from pediatric symptomatic subjects referred for PSG at the University of Chicago (UofC). Unlike the CHAT dataset, the UofC dataset did not include an annotated hypnogram with sleep stages in the raw signal dataset (see Supplemental Table 2). Consequently, this dataset was only used to assess the ability of the deep-learning models to estimate the TST in an external dataset and to evaluate the agreement between the Grad-CAM explanations obtained from UofC and CHAT datasets. Similar to the CHAT database, the C4-M1 EEG channel from each subject was used in the UofC dataset. The original data in the UofC dataset were acquired at fs of either 200 Hz or 500 Hz. Prior to applying the proposed methodology, these data were also resampled to a standardized frequency of 125 Hz.

Table 1
Clinical and polysomnographic data of the children in the CHAT database.

	CHAT (training)	CHAT (validation)	CHAT (test)
Subjects (*n*)	818	409	410
Age (years)	7 [6; 8]	7 [6; 8]	7 [6; 8]
Males (*n*)	387 (47.3%)	203 (49.6%)	168 (45.4%)
BMI (kg/m^2)	17.3 [15.3; 21.7]	17.3 [15.7, 21.4]	17.3 [15.5, 21.1]
AHI (e/h)	2.6 [1.1; 6.5]	2.5 [1.2; 5.8]	2.3 [1.1; 5.2]
Wake (*n*)	235705 (24.1%)	114918 (23.7%)	114675 (23.5%)
N1 (*n*)	60559 (6.2%)	28723 (5.9%)	30087 (6.2%)
N2 (*n*)	314261 (32.1%)	156344 (32.3%)	158005 (32.4%)
N3 (*n*)	231314 (23.6%)	116363 (24.0%)	116887 (23.9%)
REM (*n*)	136541 (14.0%)	68468 (14.1%)	68785 (14.1%)
TRT (min)	587 [545, 648]	584 [548, 637]	586 [545, 646]
TST (min)	463 [423, 493]	461 [427, 489]	463 [427, 494]

Data presented as median [interquartile range] or n (%).
BMI = body mass index, AHI = apnea-hypopnea index; CHAT = Childhood Adenotonsillectomy Trial, e/h = events per hour, REM: rapid eye movement, TRT: total recording time, TST: total sleep time.

2.2. Deep-learning architectures

This study evaluated the capability of three distinct deep-learning architectures to determine the probability of each 30-s EEG epoch belonging to each sleep stage (W/N1/N2/N3/REM). The main components of each architecture are described next.

- *Standard CNN*. CNN is the most widely-used deep-learning algorithm for time series processing [31]. The CNN architecture utilized in this study was adapted from the network proposed by Sors et al. [10], which aimed to classify sleep stages in adult OSA patients. The proposed CNN architecture receives as input the EEG 30-s epoch to be classified, concatenated along with two preceding and one posterior epochs, resulting in a 120-s input segment. This input segment is processed through 12 convolutional blocks, each one composed of the following layers: a 1-D convolution, batch normalization (BN), activation, and dropout [10]. In the present study, the architecture proposed by Sors et al. [10] was improved by adding batch normalization and dropout layers to minimize overfitting. The dropout layer within each block used a probability of 0.1, which was empirically determined as the optimal probability that maximized the accuracy in the validation set.
- *CNN-Inception*. The configuration of CNN-Inception is based on the EEG-Inception network originally developed by Santamaría-Vázquez et al. [32] for the detection of event-related potentials [32]. It is a CNN architecture that incorporates inception modules, enabling parallel processing of the input data (i.e., the 120-s input EEG segment, as in the standard CNN) using convolutional layers with different filter size to learn features at different time scales/resolutions [33]. First, two inception modules are used to process the input. Each module consists of three branches with a convolutional block (1-D convolution, BN, and activation) to learn features at 3 distinct temporal scales: 500 ms, 250 ms, and 125 ms. The outputs from these branches are concatenated and subsequently average-pooled. Then, two convolutional blocks with average-pooling are used to continue extracting important features before reaching the final output. In this study, some minor modifications were made to the network: (*i*) depth-wise 2D convolutions were removed since the input is 1D (single-channel EEG); (*ii*) the number of convolutional filters and the average-pooling factor are multiplied by 2 due to the longer input size in seconds in our study (120-s *vs.* 1-s); (*iii*) the dropout layer was removed as a dropout rate of 0.0 (*i.e.*, no dropout) yielded the highest validation accuracy.
- *CNN-RNN*. The configuration of CNN-RNN is based on the deep neural network developed by Korkalainen et al. [9] for sleep stage detection using EEG signals [9]. CNN-RNN processes a sequence of 100 EEG epochs of 30-s by combining a CNN with an RNN. Each 30-s epoch is first processed individually through a time distributed layer containing a CNN. The CNN consists of six convolutional blocks (1D convolution, BN, and activation), two max-pooling layers and a global average (GAP) layer, which extract the EEG features from each epoch associated with sleep stages. The time distributed CNN is subsequently fed into an RNN composed of a bidirectional Gate Recurrent Unit (GRU), which learns the temporal distribution of sleep stages. The choice of GRU over LSTM was made due to similar performance with a lower computational load. In comparison to Korkalainen et al. [9], the gaussian dropout layer and dropout at the input of the GRU layer were removed (probabilities set as 0.0), while the recurrent dropout in GRU layer was set as 0.75, as these values resulted in the highest performance on the validation set.

The three architectures were trained on a NVIDIA GeForce RTX 2080 GPU, using the following configuration [34]: He-normal method for weights and biases initialization; Adam algorithm with an initial learning rate of 0.001 for weights and biases optimization; categorical cross entropy as the objective function to minimize; batch sizes of 128

(CNN and CNN-Inception) and 16 (CNN-RNN) using 50 reading queues from different patients to efficiently feed training data into the GPU memory in random order [10]; reduction of the learning rate by a factor of 2 after 5 epochs of non-improvement in the validation loss; early stopping after 20 epochs of non-improvement, restoring the model to the best weights determined by the validation set.

2.3. Explainable artificial intelligence: Grad-CAM

Class Activation Mapping (CAM) was initially proposed by Zhou et al. [23] as a XAI technique capable of identifying discriminative regions that significantly influence the predicted output of CNNs used for image classification. However, CAM is limited to the last convolutional layer of CNNs, where GAP feature maps are followed by a final softmax layer [23]. Grad-CAM is an enhanced version of CAM that uses the gradient information flowing into a specified convolutional layer to understand the importance of each input element in the decision-making process of the network. This makes Grad-CAM applicable to any CNN-based architecture [23]. To generate the class-discriminative localization map (heatmap), Grad-CAM calculates the gradients of the output of the target class y^c with respect to the 2-D feature maps A^k_{ij} of the chosen convolutional layer, i.e., $\frac{\partial y^c}{\partial A^k_{i,}}$. These gradients are then averaged to get the weights α^c_k, which capture the importance of each feature map k for the target class c:

$$\alpha^c_k = \frac{1}{Z}\sum_i \frac{\partial y^c}{\partial A^k_{i,}} \quad (1)$$

where Z is the number of feature maps (filters) in the layer. Subsequently, the Grad-CAM heatmap is derived by performing a weighted combination of the feature maps, which is then followed by a Rectified Linear Unit (ReLU) activation:

$$L^c_{Grad-CAM} = ReLU\left(\sum_k \alpha^c_k \bullet A^k\right) \quad (2)$$

As a result, a heatmap of the same dimensions as the feature maps of the corresponding convolutional layer is generated. This heatmap is then normalized and resized to facilitate a comprehensive joint visualization with the raw EEG signal [23]. In this study, we computed the average of the normalized and resized Grad-CAM heatmaps obtained for each layer, as we believe this approach enhances the identification of EEG patterns associated with different sleep stages.

2.4. Statistical analysis

The overall performance of the deep-learning architectures for automatic sleep staging was assessed by means of confusion matrices (5-class). These matrices were used to compute the 5-class accuracy (Acc), Cohen's kappa index (kappa), macro F1-score (MF1), and per-class F1-score (F1). Additionally, the TST was calculated for each patient based on automatic sleep scoring and compared with the TST obtained from PSG data in both the CHAT test set and the external UofC set. To assess the agreement of estimated TST, Bland-Altman plots and the intra-class correlation coefficient (ICC) were used.

3. Results

3.1. Deep-learning models performance

Fig. 2 shows the confusion matrices of the three deep-learning models (CNN, CNN-Inception, and CNN-RNN), while Table 2 shows the performance metrics of these models in the CHAT test set. Among the models, the CNN architecture demonstrated the highest overall performance, achieving an Acc of 86.9%, a kappa of 0.827, and a MF1 of 82.7%. This model outperformed CNN-RNN (Acc = 86.0%, kappa = 0.815, and MF1 = 81.4%) and CNN-Inception (Acc = 84.3%, kappa = 0.791, and MF1 = 76.6%) models. Given that class imbalance among sleep stages could potentially impact the outcomes of our approach, we have conducted additional analyses to address this issue. In section 1 of the supplementary material, we provide a comparison of the performance of the proposed CNN architecture with two CNNs trained using a batch-balance configuration: (*i*) CNN with batch-balance in the training set (CNN_{BBT}); (*ii*) CNN with batch-balance in the training and validation sets (CNN_{BBTV}). This comparison shows that implementing a balance strategy does not lead to an improvement in model performance.

Table 3 shows the diagnostic performance of the CNN model in the test set by OSA severity, gender, and CHAT subgroups. Regarding gender and OSA severity, a similar performance was observed across different subgroups. Conversely, performance metrics were slightly higher in follow-up than in nonrandomized and baseline subjects in the CHAT test set.

Supplementary Fig. 2 shows the Bland-Altman plot of the TST derived from automated CNN-based scoring against those obtained during PSG in the UofC set, together with the obtained in the CHAT test set (see section 2.2 of the Supplementary Material). The TST derived from the CNN model exhibited a high performance in the UofC set, albeit with a lower intra-class correlation coefficient (ICC) and a higher 95% confidence interval compared to the CHAT test set (0.856 *vs*. 0.772 and 122.19 min *vs*. 191.14 min).

3.2. Grad-CAM heatmaps interpretation of the CNN model

Figs. 3 and 4 show some representative examples of Grad-CAM heatmaps obtained for EEG epochs in the test set rightly predicted as W, N1, N2, and REM sleep stages. For each heatmap, a zoom in a relevant region of the heatmap is included at the right, together with the short-time Fourier transform (STFT), which better shows the time-frequency characteristics of the EEG patterns that the CNN is focusing on to make the prediction. The STFT is only shown for the 0–30 Hz

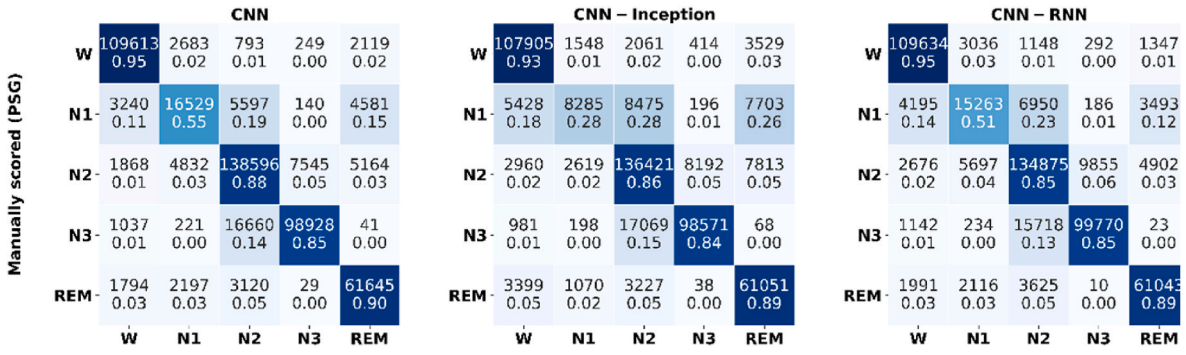

Fig. 2. Confusion matrices of the CNN, CNN-Inception and CNN-RNN models in the CHAT test set.

Table 2
Diagnostic performance of CNN, CNN-Inception, and CNN-RNN models to automatically classify sleep stages.

	Overall Metrics			Per-class F1-score (F1)				
	Acc (%)	kappa	MF1(%)	W	N1	N2	N3	REM
CNN	**86.9**	**0.827**	**82.7**	**94.1**	**58.5**	**85.9**	**88.4**	86.6
CNN-Inception	84.3	0.791	76.6	91.4	37.8	83.9	87.9	82.0
CNN-RNN	86.0	0.815	81.4	93.3	54.1	84.2	87.9	**87.5**

Acc = Accuracy, CNN = Convolutional neural network, MF1 = macro F1-score, N1: level 1 of non-rapid eye movement (NREM) sleep, N2: level 2 of NREM sleep, N3: level 3 of NREM sleep, REM: rapid eye movement.

Table 3
Diagnostic performance of the CNN model in the test set by sex, OSA severity, and CHAT subgroups.

		Overall Metrics			Per-class F1-score (F1)				
		Acc (%)	kappa	MF1(%)	W	N1	N2	N3	REM
Sex	**Males**	86.7	0.825	82.4	**94.2**	58.1	85.2	**88.5**	86.1
	Females	**87.2**	**0.830**	**82.9**	94.0	**58.8**	**86.4**	88.4	**87.0**
OSA severity	**No OSA**	86.8	0.826	82.1	94.2	56.4	**86.0**	88.2	85.6
	Mild OSA	**87.1**	0.829	**83.0**	94.1	**59.1**	**86.0**	88.1	**87.5**
	Moderate OSA	**87.1**	**0.832**	82.8	**95.3**	58.4	85.3	88.8	86.1
	Severe OSA	86.2	0.818	82.4	92.2	59.3	85.6	**89.9**	84.8
CHAT subgroup	**Baseline**	85.9	0.814	81.6	92.4	57.9	84.4	87.8	85.6
	Follow-up	**88.6**	**0.850**	**84.4**	**95.6**	**60.4**	**87.9**	**90.0**	**88.1**
	Non-randomized	86.6	0.824	82.4	94.4	57.8	85.6	88.0	86.4

Acc = Accuracy, CHAT = Childhood Adenotonsillectomy Trial, CNN = Convolutional neural network, EEG = Electroencephalogram, Grad-CAM = Gradient-weighted class activation mapping, MF1 = macro F1-score, N1: level 1 of non-rapid eye movement (NREM) sleep, N2: level 2 of NREM sleep, N3: level 3 of NREM sleep, OSA=Obstructive Sleep Apnea, REM: rapid eye movement.

region to better visualize the time-frequency EEG patterns. The darker the color of the heatmap, the more important that region is in the final decision taken by the CNN. Notice that the heatmaps are highlighting well-known EEG features included in the scoring rules of the AASM related to each stage [2]: alpha rhythm and eye blinks (W); low-amplitude mixed frequency (LAMF) activity and vertex waves (N1); K-complexes and sleep spindles (N2); slow waves (N3); rapid eye movements and sawtooth waves (REM). Similarly, Grad-CAM heatmaps are also highlighting in the UofC set the same well-known EEG features related to each stage (see section 2.3 of the Supplementary Material), although this dataset does not contain the hypnogram to properly confirm the prediction.

Fig. 5 shows some interesting EEG patterns highlighted by Grad-CAM as important for the detection of W, N1, N2, and REM sleep. These EEG features, despite being important to sleep researchers, are not currently incorporated into the scoring rules of the AASM for identifying the various sleep stages. This suggests the potential use of these features to improve the reliability and reduce variability in sleep scoring. Regarding the epochs misclassified by the CNN, Figs. 6 and 7 present representative examples of Grad-CAM explanations corresponding to the most common errors observed in the confusion matrices. It is important to note that these misclassified epochs contain EEG features that correspond to both the predicted and the scored sleep stage from PSG. Furthermore, some of these epochs are located near sleep transitions, which introduces ambiguity among sleep technicians when scoring such epochs [35].

For the sake of completeness of the analysis, section 3 of the supplementary material also includes a visualization of the raw feature maps extracted by each layer of the CNN for the EEG segments in Fig. 3 (b) and (f). Although the raw feature maps may allow to discern which information is extracted in each layer of the CNN, its interpretation is far more difficult compared to the provided by Grad-CAM heatmaps. Despite showing the EEG patterns preserved in the convolutional layers (see Supplemental Figs. 4–11), the visualization of raw feature maps does not highlight the relative importance of this information in the predicted sleep stage, while Grad-CAM provides a class-discriminative localization of the EEG important features. Another key advantage of the proposed XAI analysis based on Grad-CAM is that it results in a single heatmap per segment, while the use of raw feature maps would require visualizing the feature maps extracted in each convolutional layer.

4. Discussion

This study aimed to develop an accurate deep-learning model based on CNN for sleep staging in pediatric OSA patients from one single EEG channel**.** We also provided an interpretation of the stage-related EEG features identified by the CNN model using a XAI approach based on Grad-CAM. Our XAI-based approach not only provided explanations for the EEG features considered by the CNN to predict each 30-s sleep stage but also enabled us to analyze uncertain epochs and propose novel patterns for sleep scoring based on single-channel EEG data. To our knowledge, this is the first explainable deep-learning approach applied to sleep scoring in pediatric OSA patients.

4.1. Pediatric sleep staging performance

The three proposed deep-learning architectures reached high performances to automatically classify sleep stages in pediatric OSA patients using a single EEG channel (C4-M1). Their 5-class Acc, kappa, and MF1 values ranged from 84.3% to 86.9%, 0.791 to 0.827, and 76.6%–82.7%, respectively. According to Lee et al. [36], inter-rater agreement in manual sleep scoring in adult patients, measured by kappa, is around 0.76 (95% CI 0.71–0.81), which highlights the usefulness of our approaches for automated pediatric sleep staging. Interestingly, the CNN model demonstrated the highest performance with an Acc of 86.9%, a kappa of 0.827, and a MF1 of 82.7%. In this respect, the higher performance of CNN *vs*. CNN-Inception suggests that a stack of convolutional layers with a larger filter size is more effective in extracting feature maps compared to the inception modules, which agrees with the findings of Supratak et al. [37]. Conversely, the slightly higher performance of CNN *vs*. CNN-RNN indicates that the contextual information required to assign each 30-s segment to a sleep stage can be captured adequately by considering just two preceding epochs and one posterior epoch, and that long-term interactions in the EEG do not significantly contribute to sleep staging.

Analyzing the per-class performance (see Table 2), the highest F1-scores were obtained in W stage, while the lowest ones were those

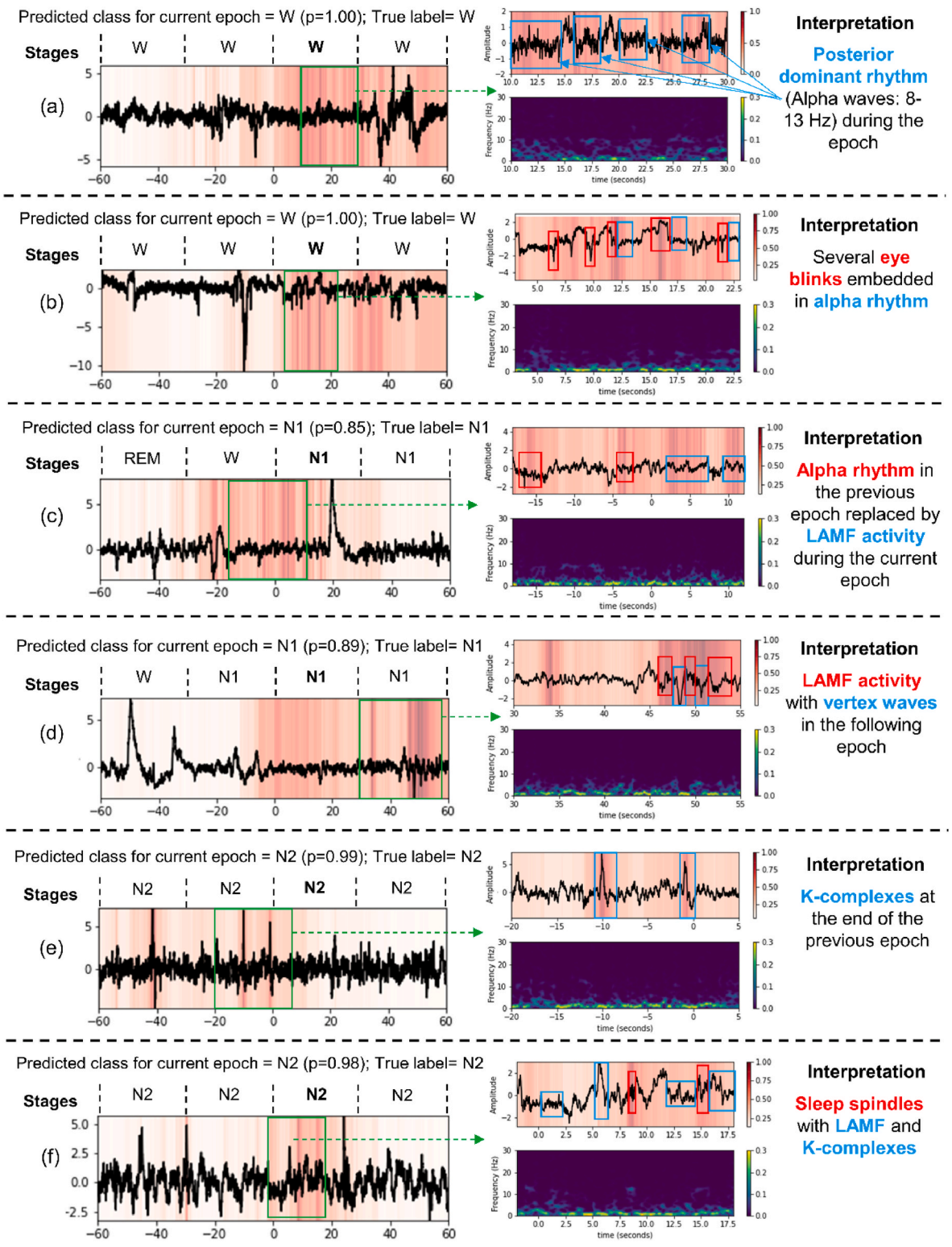

Fig. 3. Grad-CAM visualizations for some representative examples in the CHAT test set rightly predicted as: (a) W; (b) W; (c) N1; (d) N1; (e) N2; (f) N2.

corresponding to the N1 stage. This finding is consistent with state-of-the-art approaches in automatic sleep staging using single-channel EEG for both adult and pediatric OSA subjects [9–14,16]. Our results also align with the American Academy of Sleep Medicine (AASM) Inter-scorer Reliability Program for sleep stage scoring, which reported average agreement scores of 84.1% (Wake), 63.0% (N1), 85.2% (N2), 67.4% (N3), and 90.5% (REM) [35]. According to Rosenberg et al. [35], most of the disagreements occurred during transitions between sleep stages, which helps to explain the higher disagreement in N1 scoring and, consequently, the results obtained from automatic sleep staging approaches, as the N1 stage typically has a lower bout length (number of consecutive 30-s epochs scored as N1) compared to the other stages [35].

Regarding the epochs misclassified by the deep-learning models, Korkalainen et al. [9] and Phan et al. [16] also suggested that a significant portion of these epochs corresponds to sleep stage transitions.

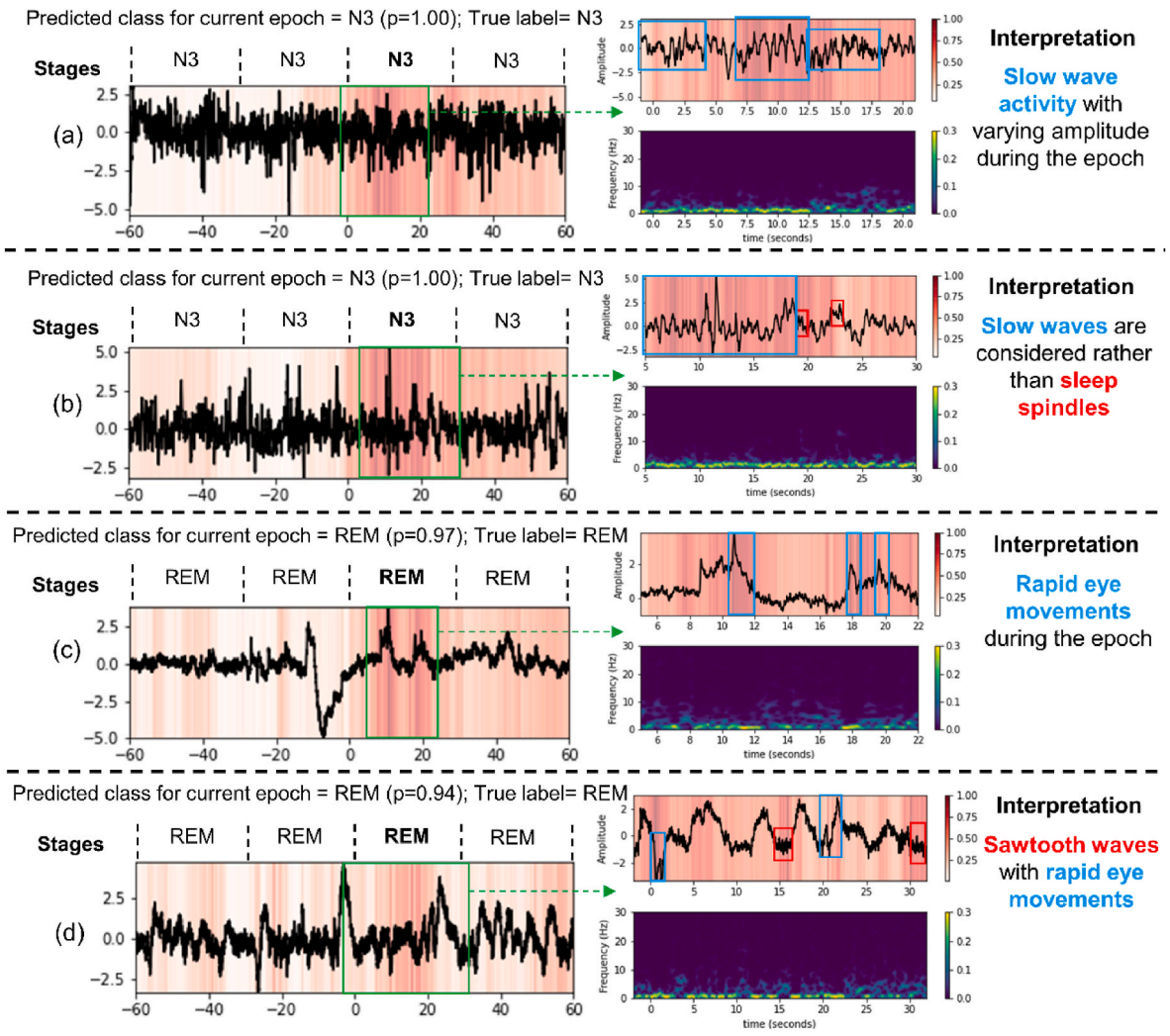

Fig. 4. Grad-CAM visualizations for some representative examples in the CHAT test set rightly predicted as (a) N3; (b) N3; (c) REM; (d) REM.

Particularly, Phan et al. [16] found that more than 60% of automatic pediatric sleep scoring errors occur within four epochs of a stage transition. To support this observation, Fig. 8 (a) displays the histogram illustrating the distance between epochs misclassified by the CNN model and the nearest predicted transitioning epoch, where transitioning epochs are defined as those in which the predicted sleep stage changes. Our analysis reveals that 29.2% of misclassified epochs are transitioning epochs, whereas 55.1% of the errors occur within one epoch of a sleep stage transition. Conversely, Fig. 8 (b) plots the distribution of the output probabilities of the CNN for misclassified epochs and all epochs. Here, the predicted probability distribution of misclassified epochs shows a notably different pattern, as derived from the interquartile range. Specifically, while 75% of the epochs are predicted with an output probability higher than 80%, nearly 75% of the misclassified epochs have a lower probability prediction, indicating a natural uncertainty threshold. Consequently, to further enhance the automatic results of pediatric sleep staging and improve the diagnostic process of pediatric OSA, sleep technicians should pay particular attention to epochs in close temporal proximity to sleep stage transitions and identify epochs with a low probability for the predicted stage using our proposed approach.

Regarding human subject's subgroup analysis, we found that the performance of our automatic approaches was not affected by either sex or OSA severity (see Table 3). This aligns with the findings of Bersch et al. [38], who reported that sex information does not improve sleep staging performance in adults, suggesting that the deep-learning models can effectively learn stage-related EEG patterns regardless of sex. Conversely, Korkalainen et al. [9] observed a decline in the performance of EEG-based automatic sleep staging with increasing OSA severity in adults, while Somaskandhan et al. [8] reported comparable performance between OSA symptomatic and control patients in a preadolescent age population (10–13 years). These findings confirm that pediatric OSA may not induce as pronounced changes in sleep macrostructure as seen in adults [39]. Regarding the external validation of our proposed approach, we obtained a high agreement (ICC = 0.772) between the estimated TST and actual TST from PSG in the UofC set, even though the hypnogram was not immediately annotated in the recordings. This underscores the potential usefulness of our approach to derive the sleep stages and overall TST in self-administrated PSG studies from single-channel EEG recordings, leading to an automatic diagnosis of pediatric OSA. Nonetheless, generalizability could be potentially improved in future studies by incorporating a broader range of pediatric sleep datasets encompassing different sleep disorders, as well as expanded datasets obtained from children across a wider age range.

4.2. *Explaining the decisions taken by the CNN*

There are some very recent studies proposing XAI approaches to identify those EEG patterns considered by deep-learning models to predict each sleep stage in adult subjects [7,11,24–27]. To the best of our knowledge, this is the first study explaining the decisions taken by

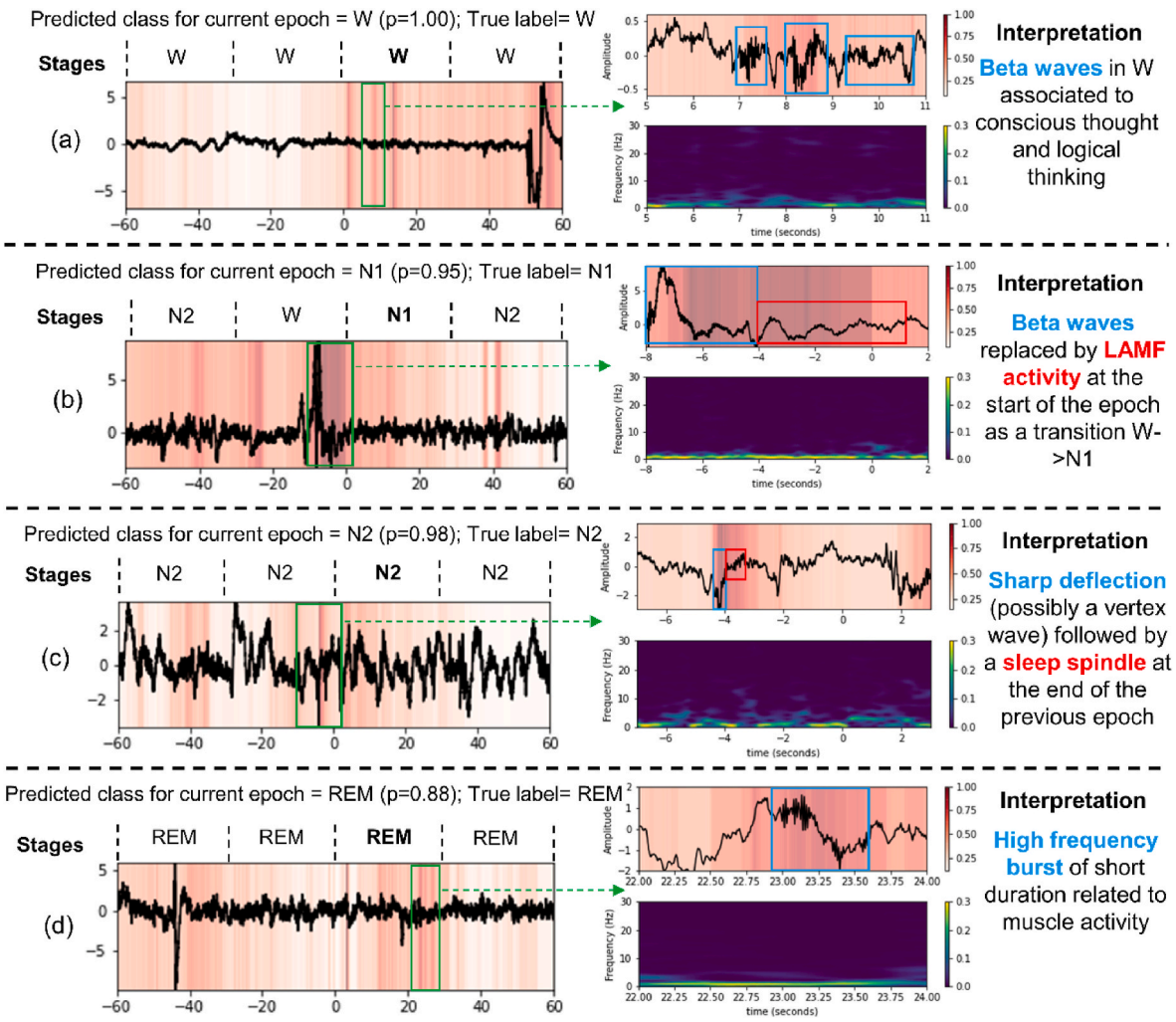

Fig. 5. Heatmaps for some interesting EEG patterns highlighted by Grad-CAM related to: (a) W; (b) N1; (c) N2; (d) REM.

an automatic deep-learning model for either pediatric sleep staging or pediatric sleep staging in OSA subjects. In this study, the proposed XAI approach based on Grad-CAM was not only used to identify those EEG patterns used by the CNN to predict each sleep stage, but also to identify and analyze epochs with a high likelihood to be misclassified, as well as to propose novel hallmarks for sleep stage detection in single-channel EEG overnight recordings. This automatic and explainable sleep scoring tool could be of great usefulness in the clinical practice for different reasons: (*i*) it would enable sleep technologists to visualize an interpret each predicted sleep stage, aligning with the recommendations of the EU regarding AI-based systems [22]; (*ii*) it could aid in the comprehensive annotation of 30-s epochs in independent sleep datasets, thereby improving the training of sleep scorers, minimizing inter-scorer variability, and enhancing the sleep scoring process by highlighting stage-related EEG patterns; (*iii*) it can be easily deployed on remote processing servers or portable monitoring devices using TensorFlow Lite [40], providing the sleep stage predictions per subject within seconds and the Grad-CAM heatmaps within minutes.

Grad-CAM has emerged as a widely used XAI algorithm in biomedical signals and images to provide visual explanations that highlight discriminative features relevant to CNN-based model predictions [41–43]. In the EEG signal processing field, Grad-CAM has been used not only to explain automatic sleep scoring models [7,26], but also to identify important EEG features associated with various healthcare applications such as cardiac arrest [44], schizophrenia [45], Alzheimer [46], emotion recognition [47,48], or brain computer interfaces [49, 50]. In contrast to these studies, which visualized Grad-CAM heatmaps at a specific layer of the network [7,26,44–50], we averaged Grad-CAM normalized and resized heatmaps obtained from all the layers. During the study design, we evaluated the impact of obtaining Grad-CAM heatmaps at the output of each convolutional layer. Our observations indicated that heatmaps from the initial layers predominantly focused on short-duration patterns, while those from the intermediate and final layers highlighted longer-duration EEG features. Consequently, we decided to generate averaged heatmaps across all layers. Looking at Figs. 3–7, it becomes evident that Grad-CAM heatmaps effectively outline EEG waveforms with different time-frequency characteristics. These include short-duration and low-frequency patterns (e.g., K-complexes or eye blinks), short-duration and high-frequency patterns (e.g., spindles or EMG bursts), long-duration and low-frequency patterns (e.g., slow or sawtooth waves), and long-duration and high-frequency patterns (e.g., beta waves or alpha rhythm). Consequently, our approach holds potential for detecting and analyzing EEG features related to cognitive development in children, such as sleep spindles, arousals, or cyclic alternating patterns [51–53]. Phan et al. [11], Kuo et al. [26], and Dutt et al. [7] have recently proposed the only three interpretable deep-learning sequence models for sleep staging, namely SleepTransformer, SNet, and SleepXAI, respectively. By leveraging attention scores at both the epoch and sequence level, Phan et al. [11] demonstrated that the transformer effectively highlights sleep-relevant EEG features in adult OSA patients. Conversely, Kuo et al. [26] and Dutt et al. [7] showed that Grad-CAM heatmaps derived from the last convolutional

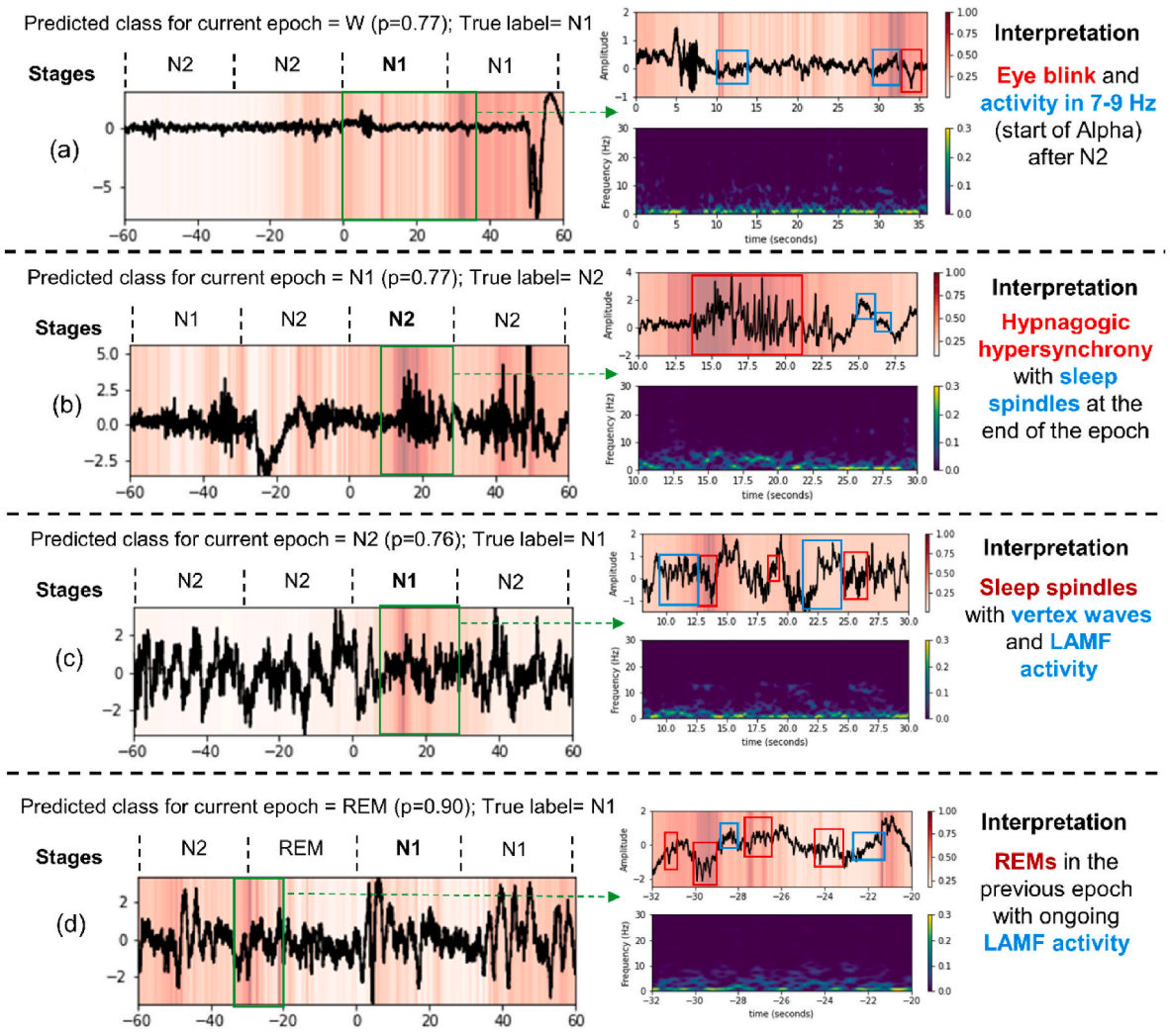

Fig. 6. Grad-CAM explanations corresponding to the following common errors in the CHAT test set made by the CNN: (a) N1 epoch predicted as W; (b) N2 epoch predicted as N1; (c) N1 epoch predicted as N2; (d) N1 epoch predicted as REM.

layer of SNet and SleepXAI, respectively, effectively identify the EEG patterns contributing to each predicted sleep stage in healthy subjects [7,26] and insomnia patients [26]. In light of these findings, future studies should explore comparisons between different XAI methods for pediatric sleep scoring.

Apart from those included in the standard sleep scoring rules, we have also identified some new EEG features extracted by the CNN that contribute to determine the sleep stage (see Fig. 5). These newly discerned EEG-related features for classifying sleep stages from a single EEG channel are summarized in Table 4. Interestingly, the AASM criteria only considers beta waves (13–30 Hz) for scoring arousals but not as an indicator of W or W-> N1 transitions [2], despite their prevalence in the wake state during conscious thought and logical thinking [54]. Furthermore, we have observed that spindles following sharp EEG deflections (probably a vertex wave) serve as strong indicators of N2 scoring. While vertex waves can also occur in N1 stage [2], spindles are commonly embedded in slow oscillations, such as K-complexes, which may also occur in slow wave sleep [55]. Therefore, these specific EEG patterns emerges prove highly valuable for sleep scoring. Lastly, short-duration high-frequency EEG bursts associated with scalp muscles can be used as strong indicators of REM sleep stage. This finding holds particular significance in situations where EMG derivations, which the AASM considers essential for defining REM sleep stage [2], are unavailable. Hence, XAI methods like Grad-CAM could contribute to improving sleep scoring rules and establishing new guideline criteria for scoring sleep stages based on single-channel EEG recordings.

The proposed explainability approach also enables the interpretation of epochs misclassified by the CNN. As previously mentioned, Grad-CAM heatmaps reveal that these epochs (see Figs. 6 and 7) contain EEG patterns related to different stages, as well as transitions between sleep stages. For example, these heatmaps illustrate instances where slow waves and K-complexes can be confounded, or where spindles are present in epochs not predicted as N2/N3 preceding/following N2 stage. In this context, our XAI approach can aid sleep technicians in reviewing manual scoring and assist in evaluating doubtful epochs, such as those occurring during sleep transitioning or epochs with a low probability of the predicted stage (see Fig. 8). Consequently, our approach contributes to improving the sleep scoring process. The presence of mixed patterns from several stages within a single 30-s epoch was also highlighted by Korkalainen et al. [56], who suggested to use epochs of shorter duration for sleep scoring in adult OSA patients, particularly for assessing sleep fragmentation. To this end, XAI methods can help define the optimal duration of sleep epochs or treat the whole recording as a continuum rather than dividing it into epochs [57,58].

4.3. Comparison with previous studies

There are multiple studies applying deep-learning algorithms for automatic sleep staging [15], including some that have used datasets from OSA patients [4–6,8–14,16]. Table 5 summarizes the comparison

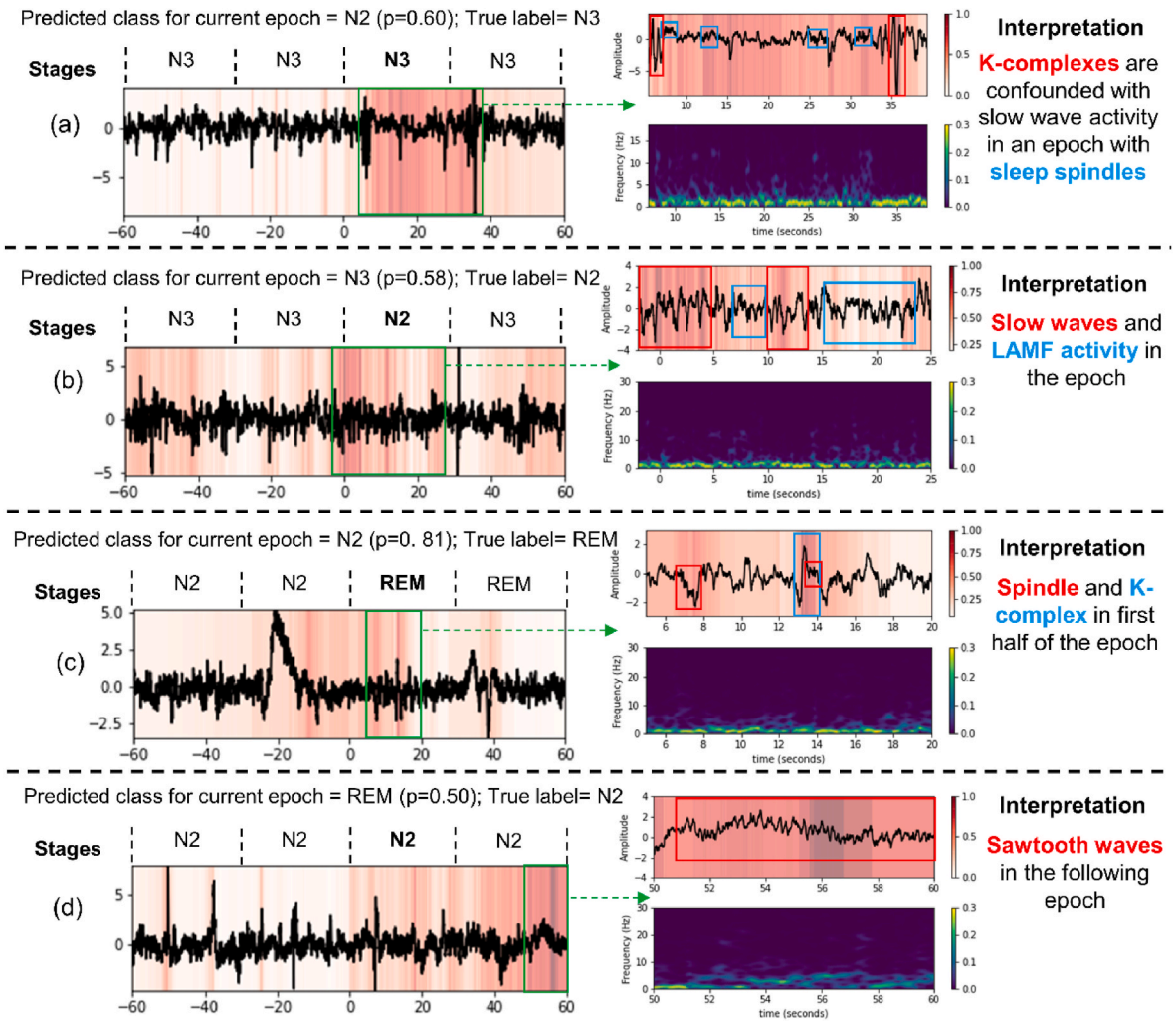

Fig. 7. Grad-CAM explanations corresponding to the following common errors in the CHAT test set made by the CNN: (a) N3 epoch predicted as N2; (b) N2 epoch predicted as N3; (c) REM epoch predicted as N2; (d) N2 epoch predicted as REM.

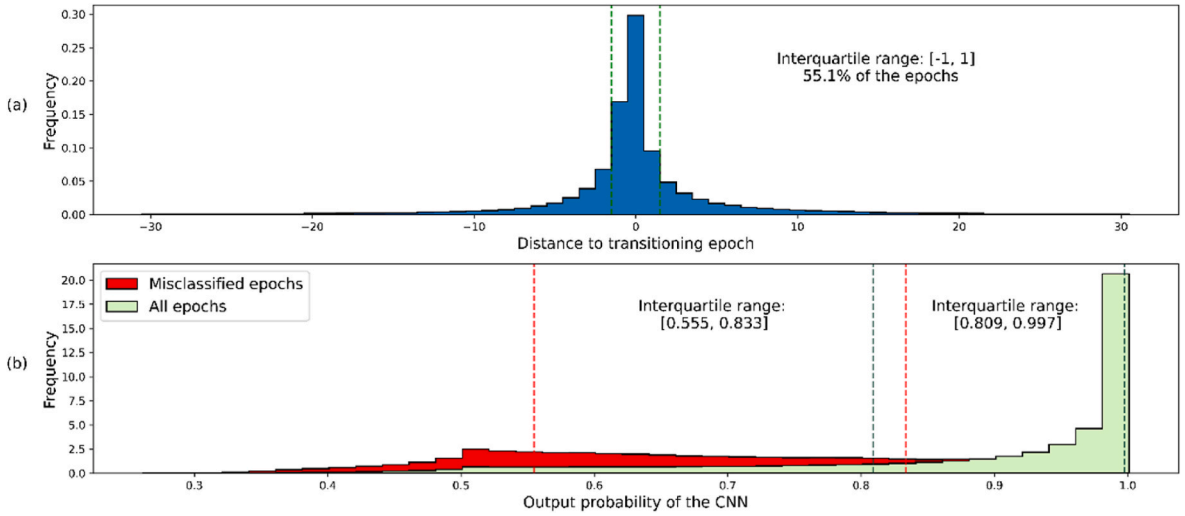

Fig. 8. Histograms of: (a) the distance of misclassified epochs to the closest transitioning epoch; (b) the output probability of the CNN for misclassified epochs and all epochs.

Table 4
Proposal of novel EEG-related patterns to distinguish sleep stages.

EEG pattern	Sleep stage	Interpretation
Beta waves	W	It can occur in children with eyes opened who are in conscious thought and logical thinking (e.g., scanning the environment)
Beta waves replaced by LAMF activity	N1	It can occur in the following situations: (*i*) children are falling asleep (transition W-> N1); (*ii*) arousal associated to an apnea/ hypopnea (transition N2/N3/REM-> N1)
Sharp EEG deflection followed by a spindle	N2	It is a strong indicator to score N2 stage instead of N1 stage (N1 contains vertex waves) or N3 (N3 contains spindles)
Short-duration high frequency burst	REM	It can be due to transient activity from the scalp muscles and are a stronger indicator of REM stage in absence of EMG derivations

EEG = electroencephalogram, EMG = electromyogram, LAMF = low-amplitude mixed-frequency, REM: rapid eye movement.

between our proposed approach and previous studies focused on the automatic detection of sleep stages using single-channel EEG tracings in both adult and pediatric OSA patients [6,9–13,16]. The first studies focused on sleep staging in adult OSA patients [9–13], reaching 5-class accuracies ranging from 82.9% to 87.7% and 5-class kappa values ranging from 0.77 to 0.828. Sors et al. [10] and Phan et al. [11] also provided insights into the interpretability of their deep-learning models. In contrast to these studies, our study provides the first explainable deep-learning model for sleep staging in pediatric OSA patients. Children exhibit distinct cardiorespiratory and neurophysiological activity during sleep, resulting in different patterns and sleep scoring criteria compared to adults [2].

Recently, Wang et al. [6] and Phan et al. [16] applied deep-learning algorithms for automatic pediatric OSA sleep staging from EEG [6,16]. It is noteworthy that Phan et al. [16] used the same CHAT database as our study. Using individual sequence models (CNN-RNN) previously pre-trained on an adult dataset [16], they reached a similar performance than our CNN-based approach in the follow-up (Acc: 88.3%–88.7% *vs.* 88.6%; kappa: 0.843–0.849 *vs.* 0.850) and non-randomized groups (Acc: 86.7%–87.0% *vs.* 86.6%; kappa: 0.822–0.828 *vs.* 0.824). Conversely, their performance slightly improved (by less than 1% Acc) when using an average ensemble of the 6 individual models, as well as when using an additional EOG channel. Moreover, they did not propose any new EEG sleep stage-related features.

In the present study, we contribute a new standard CNN model that achieves comparable performance to an ensemble of sequential deep-learning models while also being evaluated for estimating the TST in an external dataset. As a result, our proposal is easier to interpret, integrate, and test in portable monitoring devices with limited computational requirements. We also contribute here with an XAI analysis methodology offers insights into the EEG patterns considered by the CNN for predicting each sleep stage using Grad-CAM, including the interpretation of doubtful epochs. In addition, we propose novel EEG-related features for sleep scoring, thereby enhancing its clinical applicability.

4.4. Limitations and future work

It is important to acknowledge some limitations of our study. First, it is important to note that the interpretability and visualization approach based on Grad-CAM is not the only way to perform XAI analysis. While we have demonstrated the effectiveness of Grad-CAM heatmaps in identifying EEG patterns that contributing to stage predictions, future studies may explore alternative XAI and visualization techniques. Particularly, efforts to develop automated tools capable of deriving novel stage-related EEG patterns should be encouraged. Successive experiments should also validate the applicability of our explainable deep-learning model for widespread use in populations of all ages (both adults and children), as it is anticipated that the adult population would be more amenable to analysis compared to children, especially in cases involving OSA. In this respect, it would also be very interesting to integrate the proposed solution in a comprehensive software suite for EEG signal acquisition and processing such as Medusa ©, a novel open-source Python-based ecosystem developed by members of our research group that supports real-time processing and visualization [59]. Regarding external validation, the UofC dataset lacks annotation files with sleep stages. Therefore, obtaining additional annotated pediatric sleep datasets would be desirable to enhance the generalizability of our findings. Furthermore, ambulatory EEG recordings acquired with portable devices at home would further contribute to the broader applicability of our methodology. Another interesting future goal would be to assess the proposed methodology combining deep-learning and

Table 5
Diagnostic performance of state-of-the-art approaches in automatic sleep staging in both adult and pediatric OSA subjects from single-channel EEG.

			Methodology		Sleep staging metrics		
Study	Subjects		Deep learning	XAI	Acc (%)	kappa	MF1 (%)
Sors et al. [10]	5793 adults (SHHS)		CNN	Class-wise EEG patterns with synthetic inputs	86.8	0.810	78.5
Seo et al. [12]	5791 adults (SHHS)		CNN-RNN (IITNet)	–	86.7	0.81	79.8
Korkalainen et al. [9]	891 adults		CNN-RNN	–	82.9	0.77	–
Leino et al. [13]	135 adults		CNN-RNN	–	79.7	0.73	–
Phan et al. [11]	5791 adults (SHHS)		Transformer (SleepTransformer)	EEG heatmaps and epoch influence with self-attention weights	87.7	0.828	80.1
Wang et al. [6]	344 children		CNN	–	87.7	0.782	80.1
Phan et al. [16]	1626 children (CHAT)	Follow-up	Sequence models (CNN-RNN)	–	88.3–88.6	0.843–0.849	81.5–85.2
		Non-randomized	Sequence models (CNN-RNN)		86.7–87.0	0.822–0.828	80.0–83.6
		Follow-up	Average ensemble of 6 sequence models		89.2	0.857	85.3
		Non-randomized	Average ensemble of 6 sequence models		87.7	0.837	83.8
This study	**1637 children (CHAT)**	**Baseline**	**CNN**	**EEG heatmaps using GradCam: stage-related EEG features and error analysis**	**85.9**	**0.814**	**81.6**
		Follow-up	**CNN**		**88.6**	**0.850**	**84.4**
		Non-randomized	**CNN**		**86.6**	**0.824**	**82.4**

Acc = Accuracy, CHAT = Childhood Adenotonsillectomy Trial, CNN = Convolutional neural network, EEG = Electroencephalogram, Grad-CAM = Gradient-weighted class activation mapping, MF1 = macro F1-score, RNN=Recurrent neural network, SHHS= Sleep heart health study, XAI = Explainable artificial intelligence.

XAI to detect apnea/hypopnea events and subsequently identify novel EEG patterns related to apneas and hypopneas. Similarly, the proposed methodology could be extended to cardiorespiratory signals, which have been frequently proposed as a simplified alternative to PSG for the diagnosis of both adult [60,61] and pediatric [62,63] OSA.

5. Conclusion

In summary, we obtained an accurate CNN-based deep-learning model for automatic sleep staging in children while using a single channel EEG. Our model outperformed CNN-Inception and CNN-RNN architectures when evaluated on a database of 1637 EEG recordings. Furthermore, a XAI approach based on Grad-CAM allowed us to identify those EEG features associated with each predicted sleep stage. In particular, the specific hallmarks identified for sleep stage detection in the C4 EEG channel include beta waves (W), beta waves followed by LAMF activity (N1), vertex wave followed by a spindle (N2), and short-duration high frequency bursts (REM). Furthermore, Grad-CAM heatmaps enabled the identification and further analysis of epochs with a high likelihood to be misclassified, thereby facilitating the proposal of new criteria for sleep scoring that would reduce inter-rater variability and ambiguity. Additionally, we demonstrated that the CNN model can be used to estimate the TST in external unannotated sleep datasets, while also reliably identifying sleep stage-related EEG features. Our results show that the integrated collection of overnight single-channel EEG recordings and their automated processing by our explainable deep-learning model will yield a highly accurate, interpretable, and widely implementable tool for the automated detection of sleep stages in children with clinical suspicion of OSA. Future research is needed to further validate the applicability of our proposed solution for widespread use in populations of all ages, as well as incorporate the proposed solution in Medusa ©, a full-scale software application for EEG signal acquisition and processing. This will ultimately favor a timely and objective diagnosis of OSA.

Data availability

Polysomnography data from the CHAT database utilized in this study is available upon request through the National Sleep Research Resource website (https://www.sleepdata.org/datasets/chat). Sleep recordings from the UofC database are not publicly available but can be obtained upon reasonable request to the authors.

Authorship contribution statement

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

Ethical approval

This work has been carried out according to the Declaration of Helsinki. The clinical trial identifier of the CHAT database is available in NCT00560859 and a written consent for parental permission for the research was obtained from each pediatric subject as part of the research protocol, which can be found in the supplementary material of Marcus et al. [29]. In the UofC dataset, the legal caretakers of the children signed a written informed consent, and the Ethics Committee of the Comer Children's Hospital of the University of Chicago approved the protocol of the study (#11-0268-AM017, #09-115-B-AM031, and #IRB14-1241).

Declaration of competing interest

None Declared.

Acknowledgements

This work was supported by 'Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación/10.13039/501100011033/', ERDF A way of making Europe, and NextGenerationEU/PRTR under projects PID2020-115468RB-I00 and PDC2021-120775-I00, by 'Sociedad Española de Neumología y Cirugía Torácica (SEPAR)' under project 649/2018, 'Sociedad Española de Sueño (SES)' under project "Beca de Investigación SES 2019", and by 'CIBER -Consorcio Centro de Investigación Biomédica en Red-' (CB19/01/00012) through 'Instituto de Salud Carlos III', as well as under the project Tattoo4Sleep from 2022 CIBER-BBN Early Stage Plus call. 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). G. C. Gutiérrez-Tobal was supported by a post-doctoral grant from the University of Valladolid. D. Álvarez is supported by a "Ramón y Cajal" grant (RYC2019-028566-I) from the 'Ministerio de Ciencia e Innovación - Agencia Estatal de Investigación' co-funded by the European Social Fund. L. Kheirandish-Gozal and D. Gozal are supported by the Leda J. Sears Foundation for Pediatric Research.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2023.107419.

References

[1] M.J. Sateia, International classification of sleep disorders-third edition, Chest 146 (2014) 1387–1394, https://doi.org/10.1378/chest.14-0970.
[2] R.B. Berry, R. Brooks, C.E. Gamaldo, S.M. Harding, C.L. Marcus, B.V. Vaughn, The AASM manual for the scoring of sleep and associated events, Am. Acad. Sleep Med. 53 (2018) 1689–1699.
[3] L. Fiorillo, A. Puiatti, M. Papandrea, P.L. Ratti, P. Favaro, C. Roth, P. Bargiotas, C. L. Bassetti, F.D. Faraci, Automated sleep scoring: a review of the latest approaches, Sleep Med. Rev. 48 (2019), 101204, https://doi.org/10.1016/j.smrv.2019.07.007.
[4] A.N. Olesen, P. Jørgen Jennum, E. Mignot, H.B.D. Sorensen, Automatic sleep stage classification with deep residual networks in a mixed-cohort setting, Sleep 44 (2021) 1–12, https://doi.org/10.1093/sleep/zsaa161.
[5] M. Perslev, S. Darkner, L. Kempfner, M. Nikolic, P.J. Jennum, C. Igel, U-Sleep, Resilient high-frequency sleep staging, Npj Digit. Med. 4 (2021) 72, https://doi.org/10.1038/s41746-021-00440-5.
[6] H. Wang, G. Lin, Y. Li, X. Zhang, W. Xu, X. Wang, D. Han, Automatic sleep stage classification of children with sleep-disordered breathing using the modularized network, Nat. Sci. Sleep 13 (2021) 2101–2112, https://doi.org/10.2147/NSS.S336344.
[7] M. Dutt, S. Redhu, M. Goodwin, C.W. Omlin, SleepXAI: an explainable deep learning approach for multi-class sleep stage identification, Appl. Intell. (2022), https://doi.org/10.1007/s10489-022-04357-8.
[8] P. Somaskandhan, T. Leppänen, P.I. Terrill, S. Sigurdardottir, E.S. Arnardottir, K. A. Ólafsdóttir, M. Serwatko, S. Sigurðardóttir, M. Clausen, J. Töyräs, H. Korkalainen, Deep learning-based algorithm accurately classifies sleep stages in preadolescent children with sleep-disordered breathing symptoms and age-matched controls, Front. Neurol. 14 (2023) 1–12, https://doi.org/10.3389/fneur.2023.1162998.
[9] H. Korkalainen, T. Leppanen, J. Aakko, S. Nikkonen, S. Kainulainen, A. Leino, B. Duce, I.O. Afara, S. Myllymaa, J. Toyras, Accurate deep learning-based sleep staging in a clinical population with suspected obstructive sleep apnea, IEEE J. Biomed. Heal. Informatics. 24 (2019), https://doi.org/10.1109/JBHI.2019.2951346, 1–1.
[10] A. Sors, S. Bonnet, S. Mirek, L. Vercueil, J.F. Payen, A convolutional neural network for sleep stage scoring from raw single-channel EEG, Biomed. Signal Process Control 42 (2018) 107–114, https://doi.org/10.1016/j.bspc.2017.12.001.

[11] H. Phan, K.B. Mikkelsen, O. Chen, P. Koch, A. Mertins, M. De Vos, SleepTransformer: automatic sleep staging with interpretability and uncertainty quantification, IEEE Trans. Biomed. Eng. 69 (2022) 2456–2467, https://doi.org/10.1109/TBME.2022.3147187.
[12] H. Seo, S. Back, S. Lee, D. Park, T. Kim, K. Lee, Intra- and inter-epoch temporal context network (IITNet) using sub-epoch features for automatic sleep scoring on raw single-channel EEG, Biomed. Signal Process Control 61 (2020), 102037, https://doi.org/10.1016/j.bspc.2020.102037.
[13] A. Leino, H. Korkalainen, L. Kalevo, S. Nikkonen, S. Kainulainen, A. Ryan, B. Duce, K. Sipila, J. Ahlberg, J. Sahlman, T. Miettinen, S. Westeren-Punnonen, E. Mervaala, J. Toyras, S. Myllymaa, T. Leppanen, K. Myllymaa, Deep learning enables accurate automatic sleep staging based on ambulatory forehead EEG, IEEE Access 10 (2022) 26554–26566, https://doi.org/10.1109/ACCESS.2022.3154899.
[14] S. Biswal, H. Sun, B. Goparaju, M. Brandon Westover, J. Sun, M.T. Bianchi, Expert-level sleep scoring with deep neural networks, J. Am. Med. Inf. Assoc. 25 (2018) 1643–1650, https://doi.org/10.1093/jamia/ocy131.
[15] O. Faust, H. Razaghi, R. Barika, E.J. Ciaccio, U.R. Acharya, A review of automated sleep stage scoring based on physiological signals for the new millennia, Comput. Methods Progr. Biomed. 176 (2019) 81–91, https://doi.org/10.1016/j.cmpb.2019.04.032.
[16] H. Phan, A. Mertins, M. Baumert, Pediatric Automatic Sleep Staging: a comparative study of state-of-the-art deep learning methods, IEEE Trans. Biomed. Eng. 69 (2022) 3612–3622, https://doi.org/10.1109/TBME.2022.3174680.
[17] A. V Benjafield, P.R. Eastwood, R. Heinzer, M.J. Morrell, U. Federal, D.S. Paulo, S. Paulo, K. Valentine, Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis, Lancet Respir. Med. 7 (2020) 687–698, https://doi.org/10.1016/S2213-2600(19)30198-5.Estimation.
[18] T. Penzel, X. Zhang, I. Fietze, Inter-scorer reliability between sleep centers can teach us what to improve in the scoring rules, J. Clin. Sleep Med. 9 (2013) 89–91, https://doi.org/10.5664/jcsm.2352.
[19] A. Lo Bue, A. Salvaggio, G. Insalaco, Obstructive sleep apnea in developmental age. A narrative review, Eur. J. Pediatr. 179 (2020) 357–365, https://doi.org/10.1007/s00431-019-03557-8.
[20] C.L. Marcus, L.J. Brooks, K.A. Draper, D. Gozal, A.C. Halbower, J. Jones, M. S. Schechter, S.H. Sheldon, K. Spruyt, S.D. Ward, C. Lehmann, R.N. Shiffman, A. A. of Pediatrics, Diagnosis and management of childhood obstructive sleep apnea syndrome, Pediatrics 130 (2012) 576–584, https://doi.org/10.1542/peds.2012-1671.
[21] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160, https://doi.org/10.1109/ACCESS.2018.2870052.
[22] R. Hamon, H. Junklewitz, I. Sanchez, Robustness and Explainability of Artificial Intelligence, Publ. Off. Eur. Union, 2020.
[23] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, in: Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 618–626. http://arxiv.org/abs/1610.02391.
[24] M. Rossi, D. Sala, D. Bovio, C. Salito, G. Alessandrelli, C. Lombardi, L. Mainardi, P. Cerveri, SLEEP-SEE-THROUGH: explainable deep learning for sleep event detection and quantification from wearable somnography, IEEE J. Biomed. Heal. Informatics. 27 (2023) 3129–3140, https://doi.org/10.1109/JBHI.2023.3267087.
[25] A.R. Troncoso-García, M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, Explainable machine learning for sleep apnea prediction, Procedia Comput. Sci. 207 (2022) 2924–2933, https://doi.org/10.1016/j.procs.2022.09.351.
[26] C.E. Kuo, G.T. Chen, P.Y. Liao, An EEG spectrogram-based automatic sleep stage scoring method via data augmentation, ensemble convolution neural network, and expert knowledge, Biomed. Signal Process Control 70 (2021), 102981, https://doi.org/10.1016/j.bspc.2021.102981.
[27] L.D. Barnes, K. Lee, A.W. Kempa-Liehr, L.E. Hallum, Detection of sleep apnea from single-channel electroencephalogram (EEG) using an explainable convolutional neural network (CNN), PLoS One 17 (2022) 1–18, https://doi.org/10.1371/journal.pone.0272167.
[28] S. Redline, R. Amin, D. Beebe, R.D. Chervin, S.L. Garetz, B. Giordani, C.L. Marcus, R.H. Moore, C.L. Rosen, R. Arens, D. Gozal, E.S. Katz, R.B. Mitchell, H. Muzumdar, H.G. Taylor, N. Thomas, S. Ellenberg, The childhood adenotonsillectomy trial (CHAT): rationale, design, and challenges of a randomized controlled trial evaluating a standard surgical procedure in a pediatric population, Sleep 34 (2011) 1509–1517, https://doi.org/10.5665/sleep.1388.
[29] C.L. Marcus, R.H. Moore, C.L. Rosen, B. Giordani, S.L. Garetz, H.G. Taylor, R. B. Mitchell, R. Amin, E.S. Katz, R. Arens, S. Paruthi, H. Muzumdar, D. Gozal, N. H. Thomas, J. Ware, D. Beebe, K. Snyder, L. Elden, R.C. Sprecher, P. Willging, D. Jones, J.P. Bent, T. Hoban, R.D. Chervin, S.S. Ellenberg, S. Redline, A randomized trial of adenotonsillectomy for childhood sleep apnea, N. Engl. J. Med. 368 (2013) 2366–2376, https://doi.org/10.1056/NEJMoa1215881.
[30] C. Iber, S. Ancoli-Israel, A. Chesson, S.F. Quan, The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specification, J. Clin. Sleep Med. 3 (2007) 752, https://doi.org/10.1017/CBO9781107415324.004.
[31] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.A. Muller, Deep learning for time series classification: a review, Data Min. Knowl. Discov. 33 (2019) 917–963, https://doi.org/10.1007/s10618-019-00619-1.
[32] E. Santamaria-Vazquez, V. Martinez-Cagigal, F. Vaquerizo-Villar, R. Hornero, EEG-inception: a novel deep convolutional neural network for assistive ERP-based brain-computer interfaces, IEEE Trans. Neural Syst. Rehabil. Eng. 28 (2020) 2773–2782, https://doi.org/10.1109/TNSRE.2020.3048106.
[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, Arxiv (2015). http://arxiv.org/abs/1409.4842v1.
[34] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.
[35] R.S. Rosenberg, S. Van Hout, The American Academy of Sleep Medicine inter-scorer reliability program: respiratory events, J. Clin. Sleep Med. 10 (2014) 447–454, https://doi.org/10.5664/jcsm.3630.
[36] Y.J. Lee, J.Y. Lee, J.H. Cho, J.H. Choi, Interrater reliability of sleep stage scoring: a meta-analysis, J. Clin. Sleep Med. 18 (2022) 193–202, https://doi.org/10.5664/jcsm.9538.
[37] A. Supratak, Y. Guo, TinySleepNet, An efficient deep learning model for sleep stage scoring based on raw single-channelchannel EEG, in: 2020 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., IEEE, 2020, pp. 641–644, https://doi.org/10.1109/EMBC44109.2020.9176741.
[38] E. Bresch, U. Großekathöfer, G. Garcia-Molina, Recurrent deep neural networks for real-time sleep stage classification from single channelchannel EEG, Front. Comput. Neurosci. 12 (2018) 1–12, https://doi.org/10.3389/fncom.2018.00085.
[39] S. Scholle, G. Zwacka, Arousals and obstructive sleep apnea syndrome in children, Clin. Neurophysiol. 112 (2001) 984–991, https://doi.org/10.1016/S1388-2457(01)00508-9.
[40] TensorFlow Lite | ML on Mobile and Edge Devices, 2022. https://www.tensorflow.org/lite.
[41] H.W. Loh, C.P. Ooi, S. Seoni, P.D. Barua, F. Molinari, U.R. Acharya, Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022), Comput, Methods Programs Biomed 226 (2022), 107161, https://doi.org/10.1016/j.cmpb.2022.107161.
[42] B.H.M. van der Velden, H.J. Kuijf, K.G.A. Gilhuijs, M.A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, Med, Image Anal 79 (2022), https://doi.org/10.1016/j.media.2022.102470, 102470.
[43] G. Yang, Q. Ye, J. Xia, Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond, Inf. Fusion 77 (2022) 29–52, https://doi.org/10.1016/j.inffus.2021.07.016.
[44] S. Jonas, A.O. Rossetti, M. Oddo, S. Jenni, P. Favaro, F. Zubler, EEG-based outcome prediction after cardiac arrest with convolutional neural networks: performance and visualization of discriminative features, Hum. Brain Mapp. 40 (2019) 4606–4617.
[45] Z. Aslan, M. Akin, A deep learning approach in automated detection of schizophrenia using scalogram images of EEG signals, Phys. Eng. Sci. Med. 45 (2022) 83–96, https://doi.org/10.1007/s13246-021-01083-2.
[46] F.C. Morabito, C. Ieracitano, N. Mammone, An explainable Artificial Intelligence approach to study MCI to AD conversion via HD-EEG processing, Clin. EEG Neurosci. 54 (2023) 51–60.
[47] F. Wang, S. Wu, W. Zhang, Z. Xu, Y. Zhang, C. Wu, S. Coleman, Emotion recognition with convolutional neural network and EEG-based EFDMs, Neuropsychologia 146 (2020), 107506, https://doi.org/10.1016/j.neuropsychologia.2020.107506.
[48] B. Liu, J. Guo, C.L.P. Chen, X. Wu, T. Zhang, Fine-grained interpretability for EEG emotion recognition: concat-aided grad-CAM and systematic brain functional network, IEEE Trans. Affect. Comput. (2023).
[49] Y. Fujiwara, J. Ushiba, Deep residual convolutional neural networks for brain–computer interface to visualize neural processing of hand movements in the human brain, Front. Comput. Neurosci. 16 (2022), 882290.
[50] Y. Yan, H. Zhou, L. Huang, X. Cheng, S. Kuang, A novel two-stage refine filtering method for EEG-based motor imagery classification, Front. Neurosci. 15 (2021) 1–9, https://doi.org/10.3389/fnins.2021.657540.
[51] P.E. Brockmann, F. Damiani, E. Pincheira, F. Daiber, S. Ruiz, F. Aboitiz, R. Ferri, O. Bruni, Sleep spindle activity in children with obstructive sleep apnea as a marker of neurocognitive performance: a pilot study, Eur. J. Paediatr. Neurol. 22 (2018) 434–439, https://doi.org/10.1016/j.ejpn.2018.02.003.
[52] N. Li, J. Wang, D. Wang, Q. Wang, F. Han, K. Jyothi, R. Chen, Correlation of sleep microstructure with daytime sleepiness and cognitive function in young and middle-aged adults with obstructive sleep apnea syndrome, Eur. Arch. Oto-Rhino-Laryngol. 276 (2019) 3525–3532, https://doi.org/10.1007/s00405-019-05529-y.
[53] P.E. Brockmann, D. Gozal, Neurocognitive consequences in children with sleep disordered breathing: who is at risk? Children 9 (2022), 1278 https://doi.org/10.3390/children9091278.
[54] P.A. Abhang, B.W. Gawali, S.C. Mehrotra, Technical aspects of brain rhythms and speech parameters, in: Introd. To EEG-And Speech-Based Emot. Recognit., Elsevier, 2016, pp. 51–79, https://doi.org/10.1016/B978-0-12-804490-2.00003-8.
[55] J. Gomez-Pilar, G.C. Gutiérrez-Tobal, J. Poza, S. Fogel, J. Doyon, G. Northoff, R. Hornero, Spectral and temporal characterization of sleep spindles - methodological implications, J. Neural. Eng. 18 (2021), https://doi.org/10.1088/1741-2552/abe8ad.
[56] H. Korkalainen, T. Leppanen, B. Duce, S. Kainulainen, J. Aakko, A. Leino, L. Kalevo, I.O. Afara, S. Myllymaa, J. Toyras, Detailed assessment of sleep architecture with deep learning and shorter epoch-to-epoch duration reveals sleep fragmentation of patients with obstructive sleep apnea, IEEE J. Biomed. Heal. Informatics. 25 (2021) 2567–2574, https://doi.org/10.1109/JBHI.2020.3043507.
[57] J.B. Stephansen, A.N. Olesen, M. Olsen, A. Ambati, E.B. Leary, H.E. Moore, O. Carrillo, L. Lin, F. Han, H. Yan, Y.L. Sun, Y. Dauvilliers, S. Scholz, L. Barateau, B. Hogl, A. Stefani, S.C. Hong, T.W. Kim, F. Pizza, G. Plazzi, S. Vandi, E. Antelmi, D. Perrin, S.T. Kuna, P.K. Schweitzer, C. Kushida, P.E. Peppard, H.B.D. Sorensen, P. Jennum, E. Mignot, Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy, Nat. Commun. 9 (2018) 1–15, https://doi.org/10.1038/s41467-018-07229-3.

[58] H. Phan, K. Mikkelsen, Automatic sleep staging of EEG signals: recent development, challenges, and future directions, Physiol. Meas. 43 (2022), https://doi.org/10.1088/1361-6579/ac6049.
[59] E. Santamaría-Vázquez, V. Martínez-Cagigal, D. Marcos-Martínez, V. Rodríguez-González, S. Pérez-Velasco, S. Moreno-Calderón, R. Hornero, MEDUSA©: a novel Python-based software ecosystem to accelerate brain-computer interface and cognitive neuroscience research, Comput. Methods Progr. Biomed. 230 (2023), https://doi.org/10.1016/j.cmpb.2023.107357.
[60] F.R. Mashrur, M.S. Islam, D.K. Saha, S.M.R. Islam, M.A. Moni, SCNN: scalogram-based convolutional neural network to detect obstructive sleep apnea using single-lead electrocardiogram signals, Comput. Biol. Med. 134 (2021), 104532, https://doi.org/10.1016/j.compbiomed.2021.104532.
[61] M. Sharma, D. Kumbhani, J. Tiwari, T.S. Kumar, U.R. Acharya, Automated detection of obstructive sleep apnea in more than 8000 subjects using frequency optimized orthogonal wavelet filter bank with respiratory and oximetry signals, Comput. Biol. Med. 144 (2022), 105364, https://doi.org/10.1016/j.compbiomed.2022.105364.
[62] J. Jiménez-García, M. García, G.C. Gutiérrez-Tobal, L. Kheirandish-Gozal, F. Vaquerizo-Villar, D. Álvarez, F. del Campo, D. Gozal, R. Hornero, A 2D convolutional neural network to detect sleep apnea in children using airflow and oximetry, Comput. Biol. Med. 147 (2022), https://doi.org/10.1016/j.compbiomed.2022.105784.
[63] A. Martín-Montero, P. Armañac-Julián, E. Gil, L. Kheirandish-Gozal, D. Álvarez, J. Lázaro, R. Bailón, D. Gozal, P. Laguna, R. Hornero, G.C. Gutiérrez-Tobal, 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.