Engineering Applications of Artificial Intelligence 162 (2025) 112562

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

An explainable deep learning approach for sleep staging in sleep apnea patients across all age subgroups from pulse oximetry signals

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

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

ARTICLE INFO

Keywords:
Age subgroups
Deep learning
Explainable artificial intelligence
Pulse oximetry
Obstructive sleep apnea
Sleep stages

ABSTRACT

Deep-learning (DL) approaches have been developed using pulse rate (PR) and blood oxygen saturation (SpO_2) recordings from pulse oximetry to streamline sleep staging, particularly for obstructive sleep apnea (OSA) patients. However, lack of interpretability and validation across patients from a wide range of ages (children, adolescents, adults, and elderly OSA individuals) are two major concerns. In this study, a DL model based on the U-Net framework (POxi-SleepNet) was tailored to accurately perform 4-class sleep staging (wake, light sleep, deep sleep, and rapid-eye movement sleep) in OSA patients across all age subgroups using PR and SpO_2 signals. An explainable artificial intelligence (XAI) methodology based on semantic segmentation via gradient-weighted class activation mapping (Seg-Grad-CAM) was also applied to quantitatively interpret the time and frequency characteristics of pulse oximetry recordings that influence sleep stage classification. Overnight PR and SpO_2 signals from 17303 sleep studies from six datasets encompassing children, adolescents, adults, and elderly OSA individuals were used. POxi-SleepNet showed high performance for sleep staging in the six databases, with accuracies between 81.5 % and 84.5 % and Cohen's kappa values from 0.726 to 0.779. It also demonstrated greater generalizability than previous studies. XAI analysis showed the key contributions of mean and variability in PR and SpO_2 amplitude, as well as changes in their spectral content across specific frequency bands (0.004–0.020 Hz, 0.020–0.100 Hz, and 0.180–0.400 Hz), for sleep stage classification. These findings indicate that POxi-SleepNet could effectively automate sleep staging and assist in diagnosing OSA across all age groups in clinical settings.

1. Introduction

Identification of sleep stages is crucial for the assessment and diagnosis of sleep-related disorders (Sateia, 2014). According to the guidelines established by the American Academy of Sleep Medicine (AASM), each 30-s non-overlapping epoch of sleep recordings must be classified as: wake (W), three levels of non-rapid eye movement (non-REM) sleep (N1, N2, and N3), or REM sleep (Berry et al., 2018). Sleep technologists conduct this scoring manually by analyzing electroencephalography (EEG), electrooculography (EOG), and submental electromyography signals, along with cardiorespiratory data collected during overnight polysomnography (PSG) (Berry et al., 2018). Despite being the gold standard, PSG is complex, expensive, highly intrusive, time-consuming, and of limited availability, which delays the diagnosis of sleep disorders (Chang et al., 2023). In addition, manual sleep staging is a tedious and time-intensive procedure, taking on average 1.5–2 h per PSG study, and is subject to considerable inter-scorer variability, thereby potentially compromising diagnostic accuracy (Fiorillo et al., 2019). Therefore, the adoption of automated sleep staging, utilizing a minimal number of channels, is recommended to improve consistency, streamline the process, and lower associated costs.

Recent progress in artificial intelligence (AI) methodologies have led

* Corresponding author. Biomedical Engineering Group, Facultad de Medicina, Av. Ramón y Cajal, 7, 47003, Valladolid, Spain.http://www.gib.tel.uva.es
E-mail addresses: fernando.vaquerizo@uva.es (F. Vaquerizo-Villar), gonzalocesar.gutierrez@uva.es (G.C. Gutiérrez-Tobal), daniel.alvarez.gonzalez@uva.es (D. Álvarez), adrian.martin@uva.es (A. Martín-Montero), gozal@marshall.edu (D. Gozal), roberto.hornero@uva.es (R. Hornero).

https://doi.org/10.1016/j.engappai.2025.112562
Received 25 September 2024; Received in revised form 21 May 2025; Accepted 29 September 2025
Available online 4 October 2025

to the proposal of multiple alternatives for sleep staging relying on the automated processing of a reduced subset of PSG signals (Faust et al., 2019). Among these methods, pulse oximetry has been extensively researched as a surrogate for PSG in sleep scoring and the diagnosis of sleep disorders (Baumert et al., 2022; del Campo et al., 2018; Imtiaz, 2021), as it can monitor individuals at their home using low-cost wearable devices. These devices consist of a non-invasive watch or finger probe equipped with an optical sensor that measures the photoplethysmography (PPG) signal, which is utilized to calculate pulse rate (PR) and blood oxygen saturation (SpO_2) (Chan et al., 2013). Overnight SpO_2 and PR raw data are stored by the pulse oximeters, facilitating offline analysis for the assessment of sleep quality and any other relevant abnormalities. The time and frequency characteristics of PPG and derived PR and SpO_2 signals exhibit changes across different sleep stages, prompting recent investigations into the application of AI methods for the automatic scoring of sleep stages using only pulse oximetry signals (Baumert et al., 2022; Imtiaz, 2021).

A considerable proportion of these works have focused on patient cohorts with obstructive sleep apnea (OSA) (Huttunen et al., 2021, 2022; Korkalainen et al., 2020; Kotzen et al., 2023; Radha et al., 2021; Sridhar et al., 2020; Vaquerizo-Villar et al., 2024; Wulterkens et al., 2021), a prevalent condition impacting approximately 1 billion individuals worldwide and affecting the entire lifespan (Benjafield et al., 2020; Chang et al., 2023). The diagnosis of OSA depends on the apnea-hypopnea index (AHI) that measures the number of apneas and hypopneas per hour of sleep (Berry et al., 2018; Chang et al., 2023), thereby emphasizing the importance of accurately scoring sleep stages in order to calculate the total sleep time (TST) and the clustering of any particular abnormality within a given sleep stage. Several investigations have targeted sleep scoring in adult (Huttunen et al., 2021, 2022; Korkalainen et al., 2020; Radha et al., 2021; Sridhar et al., 2020) or pediatric OSA (Haimov et al., 2025a; Vaquerizo-Villar et al., 2024) cohorts using pulse oximetry signals, but only Kotzen et al. (2023), Nam et al. (2024) and Wulterkens et al. (2021) approached sleep staging in patients across various age groups. However, the number of children/adolescents ($n = 73$ (3.1 %) out of 2380 (Kotzen et al., 2023), $n = 80$ (3.2 %) out of 2488 (Nam et al., 2024), and $n = 54$ (6.5 %) out of 835 (Wulterkens et al., 2021)), was limited. Including a larger proportion of pediatric individuals is essential because pediatric OSA subjects have specific considerations in terms of etiology, diagnosis, and treatment compared to adults (Chang et al., 2023), alongside substantial developmental differences in sleep architecture (e.g., higher proportion of deep sleep) and cardiorespiratory activity (e.g., less recurrent an profound desaturations and bradycardias) (Berry et al., 2018; Goh et al., 2000; Guilleminault et al., 1984). Given that OSA affects the entire lifespan, there is an inherent interest in developing automatic sleep scoring models that encompasses all individuals irrespective of their age.

In the last few years, automatic sleep scoring is living substantial progresses, largely attributed to advancements in deep-learning (DL) approaches, which have demonstrated efficacy in automatically learning stage-related features from raw data (Faust et al., 2019). Initial DL-based approaches performed sleep staging mainly using convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which were fed with sequences of 30-s epochs from PPG, PR and/or SpO_2 recordings (Huttunen et al., 2021; Korkalainen et al., 2020). More recent studies have shifted towards utilizing entire overnight recordings for sleep staging, thereby enabling the consideration of the whole-night dynamics of pulse oximetry during sleep to score sleep stages in each input sample (Casal et al., 2021, 2022; Huttunen et al., 2022; Kotzen et al., 2023; Sridhar et al., 2020; Vaquerizo-Villar et al., 2024). These studies focused on encoder-decoder networks comprising convolutional layers, inspired on the U-Net framework (Ronneberger et al., 2015), initially designed for image segmentation, yet adaptable for point-wise prediction (i.e., per-sample identification) of sleep stages (Fiorillo et al., 2023a; Huttunen et al., 2022; Perslev et al., 2021; Sridhar et al., 2020; Vaquerizo-Villar et al., 2024). Perslev et al. (2021) and Fiorillo et al. (2023a) also showed the generalizability of U-Net architectures for sleep scoring across various age groups and sleep disorders from whole-night EEG recordings.

Despite the success of DL algorithms in sleep staging, their "black box" nature hinders acceptance in healthcare (Adadi and Berrada, 2018; Yang et al., 2022), where understanding the rationale behind predictions is crucial. This challenge extends beyond sleep staging; for instance, in health recommendation systems, the lack of transparency can impede trust and adoption, even when the system offers accurate predictions (Chinnasamy et al., 2023a). In this context, explainable artificial intelligence (XAI) methodologies have recently emerged to introduce transparent and understandable recommendations to healthcare professionals and patients based on complex AI-based models (including DL) (Adadi and Berrada, 2018; Yang et al., 2022). In the field of sleep staging, several recent works have proposed XAI methodologies to visually discern those EEG patterns considered by DL models (Dutt et al., 2022; Kuo et al., 2021; Phan et al., 2022; Vaquerizo-Villar et al., 2023) for sleep staging. Nonetheless, only Nam et al. (2024) and Vaquerizo-Villar et al. (2024) have used XAI approaches to interpret DL models aimed at sleep staging from pulse oximetry signals. In this respect, a recent study by Nam et al. (2024) analyzed the local attention module to provide interpretability to the DL model predictions. Similarly, a recent conference paper developed by our own group signals (Vaquerizo-Villar et al., 2024) has provided some qualitative insights on the sleep scoring process through the application of semantic segmentation via gradient-weighted class activation mapping (Seg-Grad-CAM), a common XAI algorithm for semantic segmentation tasks (e.g., sleep staging) in CNN-based models (Vinogradova et al., 2020). However, these works offered only a visual XAI-derived interpretation of some selected examples from a DL model targeting sleep staging in OSA cohorts (Nam et al., 2024; Vaquerizo-Villar et al., 2024), without providing quantitative evidence of the pulse oximetry characteristics linked with each sleep stage. A quantitative analysis interpretation could not only align the model's behavior with the current knowledge of pulse oximetry features related to each sleep stage, while also uncovering novel patterns related to these stages.

Based on the aforementioned factors, we hypothesized that the application of a U-Net DL model, along with the Seg-Grad-CAM XAI algorithm, will yield models that are not only highly accurate and generalizable but also interpretable, making them clinically applicable for automated sleep stage staging in individuals of all age groups. Consequently, the main objectives of this study were twofold: (*i*) to develop and evaluate an interpretable and generalizable DL model capable of accurately classifying W, light sleep (N1 and N2), deep sleep (N3), and REM sleep stages within the OSA context in subjects of all ages from pulse oximetry data; (*ii*) to quantitatively identify key time and frequency-domain features from PR and SpO_2 signals that contribute to sleep staging.

Fig. 1 presents a general scheme of the proposed approach. The present study presents the following major contributions:

- A new DL model based on the U-Net architecture, POxi-SleepNet, is developed for sleep scoring in OSA patients using whole-night PR and SpO_2 signals.
- A comprehensive evaluation of the generalizability of POxi-SleepNet across six independent databases encompassing all age subgroups (children, adolescents, adults, and elderly).
- An XAI methodology relying on Seg-Grad-CAM to identify the key regions of overnight PR and SpO_2 signals from each patient that POxi-SleepNet models consider to predict W, light sleep, deep sleep, and REM sleep stages.
- A quantitative time- and frequency-domain analysis of the key regions highlighted by Seg-Grad-CAM to discern the PR and SpO_2 features associated with each sleep stage.

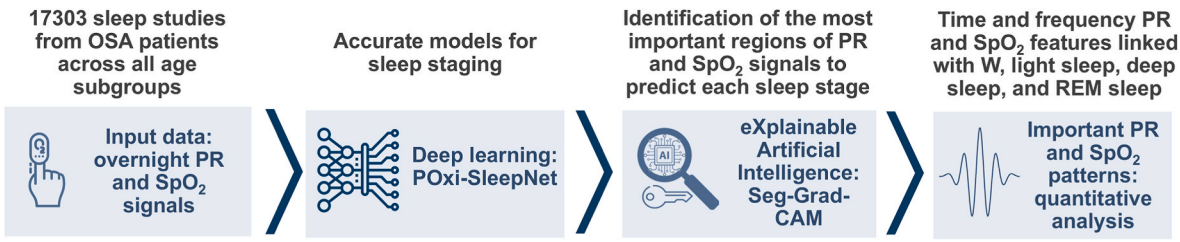

Fig. 1. Flowchart of the proposed methodology.

2. Methods

2.1. Subjects and signals

A total of 17303 PSG studies with valid pulse oximetry (PR and SpO_2) signals from 13222 participants of six different semi-public datasets composed the population under study: (*i*) the Cleveland Children's Sleep and Health Study (CCSHS), containing 515 recordings of adolescents aged 16–19.9 years (Rosen et al., 2003); (*ii*) the Cleveland Family Study (CFS), totaling 730 recordings of patients aged 7–89 years (Redline et al., 1995); (*iii*) the Childhood Adenotonsillectomy Trial (CHAT), which includes 1633 valid pulse oximetry recordings from pediatric subjects aged 5–9.9 years (Marcus et al., 2013); (*iv*) the Multi-Ethnic Study of Atherosclerosis (MESA), totaling 2056 recordings of patients aged over 54 years (Chen et al., 2015); (*v*) the Osteoporotic Fractures in Men Study (MrOS), containing 3925 recordings of patients aged over 67 years (Blackwell et al., 2011); and (*vi*) the Sleep Heart Health Study (SHHS), containing 8444 recordings of patients aged over 40 years (Quan et al., 1997). All these databases are publicly available upon request on the National Sleep Research Resource (NSRR) repository (https://sleepdata.org/datasets). The main information regarding rationale, design, primary outcomes, and sleep recordings of these databases can be consulted in sections 1.1-1.6 of the supplementary material, while the full details are available on the specific database websites at NSRR (CCSHS: https://sleepdata.org/datasets/ccshs; CFS: https://sleepdata.org/datasets/cfs; CHAT: https://sleepdata.org/datasets/chat; MESA: https://sleepdata.org/datasets/mesa; MROS: https://sleepdata.org/datasets/mros; SHHS: https://sleepdata.org/datasets/shhs).

All sleep studies (i.e., PSGs) were performed to confirm or discard OSA according to AASM guidelines (Iber et al., 2007), being the full details of PSG recordings available on the specific database websites at NSRR. Specifically, each sleep study from the six databases (CCSHS, CFS, CHAT, MESA, MrOS, and SHHS) provides PSG-derived PR and SpO_2 data, as well as annotations for sleep stages and apnea/hypopnea events (see sections 1.1-1.6 of the supplementary material). Each dataset was randomly split, on a per-subject basis, into three sets: (*i*) training (up to 50 % of subjects), used to train the POxi-SleepNet models; (*ii*) validation (approximately 25 % of the amount of training subjects), used for adjusting regularization and monitoring model convergence; and (*iii*) test (at least 50 % of subjects), used for performance evaluation and interpretation. Table 1 shows the clinical and polysomnographic data of the six databases analyzed, with further information provided in sections 1.1-1.6 of the supplementary material.

2.2. Signal preprocessing

PR and SpO_2 recordings, initially recorded at sampling frequencies (*fs*) from 1 to 512 Hz, were first resampled to a uniform *fs* of 1 Hz (Vaquerizo-Villar et al., 2024). Subject-specific standardization was then applied to normalize the baseline levels of PR and SpO_2 signals across different subjects (Casal et al., 2021). In accordance with the input size of the POxi-SleepNet models and the total recording time of the sleep studies (see Table 1), all PR and SpO_2 standardized recordings were finally padded or truncated to a common length of 12 h ($L = 43200$ samples) (Kotzen et al., 2023; Vaquerizo-Villar et al., 2024). Zero-padding or truncation was only performed at the onset of each pulse oximetry recording, with a duration of 12 h selected to ensure that only the initial wake periods were either added or removed. Each sample was then assigned to one of the sleep stages—wake, light sleep, deep sleep, or REM sleep—based on the annotations for the corresponding 30-s epochs.

2.3. DL architecture: POxi-SleepNet

POxi-SleepNet leverages the U-Net architecture, originally designed for image segmentation (Ronneberger et al., 2015), but which has shown its usefulness for sleep staging (Huttunen et al., 2022; Perslev et al., 2021; Sridhar et al., 2020). Fig. 2 illustrates the specific configuration of the POxi-SleepNet architecture used in this study for sleep

Table 1
Clinical and polysomnographic data of the six databases under study.

	CCSHS	CFS	CHAT	MESA	MrOS	SHHS
Subjects (*n*)	515	730	1633	2056	3925	8444
Age (years)	18 [17–18]	44 [23–55]	7 [6–8]	68 [62–76]	77 [73–82]	65 [56–73]
Males (*n*)	260 (50.5)	329 (45.1)	776 (47.5)	954 (46.4)	3925 (100)	3986 (47.2)
AHI (e/h)	0.8 [0.3–1.8]	5.2 [1.7–15.9]	2.5 [1.1–5.9]	18.4 [9.5–33.3]	17.0 [9.0–30.0]	13.3 [6.7–24.0]
TRT (h)	11.2 [10.8–11.7]	9.9 [9.3–10.4]	9.8 [9.1–10.8]	10.0 [9.5–12.0]	11.2 [10.0–12.4]	8.8 [8.5–9.1]
Wake (%)	29.3 [23.5–35.4]	35.8 [27.9–44.3]	22.8 [15.8–29.4]	41.7 [33.1–50.8]	46.6 [38.6–55.4]	30.4 [22.3–39.3]
Light (%)	39.4 [34.3–44.2]	38.8 [31.1–45.7]	38.7 [33.0–44.5]	41.3 [34.6–48.2]	37.1 [30.7–43.6]	43.1 [35.3–50.7]
Deep (%)	15.3 [12.0–19.2]	11.8 [6.3–17.8]	23.1 [19.5–27.8]	4.5 [1.1–9.1]	4.2 [1.3–8.3]	11.0 [5.1–16.9]
REM (%)	14.5 [11.8–17.5]	11.5 [8.0–15.1]	14.1 [11.6–16.8]	10.4 [7.0–13.8]	9.9 [7.0–13.0]	13.8 [10.1–17.4]
Training (*n*)	200 (38.8 %)	271 (37.1 %)	575 (35.2 %)	768 (37.4 %)	1409 (35.9 %)	2368 (28.0 %)
Validation (*n*)	57 (11.1 %)	94 (12.8 %)	200 (12.2 %)	260 (12.6 %)	467 (11.9 %)	778 (9.2 %)
Test (*n*)	258 (50.1 %)	365 (50.0 %)	858 (52.5 %)	1028 (50.0 %)	2049 (52.2 %)	5298 (62.7 %)

Data are presented as median [interquartile range], *n*, or *n* (%).
AHI: apnea-hypopnea index; CCSHS: Cleveland Children's Sleep and Health Study; CFS: Cleveland Family Study; CHAT: Childhood Adenotonsillectomy Trial; e/h: events per hour, h: hour; MESA: Multi-Ethnic Study of Atherosclerosis, MrOS: Osteoporotic Fractures in Men Study; REM: rapid eye movement; SHHS: Sleep Heart Health Study; TRT: total recording time.

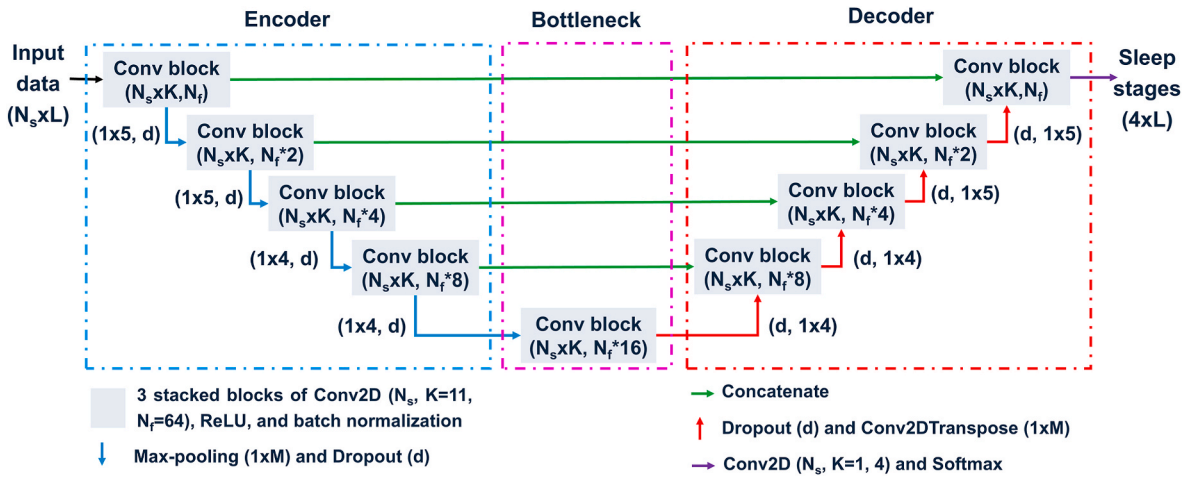

Fig. 2. Overview of the proposed POxi-SleepNet architecture for sleep staging.

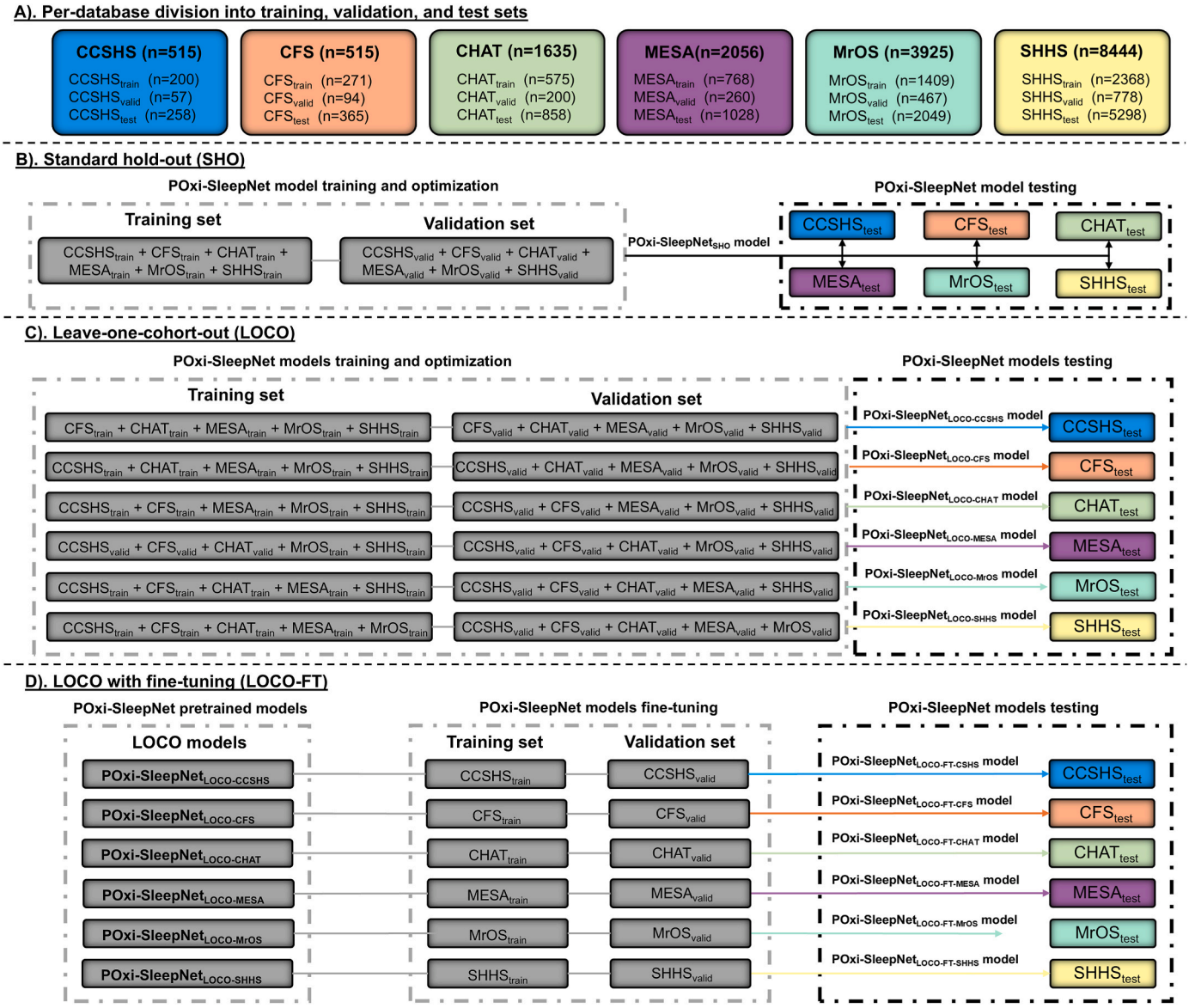

Fig. 3. Proposed approach for assessing the performance and generalization ability of the POxi-SleepNet architecture. (A) Division of the six databases (CCSHS, CFS, CHAT, MESA, MrOS, and SHHS) into training, validation, and test sets; (B) Standard hold-out strategy; (C) Leave-one-cohort-out (LOCO) without a fine-tuning strategy; (D) LOCO with fine-tuning (LOCO-FT) strategy. train: training; valid: validation.

staging, which is based on previous works (Huttunen et al., 2022; Perslev et al., 2021). In this study, three POxi-SleepNet architectures were developed with different input signal configurations: (*i*) POxi-SleepNet$_{PR\text{-}SpO2}$ (PR and SpO_2 data); (*ii*) POxi-SleepNet$_{PR}$ (only PR data); (*iii*) POxi-SleepNet$_{SpO2}$ (only SpO_2 data). Thus, the input section of POxi-SleepNet consists of N_sxL samples (12-h of data), being $N_s = 1$ (POxi-SleepNet$_{PR}$ and POxi-SleepNet$_{SpO2}$) or $N_s = 2$ (POxi-SleepNet$_{PR\text{-}SpO2}$).

The input is initially processed by an encoder composed of four layers, aimed at extracting low-level features related to sleep stages from PR and SpO_2 signals. Each layer comprises a convolutional block (conv block), followed by max-pooling and dropout layers. Within each conv block, there are three sequential sub-blocks, each performing a 2D convolution with *Nf* filters and a kernel size of N_sxK, as well as applying a rectified linear unit (ReLU) activation function and batch normalization. According to the standard U-Net design, *Nf* is initially set as 64 and is subsequently doubled at each layer of increased depth. The dropout rate (d) was empirically determined in the range 0.0–0.5 as the one that maximized model performance in the validation set. Conversely, the kernel size (N_sx3) was increased to N_sx11 (You et al., 2021), and the max-pooling factor (1x2) was also modified, increasing it to 1x5 in the first two layers and to 1x4 in the last two layers of the encoder, with the aim of enhancing the extraction of long-term features related to sleep stages while reducing computational load.

After the encoder stage, the extracted feature maps undergo further processing at the bottleneck of the network through a single conv block prior to the decoder. The decoder, which also consists of four layers, is designed to generate high-resolution feature maps and includes a dropout layer, a 2D transposed convolution (Conv2DTranspose), and a convolutional block in each layer. To ensure retention of low-level features, the output of the Conv2DTranspose of each decoder layer is concatenated with the output of the corresponding conv block of the encoder. Finally, the last layer of the network is a 2D convolution operation employing 4 filters and a kernel size of N_sx1 and a softmax activation, which generates point-wise predictions as output (4xL samples), representing the probability that each sample is associated with W, light, deep, and REM sleep stages.

2.4. POxi-SleepNet experimental setup: training and evaluation strategies

To see the performance and generalization ability of the proposed approach across databases encompassing different age subgroups, POxi-SleepNet models were trained and assessed using three different strategies: (i) a standard hold-out strategy (SHO); (ii) a leave-one-cohort-out (LOCO) without a fine-tuning approach; and (iii) a LOCO with fine-tuning (LOCO-FT) approach. These strategies are presented in Fig. 3 and are described next:

- SHO is intended to measure overall sleep staging performance of the POxi-SleepNet models. Each database is first partitioned into training/validation/test subsets (Fig. 3A), with comprehensive details of the partitioning schema presented in sections 1.1-1.6 of the supplementary material. All the training and validation subsets are then merged to compose the overall training and validation set, respectively (Fig. 3B). Subsequently, a single POxi-SleepNet model is trained and optimized using the overall training and validation sets, respectively. Finally, this model is assessed using each of the test subsets separately (Fig. 3B).
- LOCO is intended to measure the generalization ability of the POxi-SleepNet$_{PR\text{-}SpO2}$ model for prospective clinical applications. Accordingly, the training and validation subsets of all but one database are merged to compose the overall training and validation set, respectively (Fig. 3C). Subsequently, a POxi-SleepNet$_{PR\text{-}SpO2}$ model is trained and optimized using the overall training and validation sets, respectively, and the test subset of the omitted database is used for assessing model performance (Fig. 3C). This process is repeated six times, one per database, so each database is considered once as the omitted and testing database (Fig. 3C), thus measuring the generalization ability for each database.
- LOCO-FT is intended to assess the generalization ability of the POxi-SleepNet$_{PR\text{-}SpO2}$ model to new population samples. As in LOCO approach, POxi-SleepNet$_{PR\text{-}SpO2}$ model is trained and optimized using the training and validation subsets of all but one database (Fig. 3D). Subsequently, POxi-SleepNet$_{PR\text{-}SpO2}$ model is fine-tuned using the training and validation subjects from the withheld database and evaluated only on its corresponding test subset (Fig. 3D). This process is also repeated six times, with each database being considered once as the omitted and testing database (Fig. 3D).

For each of these strategies (SHO, LOCO, and LOCO-FT), the POxi-SleepNet architectures were trained using TensorFlow 2.9.0 library on an NVIDIA 3080Ti GPU. The training procedure included the following configuration: the categorical cross-entropy as the loss function to minimize, the Adam method with an initial learning rate of 10^{-4} to optimize weights and biases of the POxi-SleepNet architectures, batch sizes of 8 (POxi-SleepNet$_{PR}$ and POxi-SleepNet$_{SpO2}$) and 4 (POxi-SleepNet$_{PR\text{-}SpO2}$) with random data shuffling, a learning rate reduction by a factor of 2 after 15 epochs without improvement in the validation loss, and early stopping after 45 epochs of no improvement, being the model with the lowest validation loss chosen as the best model.

2.5. Explainable artificial intelligence: Seg-Grad-CAM

Seg-Grad-CAM is an extended version of the widely used Grad-CAM method that produces heatmaps showing the most important areas of the input data in CNN-based DL models for semantic segmentation applications (Vinogradova et al., 2020). Considering automatic scoring of sleep stages in sleep recordings as a semantic segmentation problem, Seg-Grad-CAM has been applied as an XAI method able to scrutinize and comprehend the decision-making processes of the POxi-SleepNet models in detecting each sleep stage. Given the feature maps of a chosen convolutional layer A^k ($k = 1, \ldots, Z$), an output class c, and a region of interest ROI^c, Seg-Grad-CAM first calculates the average of the gradients of the model output y_i^c ($i \in ROI^c$) with respect to all the feature maps A^k:

$$\alpha_k^c = \frac{1}{Z}\sum_{u,v}\frac{\partial \sum_{i\in ROI^c} y_i^c}{\partial A_{u,v}^k} \tag{1}$$

The class-discriminative localization map (heatmap) is then computed as a gradient-weighted combination of the feature maps, followed by a ReLU:

$$L^c = ReLU\left(\sum_k \alpha_k^c \cdot A^k\right) \tag{2}$$

This results in a heatmap of the exact dimensions as A^k, which is normalized and resized (N_sxL samples) to enable joint visualization with the input PR and SpO_2 signals (Vinogradova et al., 2020). Seg-Grad-CAM heatmaps were calculated in the last convolutional layer of each conv block in the bottleneck and decoder layers, as it enhances the identification of both low-level and high-level stage-related features (Vaquerizo-Villar et al., 2023). The final heatmap was then obtained by averaging all normalized and resized heatmaps, as it has been done in previous studies (Jiménez-García et al., 2024; Vaquerizo-Villar et al., 2023).

In this work, c denotes one of the four sleep stages (W/Light/Deep/REM) and ROI^c is the region of points scored as c by the POxi-SleepNet models, which allows to identify the most relevant areas of the PR (POxi-SleepNet$_{PR}$ model) and SpO_2 (and POxi-SleepNet$_{SpO2}$ model) data contributing to predict each sleep stage, as well as to analyze the complementarity between PR and SpO_2 signals (POxi-SleepNet$_{PR\text{-}SpO2}$ model) for sleep staging.

2.6. Quantitative analysis of the most important time and frequency stage-related PR and SpO_2 patterns

To provide a quantitative and consistent interpretation of the pulse oximetry features that drive the model to score each sleep stage, we conducted a thorough time and frequency analysis of the predominant regions within PR and SpO_2 signals highlighted by Seg-Grad-CAM heatmaps for each test patient. Since the heatmap values range from 0 (denoting minimal relevance in the prediction) to 1 (indicating high relevance in the prediction), we first selected, for each subject, the most important regions for predicting W/Light/Deep/REM stages as those where the heatmap amplitude exceeded 0.5, an empirically determined threshold. A comprehensive analysis of these regions was then performed to discern time and frequency PR and SpO_2 characteristics associated with each sleep stage:

- Time analysis. For each test patient, mean, standard deviation (SD), and root mean square of successive differences (RMSSD) were computed in the most important regions of standardized PR and SpO_2 signals derived from Seg-Grad-CAM heatmaps to predict W/Light/Deep/REM stages. These three common time-domain metrics measure temporal changes of R-R intervals and provide important insights into overall variability and short-term cardiac fluctuations (Shaffer and Ginsberg, 2017). They have been previously used to evaluate differences in heart rate variability (HRV) among sleep stages (Martín-Montero et al., 2023), and adapted here for PR and SpO_2.
- Frequency analysis. Previous studies have shown that spectral activity of cardiorespiratory signals differ among sleep stages (Martín-Montero et al., 2023; Penzel et al., 2003). For each test subject, we have first computed the continuous wavelet transform (CWT) of overnight PR and SpO_2 recordings, which offers optimal time-frequency resolution for the whole overnight recordings (Rioul and Vetterli, 1991). Specifically, CWT was calculated using the complex Morlet wavelet (Wachowiak et al., 2016) and frequencies in the range 0.001–0.400 Hz. The power spectral density (PSD) for each sleep stage was then derived as the average of the CWT in the corresponding relevant region derived from Seg-Grad-CAM heatmaps.

2.7. Statistical analysis

The POxi-SleepNet models provide probabilities for predicting each sleep stage (W/Light/Deep/REM) for each sample within the 12-h input data. These probabilities were translated into predictions of sleep stages by choosing the class with the highest probability. Given that manual sleep staging occurs every 30 s, the output label for each 30-s epoch was determined as the most prevalent predicted sleep stage within the epoch. Zero-padded areas were excluded prior to calculating performance metrics. The overall performance of the POxi-SleepNet architectures for automated sleep stage classification was evaluated using confusion matrices (4-class), which were utilized to obtain the 4-class accuracy (Acc), Cohen's kappa (kappa), macro-F1 score (MF1), and per-class F1-score (F1). In addition, the Wilcoxon signed-rank test was applied to evaluate two-by-two statistical differences in time parameters (mean, SD, and RMSSD) and in each frequency bin from the PSDs, derived from quantitative analysis among sleep stages (W/Light/Deep/REM). A p-value<0.01 was considered significant after Bonferroni correction (six comparisons).

3. Results

3.1. POxi-SleepNet performance: SHO strategy

A SHO strategy was used to assess overall sleep staging performance of the POxi-SleepNet models. Fig. 4 shows the confusion matrices of the POxi-SleepNet model trained with a using PR and SpO_2 data (POxi-SleepNet$_{PR\text{-}SpO2}$) in the six test subsets (CCSHS, CFS, CHAT, MESA, MrOS, and SHHS), whereas Table 2 shows the performance metrics of POxi-SleepNet$_{PR\text{-}SpO2}$ model in the six test sets, compared with those from POxi-SleepNet models trained using single-channel PR (POxi-SleepNet$_{PR}$) and single-channel SpO_2 data (POxi-SleepNet$_{SpO2}$). Notably, POxi-SleepNet$_{PR\text{-}SpO2}$ model showed a high performance in the six databases, with 4-class Acc values in the range 81.5 %–84.5 %, kappa values in the range 0.726–0.779, and MF1 values in the range 74.0 %–83.1 %. As expected, this model outperformed POxi-SleepNet$_{PR}$ (78.6 %–83.6 % Acc, 0.679–0.766 kappa, and 68.1 %–82.1 % MF1) and POxi-SleepNet$_{SpO2}$ (72.1 %–80.3 % Acc, 0.609–0.681 kappa, and 66.7 %–

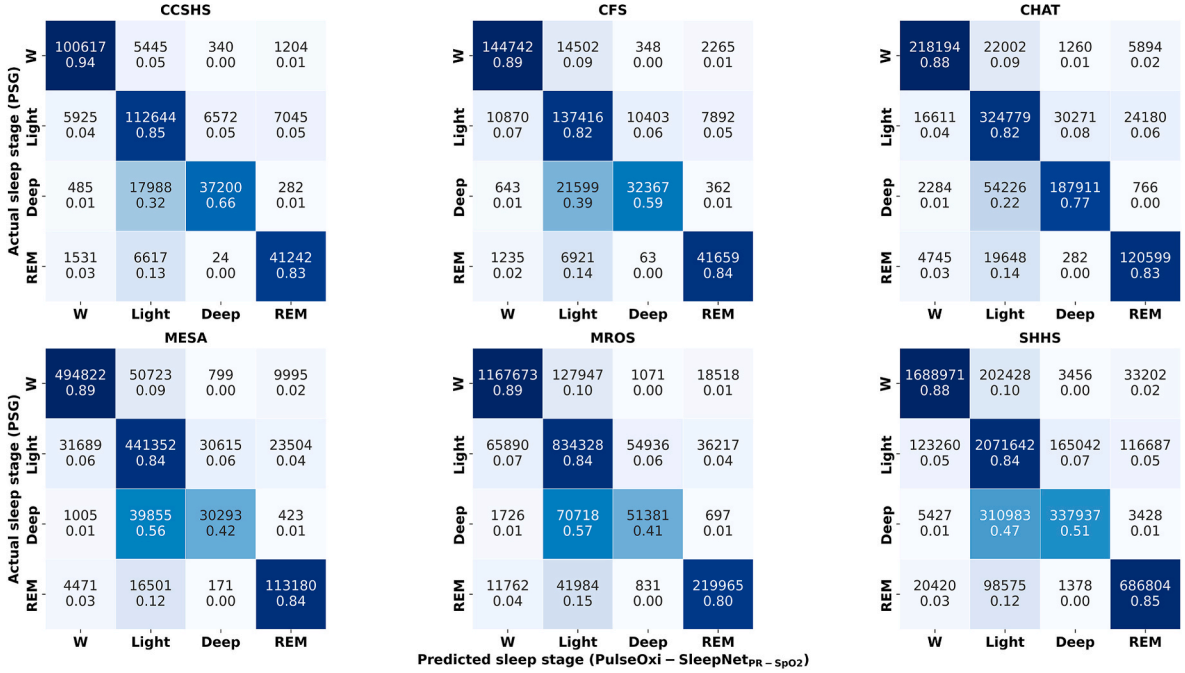

Fig. 4. Confusion matrix of the POxi-SleepNet$_{PR\text{-}SpO2}$ model in in the six test subsets (CCSHS, CFS, CHAT, MESA, MrOS, and SHHS). This matrix compares the sleep stages manually scored from PSG with the corresponding automatic assignation using the POxi-SleepNet$_{PR\text{-}SpO2}$ model.

Table 2
Diagnostic performance of POxi-SleepNet$_{PR\text{-}SpO2}$, POxi-SleepNet$_{PR}$, and POxi-SleepNet$_{SpO2}$ models in the six test subsets (CCSHS, CFS, CHAT, MESA, MrOS, and SHHS) to automatically classify sleep stages.

DB	Model	Overall Metrics			Per-class F1-score (%)			
		Acc (%)	kappa	MF1 (%)	W	Light	Deep	REM
CCSHS	**POxi-SleepNet$_{PR\text{-}SpO2}$**	**84.5**	**0.779**	**83.1**	**93.1**	**82.0**	**74.3**	**83.2**
	POxi-SleepNet$_{PR}$	83.6	0.766	82.1	92.3	81.1	74.9	80.3
	POxi-SleepNet$_{SpO2}$	74.5	0.636	72.3	86.0	72.4	65.1	65.7
CFS	**POxi-SleepNet$_{PR\text{-}SpO2}$**	**82.2**	**0.737**	**79.4**	**90.6**	**79.2**	**66.0**	**81.6**
	POxi-SleepNet$_{PR}$	79.9	0.702	77.0	88.2	77.1	65.7	77.0
	POxi-SleepNet$_{SpO2}$	76.5	0.653	72.9	86.7	73.4	61.6	70.0
CHAT	**POxi-SleepNet$_{PR\text{-}SpO2}$**	**82.4**	**0.754**	**82.7**	**89.2**	**79.6**	80.8	**81.3**
	POxi-SleepNet$_{PR}$	81.7	0.745	81.9	88.1	78.9	**81.0**	79.7
	POxi-SleepNet$_{SpO2}$	72.1	0.609	71.7	80.7	70.0	71.7	64.3
MESA	**POxi-SleepNet$_{PR\text{-}SpO2}$**	**83.7**	**0.743**	**74.7**	**90.9**	**82.1**	**45.4**	**80.4**
	POxi-SleepNet$_{PR}$	80.7	0.694	71.1	88.3	79.5	42.8	73.9
	POxi-SleepNet$_{SpO2}$	77.0	0.638	66.7	85.5	76.0	35.8	69.4
MrOS	**POxi-SleepNet$_{PR\text{-}SpO2}$**	**84.0**	**0.742**	**74.0**	**91.1**	**80.8**	**44.2**	**80.0**
	POxi-SleepNet$_{PR}$	80.4	0.679	68.1	88.4	77.6	37.0	69.6
	POxi-SleepNet$_{SpO2}$	80.3	0.681	68.8	88.0	77.3	35.9	73.9
SHHS	**POxi-SleepNet$_{PR\text{-}SpO2}$**	**81.5**	**0.726**	**77.8**	**89.7**	**80.3**	**58.0**	**83.4**
	POxi-SleepNet$_{PR}$	78.6	0.679	74.1	86.8	78.2	53.4	77.7
	POxi-SleepNet$_{SpO2}$	75.9	0.640	71.1	84.8	75.4	49.2	74.8

Acc: accuracy, CCSHS: Cleveland Children's Sleep and Health Study, CFS: Cleveland Family Study; CHAT: Childhood Adenotonsillectomy Trial; kappa: Cohen's kappa index; MESA: Multi-Ethnic Study of Atherosclerosis; MF1: macro F1-score, MrOS: Osteoporotic Fractures in Men Study; REM: rapid eye movement; SHHS: Sleep Heart Health Study; W: wakefulness.

72.9 % MF1) models in the six test databases.

In section 2 of the supplementary material, we show the high performance of the POxi-SleepNet$_{PR\text{-}SpO2}$ model for estimating the TST in the six test subsets (intra-class correlation coefficient values: 0.703–0.950; 95 % confidence intervals: 85.4 minutes–227.0 minutes). For a thorough analysis, section 3 of the supplementary material provides a comparison of the performance metrics of different POxi-SleepNet$_{PR\text{-}SpO2}$ models in the validation set according to their network structure, number of filters, and filter size (Table S1). Similarly, section 4 of the supplementary material provides the performance of the POxi-SleepNet$_{PR\text{-}SpO2}$ model for 5-class (W, N1, N2, Deep, and REM sleep) sleep staging, showing a low performance for N1 detection (Table S2 and Fig. S2) that is in line with previous studies.

In section 5 of the supplementary material, we show and discuss the sleep staging performance of the POxi-SleepNet models in the whole test cohort (Table S3) and across the six test subsets by input signal, age, sex, population, and OSA severity subgroups (Tables S4–S9). Interestingly, the POxi-SleepNet$_{PR}$ and POxi-SleepNet$_{PR\text{-}SpO2}$ models showed significantly lower performance with increasing age and OSA severity, while the POxi-SleepNet$_{SpO2}$ models performed significantly better. Conversely, sex-related differences in POxi-SleepNet models were observed in elderly patients (Table S10), with significantly higher Acc in elderly males and significantly higher MF1 in elderly females.

In section 6 of the supplementary material, we show and discuss the sleep staging performance of the POxi-SleepNet models by comorbidity and drug intake subgroups (Tables S11–S21). POxi-SleepNet models showed a significantly lower performance in patients with atrial fibrillation, congestive heart failure, diabetes, and hypertension comorbidities, as well as in patients taking diuretics, betablockers, benzodiazepines, antidepressants, and antidiabetics drugs.

Table 3
Diagnostic performance of POxi-SleepNet$_{PR\text{-}SpO2}$ models trained using SHO, LOCO, and LOCO-FT in the six test subsets (CCSHS, CFS, CHAT, MESA, MrOS, and SHHS) to automatically classify sleep stages.

DB	Model	Overall Metrics			Per-class F1-score (%)			
		Acc (%)	kappa	MF1 (%)	W	Light	Deep	REM
CCSHS	**SHO**	84.5	0.779	83.1	93.1	82.0	74.3	83.2
	LOCO	83.6	0.768	82.4	92.7	80.8	74.9	81.1
	LOCO-FT	84.9	0.785	83.8	93.2	82.3	76.6	82.9
CFS	**SHO**	82.2	0.737	79.4	90.6	79.2	66.0	81.6
	LOCO	82.2	0.737	79.8	90.2	79.3	68.8	80.7
	LOCO-FT	82.8	0.746	80.4	90.8	79.8	69.3	81.6
CHAT	**SHO**	82.4	0.754	82.7	89.2	79.6	80.8	81.3
	LOCO	76.3	0.668	76.4	84.9	74.0	71.3	75.5
	LOCO-FT	82.8	0.760	83.2	90.1	79.8	81.1	81.6
MESA	**SHO**	83.7	0.743	74.7	90.9	82.1	45.4	80.4
	LOCO	81.1	0.707	74.0	89.2	79.1	48.0	79.5
	LOCO-FT	84.1	0.746	73.1	90.8	83.0	37.5	80.9
MrOS	**SHO**	84.0	0.742	74.0	91.1	80.8	44.2	80.0
	LOCO	82.9	0.722	72.6	90.0	79.9	42.5	78.1
	LOCO-FT	84.2	0.742	73.8	91.1	81.1	42.8	80.1
SHHS	**SHO**	81.5	0.726	77.8	89.7	80.3	58.0	83.4
	LOCO	80.0	0.699	74.3	88.8	79.3	47.4	81.5
	LOCO-FT	80.6	0.712	76.7	88.9	79.7	56.2	82.0

Acc: accuracy, CCSHS: Cleveland Children's Sleep and Health Study, CFS: Cleveland Family Study; CHAT: Childhood Adenotonsillectomy Trial; kappa: Cohen's kappa index; LOCO: Leave-one-cohort-out without fine-tuning; LOCO-FT: LOCO with fine-tuning; MESA: Multi-Ethnic Study of Atherosclerosis; MF1: macro F1-score, MrOS: Osteoporotic Fractures in Men Study; REM: rapid eye movement; SHHS: Sleep Heart Health Study; SHO: Standard Hold out; W: wakefulness.

3.2. POxi-SleepNet generalization ability

The performance of POxi-SleepNet$_{PR\text{-}SpO2}$ was assessed using SHO, LOCO, and LOCO-FT strategies to evaluate its generalization ability across databases with different age subgroups. Table 3 presents a comparison of the performance of POxi-SleepNet$_{PR\text{-}SpO2}$ models across the six test subsets (CCSHS, CFS, CHAT, MESA, MrOS, and SHHS) when trained using three SHO, LOCO, and LOCO-FT strategies. Interestingly, the POxi-SleepNet$_{PR\text{-}SpO2}$ models showed a high generalization ability, with only minor differences in performance metrics (less than 3 % in Acc, 0.04 in kappa, and 4 % in MF1) across SHO, LOCO, and LOCO-FT approaches, except for the CHAT database.

3.3. Seg-Grad-CAM heatmaps interpretation of the POxi-SleepNet models

Fig. 5 shows Seg-Grad-CAM heatmaps obtained for samples of a representative subject predicted as W and Light sleep stages by the POxi-SleepNet$_{PR}$ (Fig. 5A–B) and POxi-SleepNet$_{SpO2}$ (Fig. 5C–D) models. Each

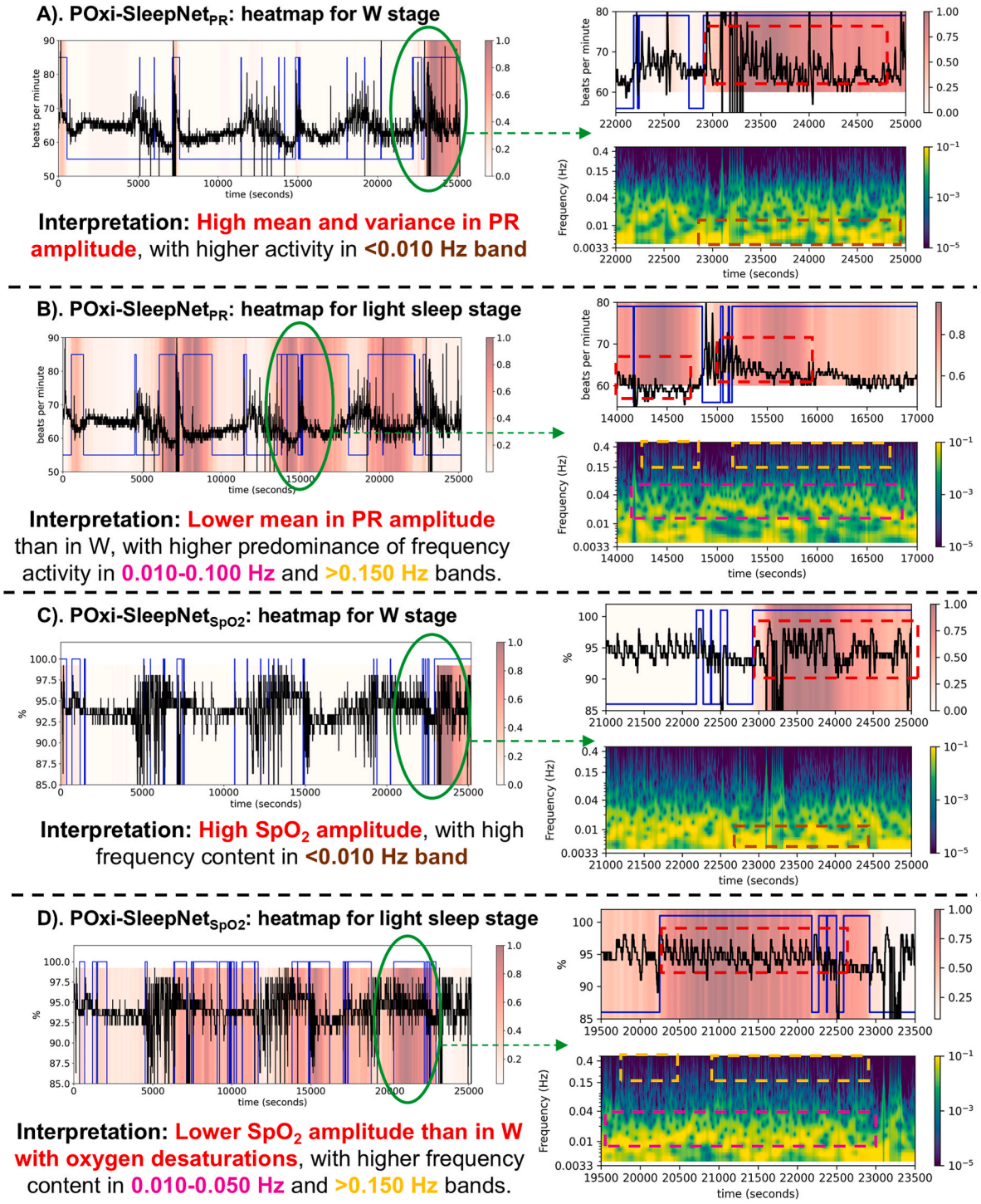

Fig. 5. Seg-Grad-CAM heatmaps obtained for samples of a representative subject (subject identifier: shhs1-205290) predicted as: (A) Wake by the POxi-SleepNet$_{PR}$ model, (B) Light sleep by the POxi-SleepNet$_{PR}$ model, (C) Wake by the POxi-SleepNet$_{SpO2}$ model, (D) Light sleep by the POxi-SleepNet$_{SpO2}$ model. Blue lines delineate the regions of interest containing samples predicted as the corresponding sleep stage.

heatmap presents a detailed zoom of a relevant area of the recording on the right, along with the scalogram derived from the CWT, which more effectively highlights the time-frequency patterns of the PR and SpO_2 recordings that the POxi-SleepNet models are focusing on to make the prediction. Darker areas in the heatmaps indicate greater relevance in the final decision taken by the POxi-SleepNet model. It is important to note that these heatmaps emphasize regions with different time and frequency PR and SpO_2 features to predict W and light sleep stages. Illustrative examples of Seg-Grad-CAM heatmaps obtained for samples classified as each sleep stage (W/Light/Deep/REM) by the POxi- POxi-SleepNet$_{PR}$ and POxi-SleepNet$_{SpO2}$ models, together with a more comprehensive visual interpretation of the time and frequency PR and SpO_2 features associated with each sleep stage, can be seen in section 5 of the supplementary material (see Figs. S3–S6).

In section 7 of the supplementary material, we also analyze the complementarity of PR and SpO_2 signals from some Seg-Grad-CAM heatmaps of the PulseOxi-SleepNet$_{PR\text{-}SpO2}$ model. This analysis shows that the POxi-SleepNet$_{PR\text{-}SpO2}$ model focuses both on PR and SpO_2 signals for sleep staging. Specifically, PR signal has a stronger influence (i.e., higher heatmap amplitude) than SpO_2 for sleep staging, with a main focus on areas of the PR signal near the regions predicted as each sleep stage (see Figs. S7–S8). Conversely, the heatmap pattern of the SpO_2 signal has a more dispersed distribution, with notable amplitudes spread across broader time regions in the sleep recording (see Figs. S7–S8).

3.4. Quantitative identification of time and frequency pulse oximetry patterns highlighted by Seg-Grad-CAM

To provide a quantitative interpretation of these findings, we subsequently performed a time and frequency analysis of the most important stage-related regions in PR and SpO_2 signals highlighted by Seg-Grad-CAM analysis of POxi-SleepNet$_{PR}$ and POxi-SleepNet$_{SpO2}$ models, respectively. The most important stage-related regions were selected as those where the heatmap amplitude exceeded 0.5.

Figs. 6A and 7A show the averaged PSDs in the 0–0.400 Hz range computed from the CWT of the most important regions of PR and SpO_2 signals, respectively, to predict each sleep stage in test subjects. Fig. S9 displays the *p*-values for each frequency of the PSDs of PR (Fig. S9A) and SpO_2 (Fig. S9B) signals. Importantly, statistically significant differences (*p*-value <0.01) are found in every frequency bin for at least five of the six sleep stage comparisons for the PR signal and in at least four of the six comparisons for the SpO_2 signal. Looking at Figs. 6A and 7A, three distinct frequency bands can be identified in PR and SpO_2 signals based on the predominance of their spectral content in each sleep stage: (*i*) 0.004–0.020 Hz (BW$_{PR1}$ and BW$_{SpO2\text{-}1}$); (*ii*) 0.020–0.100 Hz for PR and 0.020–0.050 Hz for SpO_2 (BW$_{PR2}$ and BW$_{SpO2\text{-}2}$); and (*iii*) 0.180–0.400

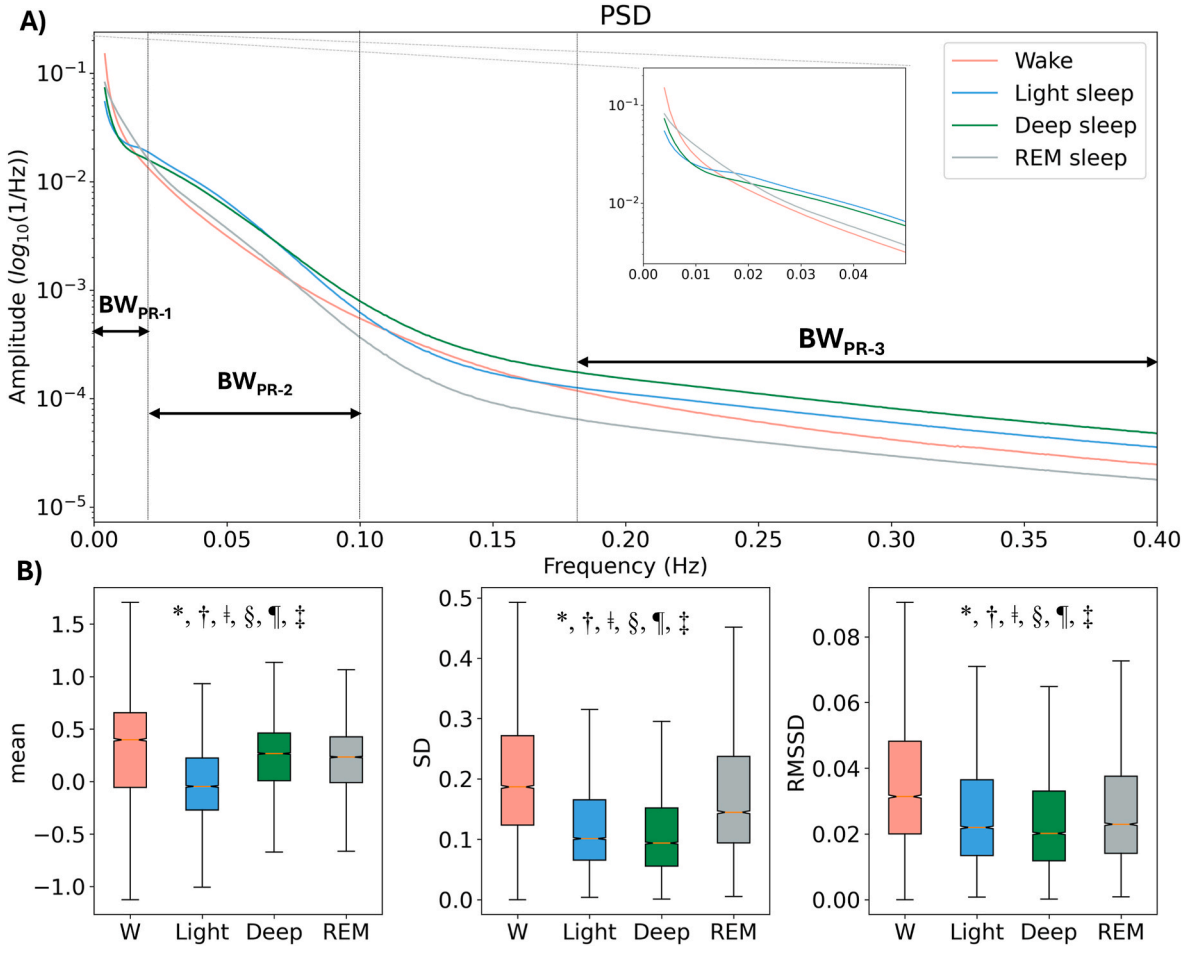

Fig. 6. Quantitative analysis for the most important regions of PR signal highlighted by Seg-Grad-CAM to predict Wake, Light sleep, Deep sleep, and REM sleep stages in test subjects. A) Averaged PSDs in the 0–0.400 Hz range. B) Boxplot distribution of the temporal features (mean, SD, and RMSSD). BW$_{PR\text{-}1}$: 0.004–0.020 Hz; BW$_{PR\text{-}2}$: 0.020–0.100 Hz; BW$_{PR\text{-}3}$: 0.180–0.400 Hz. Time-frequency parameters (PSD, as well as mean, SD, and RMSSD) were calculated from the standardized PR signal. *Statistically significant differences ($p < 0.01$, Bonferroni correction) between W and Light sleep; †Statistically significant differences ($p < 0.01$, Bonferroni correction) between W and Deep sleep; ǂStatistically significant differences ($p < 0.01$, Bonferroni correction) between W and REM sleep; §Statistically significant differences ($p < 0.01$, Bonferroni correction) between Light and Deep sleep; ¶Statistically significant differences ($p < 0.01$, Bonferroni correction) between Light and REM sleep; ‡Statistically significant differences ($p < 0.01$, Bonferroni correction) between Deep and REM sleep.

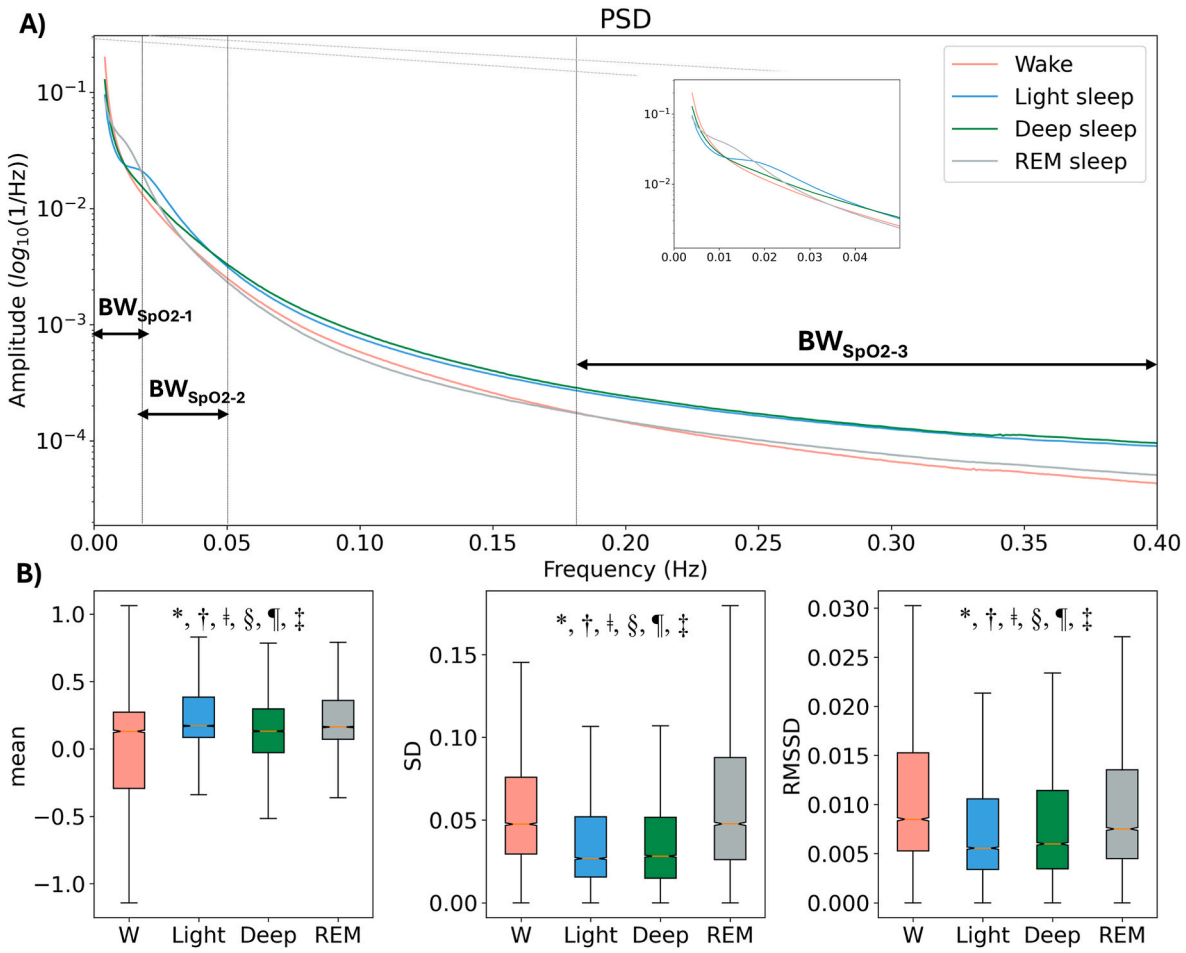

Fig. 7. Quantitative analysis for the most important regions of SpO_2 signal highlighted by Seg-Grad-CAM to predict Wake, Light sleep, Deep sleep, and REM sleep stages in test subjects. A) Averaged PSDs in the 0–0.400 Hz range. B) Boxplot distribution of the temporal features (mean, SD, and RMSSD). BW_{SpO2-1}: 0.004–0.020 Hz; BW_{SpO2-2}: 0.020–0.050 Hz; BW_{SpO2-3}: 0.180–0.400 Hz. Time-frequency parameters (PSD, as well as mean, SD, and RMSSD) were calculated from the standardized SpO_2 signal. *Statistically significant differences ($p < 0.01$, Bonferroni correction) between W and Light sleep; †Statistically significant differences ($p < 0.01$, Bonferroni correction) between W and Deep sleep; ‡Statistically significant differences ($p < 0.01$, Bonferroni correction) between W and REM sleep; §Statistically significant differences ($p < 0.01$, Bonferroni correction) between Light and Deep sleep; ¶Statistically significant differences ($p < 0.01$, Bonferroni correction) between Light and REM sleep; ǂStatistically significant differences ($p < 0.01$, Bonferroni correction) between Deep and REM sleep.

Hz (BW_{PR3} and BW_{SpO2-3}).

Figs. 6B and 7B display the boxplot distribution of the temporal features (mean, SD, and RMSSD), computed from the most important stage-related regions of PR and SpO_2 signals. Interestingly, boxplot distributions reflect statistically significant differences (p-value <0.01) in mean, SD, and RMSSD among sleep stages. In section 8 of the supplementary material, we provide the averaged PSDs and boxplot distribution of the time-domain features obtained for each test database, as well as by age, gender, and OSA severity subgroups (see Figs. S10–S43), together with a more comprehensive interpretation of the quantitative identification of time and frequency PR and SpO_2 features related to each sleep stage.

4. Discussion

In this study, we generated accurate and highly generalizable DL models based on U-Net (POxi-SleepNet) for sleep staging in OSA patients across all age subgroups (children, adolescents, adults, and elderly patients) while exclusively relying on pulse oximetry signals. We also offered a detailed interpretation of the stage-related pulse oximetry patterns identified by the POxi-SleepNet models through an XAI methodology based on Seg-Grad-CAM. This approach enabled us to delineate key regions within overnight PR and SpO_2 signals from each patient contributing to the prediction of each sleep stage, and to quantitatively identify the time and frequency characteristics of the pulse oximetry signals associated with W, light sleep, deep sleep, and REM sleep stages. This study is, to our knowledge, the first to combine DL and XAI techniques for the automated scoring of sleep stages in OSA patients across all age subgroups from pulse oximetry signals.

4.1. Sleep staging overall performance

POxi-SleepNet models achieved high performances to automatically score sleep stages in the six databases. Particularly, the POxi-SleepNet model using PR and SpO_2 data ($POxi\text{-}SleepNet_{PR\text{-}SpO2}$) showed the highest performance, with 81.5 %–84.5 % Acc, 0.726–0.779 kappa, and 74.0 %–83.1 % MF1 in the CCSHS, CFS, CHAT, MESA, MrOS, and SHHS databases. Of note, the highest F1-scores were reached for wake stage, with the deep sleep stage showing the lowest scores. This finding aligns with the state-of-the-art studies showing a considerable overlap between light sleep and deep sleep stages (Kotzen et al., 2023), probably due to the similarity of cardiovascular and respiratory patterns during non-REM sleep (Qin et al., 2021).

Manual sleep scoring is a highly subjective task. According to Lee et al. (2022), inter-rater agreement for 5-class sleep scoring is reported as a kappa of 0.76 (95 % confidence interval, 0.71–0.81). Thus, the

performance of POxi-SleepNet$_{PR\text{-}SpO2}$ for 4-class (0.726–0.779 kappa, Table 2) and 5-class sleep staging (0.675–0.762 kappa, Table S2) is not far from PSG-based sleep scoring, while pulse oximetry-based scoring has the advantage that it can be easily integrated with non-invasive wearable devices to monitor sleep stages in individuals at their home. AI-derived hypnodensity graphs, which depict the probability distribution across all sleep stages for each 30-s epoch throughout the night (Stephansen et al., 2018), have proven valuable for quantifying both sleep stage ambiguity and stability (Anderer et al., 2023a). They have proven effective not only in PSG but also in home sleep apnea testing (HSAT), where sleep staging relies primarily on cardiorespiratory signals, suggesting that ambiguities between sleep stages are reflected in both the central and autonomic nervous system activity (Anderer et al., 2023b). The usefulness of hypnodensity graphs is further reinforced when applied to multi-scored datasets, where inter-scorer variability can be explicitly modeled (Anderer et al., 2023b; Fiorillo et al., 2023b). Specifically, Fiorillo et al. (2023b) demonstrated that integrating multiple expert annotations during DL model training through label smoothing and soft-consensus distributions allows models to better adapt to the consensus of the group of scorers. This approach not only improves the performance of DL models for sleep staging but also enhances the similarity between the model-generated hypnodensity graphs and those derived from scorer consensus. Hence, the performance and clinical applicability of our proposed POxi-SleepNet$_{PR\text{-}SpO2}$ model could be enhanced by using hypnodensity graphs and multi-scored databases.

POxi-SleepNet$_{PR\text{-}SpO2}$ outperformed POxi-SleepNet$_{PR}$ and POxi-SleepNet$_{SpO2}$ models in the six test databases, which suggests that SpO_2 and PR have complementary information in sleep stage detection, as it has been suggested in previous studies (Casal et al., 2021, 2022; Vaquerizo-Villar et al., 2024). This complementarity becomes more apparent as age and OSA severity increases (see Table S3), which can be explained by impact of apneic events and sleep fragmentation induced by OSA and aging, respectively, on PR (induced sympathetic excitation) and SpO_2 (recurrent oxygen desaturations) signals (Choi et al., 2016; Edwards et al., 2010; Martín-Montero et al., 2023). In this context, Korkalainen et al. (2021) reported that the current guidelines for sleep staging may not be appropriate to analyze sleep recordings of patients with OSA and sleep fragmentation. A comprehensive discussion of sleep staging performance by input signal, age, sex, and OSA severity subsets is shown in section 5 of the Supplementary Material.

4.2. Sleep staging performance generalizability across all age subgroups

As aforementioned, the POxi-SleepNet$_{PR\text{-}SpO2}$ models showed a high generalization ability. It is important to note that fine-tuning the models to a specific cohort (i.e., LOCO-FT approach) leads to a minimal improvement in model performance, with less than 1 % in Acc and 0.02 in kappa when compared to the SHO approach. Interestingly, LOCO-FT approach resulted in a lower performance than SHO in the SHHS database, suggesting that cohort-specific sleep staging patterns are well captured by training the DL models with an extensive range of sleep databases. This aligns with Fiorillo et al. (2023b), who demonstrated that incorporating consensus information from multi-scored sleep databases improves sleep staging performance by reducing inter-rater variability, and with Fiorillo et al. (2023a), who reported that training on heterogeneous data from multiple centers consistently leads to better model performance than using a single cohort. Specifically, Fiorillo et al. (2023a) evaluated the generalization capability of U-Sleep, a U-Net-like DL network fed with one EEG and one EOG channel from 28528 PSG studies from 13 different databases encompassing all age subgroups and a wide range of sleep disorders. Their proposed DL network was able to deal with variability in EEG and EOG channel derivations, age, and sleep disorders. This supports the generalizability of DL models by highlighting their stability despite heterogeneity in clinical data.

The generalizability of our proposal is also supported by the minimal reduction in model's performance when using the LOCO approach, with less than 3 % Acc and 0.04 kappa in all but CHAT database. In the CHAT database, these differences in performance can be due to the pediatric patients it is composed of (5–10 years), who present substantial developmental differences in cardiorespiratory and neurophysiological activity, as well as in sleep architecture when compared to adults (Berry et al., 2018; Goh et al., 2000). Furthermore, the well-known cardiac cyclical variation in adults with OSA (Guilleminault et al., 1984) presents a high variability degree in the pediatric population determined by the type and extent of apneic events (Martín-Montero et al., 2023). However, these differences seem to disappear during adolescence, given the slight reduction in performance in CCSHS (16–19 years) using the LOCO approach. Our findings align with Fiorillo et al. (2023a) , who also observed a decrease in sleep staging in pediatric subjects when testing a DL model trained only with adults. This highlights the need to develop sleep-scoring models including more patients from different age subgroups to maximize its universal implementation.

4.3. Seg-Grad-CAM interpretations of the automatic sleep staging models

Some recent works have proposed XAI methodologies to visually discern those EEG (Dutt et al., 2022; Kuo et al., 2021; Phan et al., 2022; Vaquerizo-Villar et al., 2023) and pulse oximetry (Vaquerizo-Villar et al., 2024) patterns considered by DL models to score sleep stages in adult (Dutt et al., 2022; Kuo et al., 2021; Phan et al., 2022) and pediatric subjects (Vaquerizo-Villar et al., 2023, 2024). In this study, we introduce a novel implementation of an XAI method based on Seg-Grad-CAM to discern the decision-making process of a DL model and to interpret stage-related pulse oximetry patterns in OSA patients across all age subgroups. Analyzing the XAI results, it becomes apparent that the Seg-Grad-CAM-based approach can detect key regions with distinct time and frequency characteristics within overnight PR and SpO_2 signals from each patient used by the POxi-SleepNet models to predict wakefulness, light sleep, deep sleep, and REM sleep stages (Figures S2-S5). The proposed explainability approach also allows us to see the complementarity of PR and SpO_2 data for sleep staging (Figures S6-S7), suggesting that sleep staging is performed by first looking at time-frequency PR patterns close to the sample being predicted and then looking at long-term dynamics of SpO_2 (e.g., changes in baseline SpO_2 amplitude and presence and depth of oxygen desaturations) when there exist doubts in the prediction. In light of the reported findings, this automatic and interpretable sleep scoring tool could contribute to: (*i*) the visualization and interpretation of the sleep staging process by sleep technicians, meeting the recommendations of the EU for AI-based systems (Hamon et al., 2020); (*ii*) health recommendation systems integrated into remote servers or portable devices (Chinnasamy et al., 2023b), providing the automatic sleep stage predictions and Seg-Grad-CAM heatmaps per subject in a few seconds.

4.4. Identification of time and frequency pulse oximetry features for sleep scoring

In contrast to previous studies that show qualitative insights based on some hand-picked examples (Dutt et al., 2022; Kuo et al., 2021; Phan et al., 2022; Vaquerizo-Villar et al., 2023), we provide, for the first time, a quantitative interpretation of the physiological features that drive the POxi-SleepNet$_{PR}$ and POxi-SleepNet$_{SpO2}$ models to score each sleep stage. Table 4 summarizes the main time and frequency characteristics of PR and SpO_2 recordings related to each sleep stage. In the frequency domain, we found three distinct frequency bands within 0.004–0.020 Hz (BW_{PR1} and $BW_{SpO2\text{-}1}$), 0.020–0.100 Hz (BW_{PR2} and $BW_{SpO2\text{-}2}$), and 0.180–0.400 Hz (BW_{PR3} and $BW_{SpO2\text{-}3}$). Regarding $BW_{PR\text{-}1}$ and $BW_{SpO2\text{-}1}$ bands, we found that: (*i*) W stage is characterized by a high activity below 0.010 Hz, which is coherent with the macro-sleep disruption band reported by Martín-Montero et al. (2023, 2021) (0.001–0.005 Hz); (*ii*) REM sleep has the highest activity in the 0.010–0.020 Hz frequency

Table 4
Time and frequency characteristics of PR and SpO_2 signals related to each sleep stage.

Signal	Sleep stage	Features for sleep scoring	
		Time domain	Frequency domain
PR	W	↑ SD compared to non-REM sleep ↑ RMSSD than in sleep stages (light, deep, and REM sleep)	↑ activity in BW_{PR-1} than non-REM sleep, with the highest power in 0.004–0.010 Hz ↓ BW_{PR-2} and BW_{PR-3} activity than non-REM sleep
	Light sleep	↓ Mean than in deep and REM sleep ↓ SD than in W and REM sleep ↓ RMSSD than in W	↑ activity in BW_{PR-2} and BW_{PR-3} and ↓ activity in BW_{PR-1} compared to W and REM stages ↑ activity in 0.020–0.050 Hz compared to deep sleep as OSA severity increases
	Deep sleep	↑ Mean than in light sleep ↓ SD than in W and REM sleep. ↓ RMSSD than in W	↑ activity in BW_{PR-2} and BW_{PR-3} and ↓ activity in BW_{PR-1} compared to W and REM sleep. ↑ BW_{PR-3} activity compared to light sleep
	REM sleep	↑ Mean than in light sleep ↑ SD compared to non-REM sleep. ↓ RMSSD than in W	↑ activity in BW_{PR-1} and ↓ BW_{PR-2} and BW_{PR-3} activity compared to non-REM sleep. Highest activity in 0.010–0.020 Hz, particularly with increasing OSA severity ↓ BW_{PR-3} activity than in W except in children
SpO_2	W	↓ Mean value than in light and REM sleep stages (probably due to artifacts). ↑ SD compared to non-REM sleep.	↑ activity in BW_{SpO2-1} than non-REM sleep, with the highest power in 0.004–0.010 Hz ↓ BW_{SpO2-2} and BW_{SpO2-3} activity than non-REM sleep
	Light sleep	↑ Mean than in deep sleep ↓ SD than in W except in severe OSA. ↓ RMSSD than in REM sleep	↑ BW_{SpO2-3} and ↓ BW_{SpO2-1} activity compared to W and REM stages ↑ activity in BW_{SpO2-2} compared to W ↑ activity in 0.020–0.050 Hz compared to deep and REM sleep as OSA severity increases
	Deep sleep	↓ Mean than in light and REM sleep ↓ SD than in light and REM sleep	↑ BW_{SpO2-3} and ↓ BW_{SpO2-1} activity compared to W and REM sleep ↓ BW_{SpO2-2} and ↑ BW_{SpO2-3} activity compared to light sleep with increasing OSA severity
	REM sleep	↑ Mean than in deep sleep. ↑ SD compared to non-REM sleep ↑ RMSSD than in light sleep	↑ activity in BW_{SpO2-1} than non-REM sleep Highest activity in 0.010–0.020 Hz, particularly with increasing OSA severity

↑: Significant increase in activity (e.g., BW_{SpO2-3} frequency band) or metric value (e.g., mean/SD/RMSSD) compared to the referenced sleep stage(s); ↓: Significant decrease in activity (e.g., BW_{SpO2-1}) or metric value (e.g., mean/SD/RMSSD) compared to the referenced sleep stage(s); OSA: obstructive sleep apnea; PR: pulse rate; REM: rapid eye movement; RMSSD: root mean square of successive differences; SD: standard deviation; SpO_2: blood oxygen saturation; W: wake.

range, particularly with increasing OSA severity (Figures S21-S24 and Figures S38-S41), which is related to apneic events of long duration that impose more severe health outcomes to OSA patients (Anderer et al., 2023b; Bonsignore et al., 2024; Varga and Mokhlesi, 2019). Regarding BW_{PR-2} and BW_{SpO2-2}, its higher spectral content during light sleep and deep sleep is associated with the intrinsic depression of sympathetic nervous system during these sleep stages that makes it easier to differentiate OSA sympathetic excitation during these sleep stages (Martín-Montero et al., 2023). This agrees with the spectral power increase in this band during non-REM sleep (mainly light sleep) with increasing OSA severity (Figures S21-S24 and Figures S38-S41), and is also consistent with previous studies that have found spectral bands within these ranges related to OSA and its severity in both pediatric and adult subjects (Álvarez et al., 2013; Gutiérrez-Tobal et al., 2015; Martín-Montero et al., 2021; Vaquerizo-Villar et al., 2018). Finally, BW_{PR3} and BW_{SpO2-3} are characterized by a higher content during non-REM sleep (light and deep sleep), which is coherent with the respiratory-modulated bands found in previous studies (Martín-Montero et al., 2021, 2023) and reflect the characteristic parasympathetic activation during these stages (Qin et al., 2021).

In the time domain, mean and variability (SD and/or RMSSD) in PR and SpO_2 amplitude also show differences among sleep stages (Figs. 6–7). First, we found that mean PR amplitude is significantly lower on light sleep when compared to deep and REM sleep, while SpO_2 baseline amplitude is lower in deep sleep than in light and REM sleep across all subgroups. To our knowledge, these pulse oximetry patterns had not been previously reported as important for sleep staging. Regarding variability, SD in PR and SpO_2 is significantly higher during W compared to non-REM sleep and RMSSD in PR is significantly higher during W compared to sleep across all subgroups (Figures S8-S41). Similarly, REM is characterized by a higher variability in PR and SpO_2 than during non-REM sleep for all subgroups (Figures S8-S41). In the presence of high sympathetic activity and low parasympathetic activity, SD in PR is increased (Shaffer and Ginsberg, 2017). Thus, the altered sympathetic activity and reduced parasympathetic activity during W and REM seems to be behind PR higher variability compared to non-REM, where this trend is reverted.

Taken together, our results show that the model is focusing on not only well-known stage-related patterns but also on OSA-specific alterations and novel PR and SpO_2 patterns that have not been previously reported as important for sleep staging. This paves the way for developing new guidelines for annotating sleep stages in widely used at-home polygraphy studies that do not record EEG.

4.5. Comparison with previous studies

There are various recent investigations applying DL techniques to pulse oximetry signals for automatic sleep staging (Faust et al., 2019), with some of them using OSA patient datasets (Casal et al., 2021, 2022; Haimov et al., 2025a; Huttunen et al., 2021, 2022; Korkalainen et al., 2020; Kotzen et al., 2023; Nam et al., 2024; Radha et al., 2021; Sridhar et al., 2020; Vaquerizo-Villar et al., 2024; Wulterkens et al., 2021). Some studies have just approached the differentiation between wake and sleep stages (W/S) (Casal et al., 2021, 2022), whereas the majority of them have targeted 3-class (W/non-REM sleep/REM sleep) or 4-class (W/light sleep/deep sleep/REM sleep) sleep staging (Haimov et al., 2025a; Huttunen et al., 2021, 2022; Korkalainen et al., 2020; Kotzen et al., 2023; Nam et al., 2024; Radha et al., 2021; Sridhar et al., 2020). Table 5 outlines a comparison between our proposed methodology and earlier research works centered on automatic 4-class sleep staging in OSA cohorts across diverse age groups using only pulse oximetry recordings (PPG, PR, and/or SpO_2) (Huttunen et al., 2021, 2022; Korkalainen et al., 2020; Kotzen et al., 2023; Nam et al., 2024; Radha et al., 2021; Sridhar et al., 2020; Vaquerizo-Villar et al., 2024). Interestingly, the current study achieved a similar performance than the reported by Nam et al. (2024) and Kotzen et al. (2023), but with a slightly higher generalization ability, suggesting that PR and SpO_2 recordings provide the same information as PPG for sleep staging while presenting less heterogeneity among different recording devices. In contrast to PPG-based approaches, our proposal is easier to be implemented and tested in portable monitoring equipment that store only PR and SpO_2 data. Furthermore, our proposal demonstrated higher performance and generalization ability

Table 5
Diagnostic performance of state-of-the-art approaches in automatic sleep staging in OSA cohorts across various age subgroups from pulse oximetry signals.

Study	Databases (Name/Age range (subgroups))	Sleep stage balance (W/Light/Deep/REM)	AHI (e/h)	N (Total/Test)	Signals	Methodology (DL: key components/Validation/XAI)	4-class Acc/kappa		
							SHO	LOCO	LOCO-FT
Sridhar et al. (2020)	MESA/54–90 years (adults)	42/41/5/10[a]	18 [10–33][a]	2033/194	HR[c]	CNN: convolutions and dilated convolutions/SHO, LOCO/-	80/0.69	–	–
	SHHS/40–90 years (adults)	30/43/11/14[a]	13 [7–24][a]	8299/800			77/0.66	–	–
	Physionet/20–84 years (adults)[b]	29/50/10/11[b]	19 (14)[k]	993/993			–	72/0.55	–
Radha et al. (2021)	Private/41–66 years (adults)	15/45/22/18*	[j]	101/101	PPG[d]	HRV features and RNN/SHO, LOCO, LOCO-FT/-	70/0.55	72/0.55	76/0.65
Korkalainen et al. (2020)	Private/44–66[l] years adults	33/42/13/12*	16 [7–33]	894/89	PPG	CNN and RNN/SHO, LOCO-FT/-	69/0.54	–	69/0.54**
Huttunen et al. (2021)	Private/44–66 [l] years adults [e]	33/42/13/12*	16 [7–33]	877/88	PPG	CNN and RNN/LOCO-FT/-	–	–	74/0.64
Huttunen et al. (2022)	Private/44–66 [l] years adults	33/42/13/12*	16 [7–33]	877/88	PPG, SpO_2	U-Net: convolutions and dilated convolutions/SHO, LOCO-FT/	75/0.63*	–	75/0.63**
Kotzen et al. (2023)	CFS/7–89 years (251 adults, 40 adolescents, and 33 children) [f,g,h]	34/39/12/17	5 [2–16][a]	320/320	PPG[h]	SleepPPGNet: residual convolutions, dense layers, and dilated convolutions/SHO, LOCO, LOCO-FT/-	–	76/0.67	82/0.74
	MESA/54–90 years (adults)[h]	37/43/5/11	18 [10–33][a]	2054/204			84/0.75	–	84/0.75**
Nam et al. (2024)	CFS/7–89 years (251 adults, 40 adolescents, and 33 children) [f,g]	34/39/12/17	5 [2–16][a]	320/320	PPG	InsightSleepNet: local attention, InceptionTime, a time-distributed dense layer, a temporal convolutional network, and CNN modules/SHO, LOCO-FT/Attention scores: visual analysis	–	–	81/0.72
	MESA/54–90 years (adults)	37/43/5/11	18 [10–33][a]	2054/204			84/0.74	–	84/0.74**
	CAP/14–82 years (101 adults and 7 adolescents)	14/39/25/15	[j]	24/24			–	–	81/0.73
Vaquerizo et al. (2024)	CHAT/5–10 years (children)	23/39/23/14	3 [1–6]	1633/858	PR, SpO_2	U-Net: convolutions/SHO/Seg-Grad-CAM: visual analysis	78/0.70	–	–
This study	**CCSHS/16–20 years (adolescents)**	29/39/15/15	1[0–2]	515/258	PR, SpO_2	U-Net (POxi-SleepNet): convolutions/SHO, LOCO, LOCO-FT/Seg-Grad-CAM: visual analysis and quantitative analysis	85/0.78	84/0.77	85/0.79
	CFS/7–89 years (587 adults, 96 adolescents, and 47 children)	36/39/12/12	5 [2–16]	730/365			**82/0.74**	**82/0.74**	**83/0.75**
	CHAT/5–10 years (children)	23/39/23/14	3 [1–6]	1633/858			**82/0.75**	**76/0.67**	**83/0.76**
	MESA/54–90 years (adults)	42/41/5/10	18 [10–33]	2056/1028			84/0.74	**81/0.71**	**84/0.75**
	MrOS/65–90 years (adults)	47/37/4/10	17 [9–30]	3915/2049			84/0.74	83/0.72	84/0.74
	SHHS/40–90 years (adults)	30/43/11/14	13 [7–24]	8444/5298			**82/0.73**	**80/0.70**	**81/0.71**

Acc: accuracy; AHI: apnea-hypopnea index; CCSHS: Cleveland Children's Sleep and Health Study; CFS: Cleveland Family Study; CHAT: Childhood Adenotonsillectomy Trial; CNN: Convolutional neural network; e/h: events per hour of sleep; HRV: heart rate variability; kappa: Cohen's kappa index; LOCO: Leave-one-cohort-out without fine-tuning; LOCO-FT: LOCO with fine-tuning; MESA: Multi-Ethnic Study of Atherosclerosis; MrOS: Osteoporotic Fractures in Men Study; N: number of sleep studies; PPG: photoplethysmography; PR: pulse rate; REM: rapid eye movement; RNN: recurrent neural network; Seg-Grad-CAM: semantic segmentation via gradient-weighted class activation mapping; SHHS: Sleep Heart Health Study; SHO: standard hold out; SpO_2: blood oxygen saturation; W: wake; XAI: explainable artificial intelligence. * Computed from reported data; ** DL model was trained and assessed using different subjects of the same cohort; [a] Computed from as the values for CFS, MESA, and/or SHHS in our study; [b] Computed from reported data in https://physionet.org/content/challenge-2018/1.0.0/. [c] HR= Heart rate derived from electrocardiogram (ECG) signal; [d] 584 ECG recordings from the Siesta database (20–95 years) were used for pretraining in LOCO and LOCO-FT approaches; [e] A private database with 2149 PPG recordings was used for pretraining the DL model; [f] CFS database contains 750 sleep studies, but authors only used 324 (251 adults/40 adolescents/33 children), presumably those containing valid PPG signal; [g] Computed from available demographic data from those subjects from CFS containing PPG; [h] 5767 ECG recordings from SHHS were used for pretraining the DL model; [i] Computed from reported data in in https://physionet.org/content/capslpdb/1.0.0/; [j] Not available; [k] Value expressed as mean (std). [l] Value expressed as interquartile range.

than other studies, particularly Sridhar et al. (2020) in the MESA and SHHS databases, and Vaquerizo-Villar et al. (2024) in the CHAT database. This underscores the suitability of the proposed DL model, POxi-SleepNet, which was designed and tested using PR and SpO_2 recordings from 17303 sleep studies of six different datasets (including 611 adolescents and 1680 children). Additionally, we introduced a novel XAI analysis methodology that provides qualitative and quantitative identification of the PR and SpO_2 characteristics considered by the POxi-SleepNet models for detecting each sleep stage using Seg-Grad-CAM, thus improving its clinical relevance.

4.6. Limitations and future work

It is important to mention several limitations of our research. First, despite a large sample size ($n = 17303$), most of the patients were adults ($n = 15623$ *vs.* $n = 1680$ children), with only one pediatric database available. This circumstance may have caused a decline in performance on the CHAT database when using the LOCO approach. This limitation is also present in existing studies aimed at sleep staging across various age groups (Fiorillo et al., 2023a; Perslev et al., 2021), which also relied solely on CHAT for pediatric sleep scoring due to the lack of alternative publicly available pediatric sleep datasets. Thus, the inclusion of incremental pediatric datasets would be advantageous for enhancing the

generalizability of our proposal, making it more applicable for simplifying the sleep staging process across a wider range of patients. Similarly, multi-scored sleep databases could help the performance of our proposal, as shown in recent EEG-based sleep staging approaches (Fiorillo et al., 2023b). Conversely, our approach has been restricted to OSA cohorts. Thus, further research is also required to validate our proposal for sleep staging in patient cohorts encompassing other sleep disorders, such as insomnia (Tripathi et al., 2022), narcolepsy (Stephansen et al., 2018), or REM sleep disorder (Levendowski et al., 2023), for which sleep staging has already been explored in these prior studies. This step is essential to ensure the applicability of our proposal across a broad range of clinical populations with suspected sleep disorders. Another limitation is that we did not measure the uncertainty in the model's decisions. In this respect, the quantification of model's uncertainty could help to identify most of the epochs wrongly classified and subsequently enhance sleep scoring, as shown in recent EEG (Phan et al., 2022) and PPG-based (Nam et al., 2024) sleep staging approaches. Furthermore, while we have shown that Seg-Grad-CAM heatmaps can help to provide a quantitative identification of time and frequency patterns of the PR and SpO_2 recordings influencing stage predictions, other XAI methods such as SHAP could be explored in future research. SHAP has been shown to aid in developing accurate and interpretable sleep staging models (Krauss et al., 2025; Wang et al., 2025), albeit with increased computational demands. Finally, another potential future aim could be to evaluate the proposed approach using ambulatory PR and SpO_2 recordings, together with a full-scale software application that shows sleep stage predictions within a hypodensity graph. Notably, ambulatory EEG recordings have been successfully tested for automatic sleep staging in OSA patients (Kalevo et al., 2022). Building on this, our approach could potentially offer real-time sleep stage scoring and Seg-Grad-CAM heatmap visualizations for each subject, delivered to the clinicians within seconds, thereby enhancing its applicability in clinical settings.

4.7. Conclusions

In summary, a novel U-Net-based model fed with PR and SpO_2 signals (POxi-SleepNet$_{PR\text{-}SpO2}$) showed remarkably high precision performance and generalization ability in the scoring of sleep stages in 17303 sleep recordings from OSA patients across all age groups. Utilizing both PR and SpO_2 signals proved to be complementary and increased the model performance than using each signal separately. In addition, a XAI analysis based on Seg-Grad-CAM enabled to recognize and quantify the time and frequency patterns of the overnight PR and SpO_2 signals that drive the DL model to predict W, light sleep, deep sleep, and REM sleep stages. Specifically, changes in the mean and variability in PR and SpO_2 amplitude, alongside changes in the spectral power of PR and SpO_2 within 0.004–0.020 Hz, 0.020–0.100 Hz, and 0.180–0.400 Hz bands, showed differences among sleep stages. Thus, we conclude that our approach combining DL and XAI analysis to process pulse oximetry signals may ease its integration in real healthcare environments for automated sleep staging in all individuals being evaluated for suspected OSA, irrespectively of their age.

CRediT authorship contribution statement

Fernando Vaquerizo-Villar: Writing – review & editing, Validation, Methodology, Funding acquisition, Data curation, Writing – original draft, Software, Investigation, Formal analysis, Conceptualization. **Gonzalo C. Gutiérrez-Tobal:** Writing – review & editing, Methodology, Funding acquisition, Writing – original draft, Investigation, Conceptualization. **Daniel Álvarez:** Writing – original draft, Funding acquisition, Writing – review & editing, Investigation, Conceptualization. **Adrián Martín-Montero:** Writing – original draft, Formal analysis, Writing – review & editing, Methodology. **David Gozal:** Writing – review & editing, Writing – original draft, Funding acquisition, Conceptualization, Investigation, Data curation. **Roberto Hornero:** Writing – original draft, Funding acquisition, Writing – review & editing, Investigation, Conceptualization.

Code availability

The code created and used for analyzing the polysomnography data for the current study is available from the corresponding author on reasonable request.

Ethical approval

This work has been carried out according to the Declaration of Helsinki.

All datasets were obtained upon request on the National Sleep Research Resource (NSRR) repository (Zhang et al., 2018). The ClinicalTrials.gov Identifier of the CHAT and SHHS databases are NCT00560859 and NCT000052, respectively. A written informed consent was obtained from all study participants or their legal caretakers as part of the research protocol of each database, which can be consulted in Rosen et al. (2003) (CCSHS), Redline et al. (1995) (CFS), Marcus et al. (2013) (CHAT), Bild et al. (2002) and Chen et al. (2015) (MESA), Blackwell et al. (2011) and Orwoll et al. (2005) (MrOS), and Quan et al. (1997) (SHHS).

Declaration of competing interest

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

Acknowledgements

This work is part of the projects PID2023-148895OB-I00, PID2020-115468RB-I00, and CPP2022-009735, funded by MCIN/AEI/10.13039/501100011033, the 'Fondo Social Europeo Plus (FSE+)', 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 "Consorcio Centro de Investigación Biomédica en Red (CIBER) en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN)" (CB19/01/00012) through "Instituto de Salud Carlos III (ISCIII)", co-funded with European Regional Development Fund.

F. Vaquerizo-Villar is supported by a "Sara Borrell" grant (CD23/00031) from the ISCIII cofounded by the FSE+. D. Álvarez is supported by a "Ramón y Cajal" grant (RYC2019-028566-I) funded by MCIN/AEI/10.13039/501100011033 and the European Social Fund Investing in your future. D. Gozal is supported by National Institute on Aging grant AG061824.

The six public datasets used in this work correspond to research studies that also received funding for data collection. The Cleveland Children's Sleep and Health Study (CCSHS) was supported by grants from the National Institutes of Health (RO1HL60957, K23 HL04426, RO1 NR02707, M01 Rrmpd0380-39). The Cleveland Family Study (CFS) was supported by grants from the National Institutes of Health (HL46380, M01 RR00080–39, T32-HL07567, RO1-46380). The Childhood Adenotonsillectomy Trial (CHAT) was supported by the National Institutes of Health (HL083075, HL083129, UL1-RR-024134, UL1 RR024989). The Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Ancillary study was funded by NIH-NHLBI Association of Sleep Disorders with Cardiovascular Health Across Ethnic Groups (RO1 HL098433). MESA is supported by NHLBI funded contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168 and N01-HC-95169 from the National Heart, Lung, and Blood Institute, and by cooperative agreements UL1-TR-

000040, UL1-TR-001079, and UL1-TR-001420 funded by NCATS. The National Heart, Lung, and Blood Institute provided funding for the ancillary MrOS Sleep Study, "Outcomes of Sleep Disorders in Older Men," under the following grant numbers: R01 HL071194, R01 HL070848, R01 HL070847, R01 HL070842, R01 HL070841, R01 HL070837, R01 HL070838, and R01 HL070839. The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002). The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.engappai.2025.112562.

Data availability

Polysomnography data from the databases employed in this study are available upon request through the National Sleep Research Resource website (https://sleepdata.org/datasets/).

References

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

Álvarez, D., Hornero, R., Marcos, V., Wessel, N., Penzel, T., Glos, M., del Campo, F., 2013. Assessment of feature selection and classification approaches to enhance information from overnight oximetry in the context of apnea diagnosis. Int. J. Neural Syst. 23, 1350020. https://doi.org/10.1142/S0129065713500202.

Anderer, P., Ross, M., Cerny, A., Vasko, R., Shaw, E., Fonseca, P., 2023a. Overview of the hypnodensity approach to scoring sleep for polysomnography and home sleep testing. Front. Sleep 2. https://doi.org/10.3389/frsle.2023.1163477.

Anderer, P., Ross, M., Cerny, A., Vasko, R., Shaw, E., Fonseca, P., 2023b. Overview of the hypnodensity approach to scoring sleep for polysomnography and home sleep testing. Front. Sleep 2. https://doi.org/10.3389/frsle.2023.1163477.

Baumert, M., Cowie, M.R., Redline, S., Mehra, R., Arzt, M., Pépin, J.L., Linz, D., 2022. Sleep characterization with smart wearable devices: a call for standardization and consensus recommendations. Sleep 45, 1–6. https://doi.org/10.1093/sleep/zsac183.

Benjafield, A.V., Eastwood, P.R., Heinzer, R., Morrell, M.J., Federal, U., Paulo, D.S., Paulo, S., Valentine, K., 2020. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir. Med. 7, 687–698. https://doi.org/10.1016/S2213-2600(19)30198-5.Estimation.

Berry, R.B., Brooks, R., Gamaldo, C.E., Harding, S.M., Marcus, C.L., Vaughn, B.V., 2018. The AASM manual for the scoring of sleep and associated events. Am. Acad. Sleep Med. 53, 1689–1699.

Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R., Diez Roux, A.V., Folsom, A.R., Greenland, P., JacobsJr, D.R., Kronmal, R., Liu, K., others, 2002. Multi-ethnic study of atherosclerosis: objectives and design. Am. J. Epidemiol. 156, 871–881. https://doi.org/10.1093/aje/kwf113.

Blackwell, T., Yaffe, K., Ancoli-Israel, S., Redline, S., Ensrud, K.E., Stefanick, M.L., Laffan, A., Stone, K.L., 2011. Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic fractures in men sleep study. J. Am. Geriatr. Soc. 59, 2217–2225. https://doi.org/10.1111/j.1532-5415.2011.03731.x.

Bonsignore, M.R., Mazzuca, E., Baiamonte, P., Bouckaert, B., Verbeke, W., Pevernagie, D. A., 2024. REM sleep obstructive sleep apnoea. Eur. Respir. Rev. 33. https://doi.org/10.1183/16000617.0166-2023.

Casal, R., Di Persia, L.E., Schlotthauer, G., 2022. Temporal convolutional networks and transformers for classifying the sleep stage in awake or asleep using pulse oximetry signals. J. Comput. Sci. 59. https://doi.org/10.1016/j.jocs.2021.101544.

Casal, R., Di Persia, L.E., Schlotthauer, G., 2021. Classifying sleep–wake stages through recurrent neural networks using pulse oximetry signals. Biomed. Signal Process Control 63, 102195. https://doi.org/10.1016/j.bspc.2020.102195.

Chan, E.D., Chan, Michael M., Chan, Mallory M., 2013. Pulse oximetry: understanding its basic principles facilitates appreciation of its limitations. Respir. Med. 107, 789–799. https://doi.org/10.1016/j.rmed.2013.02.004.

Chang, J.L., Goldberg, A.N., Alt, J.A., Mohammed, A., Ashbrook, L., Auckley, D., Ayappa, I., Bakhtiar, H., Barrera, J.E., Bartley, B.L., Billings, M.E., Boon, M.S., Bosschieter, P., Braverman, I., Brodie, K., Cabrera-Muffly, C., Caesar, R., Cahali, M. B., Cai, Y., Cao, M., Capasso, R., Caples, S.M., Chahine, L.M., Chang, C.P., Chang, K. W., Chaudhary, N., Cheong, C.S.J., Chowdhuri, S., Cistulli, P.A., Claman, D., Collen, J., Coughlin, K.C., Creamer, J., Davis, E.M., Dupuy-McCauley, K.L., Durr, M. L., Dutt, M., Ali, M. El, Elkassabany, N.M., Epstein, L.J., Fiala, J.A., Freedman, N., Gill, K., Gillespie, M.B., Golisch, L., Gooneratne, N., Gottlieb, D.J., Green, K.K., Gulati, A., Gurubhagavatula, I., Hayward, N., Hoff, P.T., Hoffmann, O.M.G., Holfinger, S.J., Hsia, J., Huntley, C., Huoh, K.C., Huyett, P., Inala, S., Ishman, S.L., Jella, T.K., Jobanputra, A.M., Johnson, A.P., Junna, M.R., Kado, J.T., Kaffenberger, T.M., Kapur, V.K., Kezirian, E.J., Khan, M., Kirsch, D.B., Kominsky, A., Kryger, M., Krystal, A.D., Kushida, C.A., Kuzniar, T.J., Lam, D.J., Lettieri, C.J., Lim, D.C., Lin, H., Liu, S.Y.C., MacKay, S.G., Magalang, U.J., Malhotra, A., Mansukhani, M.P., Maurer, J.T., May, A.M., Mitchell, R.B., Mokhlesi, B., Mullins, A. E., Nada, E.M., Naik, S., Nokes, B., Olson, M.D., Pack, A.I., Pang, E.B., Pang, K.P., Patil, S.P., Van de Perck, E., Piccirillo, J.F., Pien, G.W., Piper, A.J., Plawecki, A., Quigg, M., Ravesloot, M.J.L., Redline, S., Rotenberg, B.W., Ryden, A., Sarmiento, K. F., Sbeih, F., Schell, A.E., Schmickl, C.N., Schotland, H.M., Schwab, R.J., Seo, J., Shah, N., Shelgikar, A.V., Shochat, I., Soose, R.J., Steele, T.O., Stephens, E., Stepnowsky, C., Strohl, K.P., Sutherland, K., Suurna, M.V., Thaler, E., Thapa, S., Vanderveken, O.M., de Vries, N., Weaver, E.M., Weir, I.D., Wolfe, L.F., Woodson, B. T., Won, C.H.J., Xu, J., Yalamanchi, P., Yaremchuk, K., Yeghiazarians, Y., Yu, J.L., Zeidler, M., Rosen, I.M., 2023. International consensus statement on obstructive sleep apnea. Int. Forum Allergy Rhinol 13, 1061–1482. https://doi.org/10.1002/alr.23079.

Chen, X., Wang, R., Zee, P., Lutsey, P.L., Javaheri, S., Alcántara, C., Jackson, C.L., Williams, M.A., Redline, S., 2015. Racial/Ethnic differences in sleep disturbances: The multi-ethnic study of atherosclerosis (MESA). Sleep 38, 877–888. https://doi.org/10.5665/sleep.4732.

Chinnasamy, P., Wong, W.-K., Raja, A.A., Khalaf, O.I., Kiran, A., Babu, J.C., 2023a. Health recommendation system using deep learning-based collaborative filtering. Heliyon 9, e22844. https://doi.org/10.1016/j.heliyon.2023.e22844.

Chinnasamy, P., Wong, W.-K., Raja, A.A., Khalaf, O.I., Kiran, A., Babu, J.C., 2023b. Health recommendation system using deep learning-based collaborative filtering. Heliyon 9, e22844. https://doi.org/10.1016/j.heliyon.2023.e22844.

Choi, E., Park, D.-H., Yu, J., Ryu, S.-H., Ha, J.-H., 2016. The severity of sleep disordered breathing induces different decrease in the oxygen saturation during rapid eye movement and non-rapid eye movement sleep. Psychiatry Investig 13, 652. https://doi.org/10.4306/pi.2016.13.6.652.

del Campo, F., Crespo, A., Cerezo-Hernández, A., Gutiérrez-Tobal, G.C., Hornero, R., Álvarez, D., 2018. Oximetry use in obstructive sleep apnea. Expet Rev. Respir. Med. 12, 665–681. https://doi.org/10.1080/17476348.2018.1495563.

Dutt, M., Redhu, S., Goodwin, M., Omlin, C.W., 2022. SleepXAI: an explainable deep learning approach for multi-class sleep stage identification. Appl. Intell. https://doi.org/10.1007/s10489-022-04357-8.

Edwards, B.A., O'Driscoll, D.M., Ali, A., Jordan, A.S., Trinder, J., Malhotra, A., 2010. Aging and sleep: physiology and pathophysiology. Semin. Respir. Crit. Care Med. 31, 618–633. https://doi.org/10.1055/s-0030-1265902.

Faust, O., Razaghi, H., Barika, R., Ciaccio, E.J., Acharya, U.R., 2019. A review of automated sleep stage scoring based on physiological signals for the new millennia. Comput. Methods Progr. Biomed. 176, 81–91. https://doi.org/10.1016/j.cmpb.2019.04.032.

Fiorillo, L., Monachino, G., van der Meer, J., Pesce, M., Warncke, J.D., Schmidt, M.H., Bassetti, C.L.A., Tzovara, A., Favaro, P., Faraci, F.D., 2023a. U-Sleep's resilience to AASM guidelines. npj Digit. Med. 6, 33. https://doi.org/10.1038/s41746-023-00784-0.

Fiorillo, L., Pedroncelli, D., Agostini, V., Favaro, P., Faraci, F.D., 2023b. Multi-scored sleep databases: how to exploit the multiple-labels in automated sleep scoring. Sleep 46. https://doi.org/10.1093/sleep/zsad028.

Fiorillo, L., Puiatti, A., Papandrea, M., Ratti, P.L., Favaro, P., Roth, C., Bargiotas, P., Bassetti, C.L., Faraci, F.D., 2019. Automated sleep scoring: a review of the latest approaches. Sleep Med. Rev. 48, 101204. https://doi.org/10.1016/j.smrv.2019.07.007.

Goh, D.Y.T., Galster, P., Marcus, C.L., 2000. Sleep architecture and respiratory disturbances in children with obstructive sleep apnea. Am. J. Respir. Crit. Care Med. 162, 682–686. https://doi.org/10.1164/ajrccm.162.2.9908058.

Guilleminault, C., Winkle, R., Connolly, S., Melvin, K., Tilkian, A., 1984. Cyclical variation of the heart rate in sleep apnoea syndrome: mechanisms, and usefulness of 24 h electrocardiography as a screening technique. Lancet 323, 126–131. https://doi.org/10.1016/S0140-6736(84)90062-X.

Gutiérrez-Tobal, G., Álvarez, D., Gomez-Pilar, J., Del Campo, F., Hornero, R., 2015. Assessment of time and frequency domain entropies to detect sleep apnoea in heart rate variability recordings from men and women. Entropy 17, 123–141. https://doi.org/10.3390/e17010123.

Haimov, S., Tabakhov, A., Tauman, R., Behar, J.A., 2025. Deep learning for pediatric sleep staging from photoplethysmography: a transfer learning approach from adults to children. IEEE Trans. Biomed. Eng. 72, 760–767. https://doi.org/10.1109/TBME.2024.3470534.

Hamon, R., Junklewitz, H., Sanchez, I., 2020. Robustness and Explainability of Artificial Intelligence. Publ. Off. Eur. Union.

Huttunen, R., Leppanen, T., Duce, B., Arnardottir, E.S., Nikkonen, S., Myllymaa, S., Toyras, J., Korkalainen, H., 2022. A comparison of signal combinations for deep learning-based simultaneous sleep staging and respiratory event detection. IEEE Trans. Biomed. Eng. 70, 1704–1714. https://doi.org/10.1109/TBME.2022.3225268.

Huttunen, R., Leppänen, T., Duce, B., Oksenberg, A., Myllymaa, S., Töyräs, J., Korkalainen, H., 2021. Assessment of obstructive sleep apnea-related sleep fragmentation utilizing deep learning-based sleep staging from photoplethysmography. Sleep 44, 1–10. https://doi.org/10.1093/sleep/zsab142.

Iber, C., Ancoli-Israel, S., Chesson, A., Quan, S.F., 2007. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specification. J. Clin. Sleep Med. 3, 752. https://doi.org/10.1017/CBO9781107415324.004.

Imtiaz, S.A., 2021. A systematic review of sensing technologies for wearable sleep staging. Sensors 21, 1–21. https://doi.org/10.3390/s21051562.

Jiménez-García, J., García, M., Gutiérrez-Tobal, G.C., Kheirandish-Gozal, L., Vaquerizo-Villar, F., Álvarez, D., del Campo, F., Gozal, D., Hornero, R., 2024. An explainable deep-learning architecture for pediatric sleep apnea identification from overnight airflow and oximetry signals. Biomed. Signal Process Control 87. https://doi.org/10.1016/j.bspc.2023.105490.

Kalevo, L., Miettinen, T., Leino, A., Westeren-Punnonen, S., Sahlman, J., Mervaala, E., Toyras, J., Leppanen, T., Myllymaa, S., Myllymaa, K., 2022. Self-applied electrode set provides a clinically feasible solution enabling EEG recording in home sleep apnea testing. IEEE Access 10, 60633–60642. https://doi.org/10.1109/ACCESS.2022.3178189.

Korkalainen, H., Aakko, J., Duce, B., Kainulainen, S., Leino, A., Nikkonen, S., Afara, I.O., Myllymaa, S., Töyräs, J., Leppänen, T., 2020. Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea. Sleep 43, 1–10. https://doi.org/10.1093/sleep/zsaa098.

Korkalainen, H., Leppanen, T., Duce, B., Kainulainen, S., Aakko, J., Leino, A., Kalevo, L., Afara, I.O., Myllymaa, S., Toyras, J., 2021. 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. Inform. 25, 2567–2574. https://doi.org/10.1109/JBHI.2020.3043507.

Kotzen, K., Charlton, P.H., Salabi, S., Amar, L., Landesberg, A., Behar, J.A., 2023. SleepPPG-Net: a deep learning algorithm for robust sleep staging from continuous photoplethysmography. IEEE J. Biomed. Heal. Inform. 27, 924–932. https://doi.org/10.1109/JBHI.2022.3225363.

Krauss, D., Richer, R., Küderle, A., Jukic, J., German, A., Leutheuser, H., Regensburger, M., Winkler, J., Eskofier, B.M., 2025. Incorporating respiratory signals for ML-based multi-modal sleep stage classification: a large-scale benchmark study with actigraphy and HRV. Sleep. https://doi.org/10.1093/sleep/zsaf091.

Kuo, C.E., Chen, G.T., Liao, P.Y., 2021. An EEG spectrogram-based automatic sleep stage scoring method via data augmentation, ensemble convolution neural network, and expert knowledge. Biomed. Signal Process. Contr. 70, 102981. https://doi.org/10.1016/j.bspc.2021.102981.

Lee, Y.J., Lee, J.Y., Cho, J.H., Choi, J.H., 2022. Interrater reliability of sleep stage scoring: a meta-analysis. J. Clin. Sleep Med. 18, 193–202. https://doi.org/10.5664/jcsm.9538.

Levendowski, D.J., Neylan, T.C., Lee-Iannotti, J.K., Timm, P.C., Guevarra, C., Angel, E., Shprecher, D., Mazeika, G., Walsh, C.M., Boeve, B.F., St Louis, E.K., 2023. The accuracy and reliability of sleep staging and sleep biomarkers in patients with isolated rapid eye movement sleep behavior disorder. Nat. Sci. Sleep 15, 323–331. https://doi.org/10.2147/NSS.S396853.

Marcus, C.L., Moore, R.H., Rosen, C.L., Giordani, B., Garetz, S.L., Taylor, H.G., Mitchell, R.B., Amin, R., Katz, E.S., Arens, R., Paruthi, S., Muzumdar, H., Gozal, D., Thomas, N.H., Ware, J., Beebe, D., Snyder, K., Elden, L., Sprecher, R.C., Willging, P., Jones, D., Bent, J.P., Hoban, T., Chervin, R.D., Ellenberg, S.S., Redline, S., 2013. A randomized trial of adenotonsillectomy for childhood sleep apnea. N. Engl. J. Med. 368, 2366–2376. https://doi.org/10.1056/NEJMoa1215881.

Martín-Montero, A., Armañac-Julián, P., Gil, E., Kheirandish-Gozal, L., Álvarez, D., Lázaro, J., Bailón, R., Gozal, D., Laguna, P., Hornero, R., Gutiérrez-Tobal, G.C., 2023. Pediatric sleep apnea: characterization of apneic events and sleep stages using heart rate variability. Comput. Biol. Med. 154, 106549. https://doi.org/10.1016/j.compbiomed.2023.106549.

Martín-Montero, A., Gutiérrez-Tobal, G.C., Kheirandish-Gozal, L., Jiménez-García, J., Álvarez, D., del Campo, F., Gozal, D., Hornero, R., 2021. Heart rate variability spectrum characteristics in children with sleep apnea. Pediatr. Res. 89, 1771–1779. https://doi.org/10.1038/s41390-020-01138-2.

Nam, B., Bark, B., Lee, J., Kim, I.Y., 2024. InsightSleepNet: the interpretable and uncertainty-aware deep learning network for sleep staging using continuous photoplethysmography. BMC Med. Inform. Decis. Mak. 24, 50. https://doi.org/10.1186/s12911-024-02437-y.

Orwoll, E., Blank, J.B., Barrett-Connor, E., Cauley, J., Cummings, S., Ensrud, K., Lewis, C., Cawthon, P.M., Marcus, R., Marshall, L.M., McGowan, J., Phipps, K., Sherman, S., Stefanick, M.L., Stone, K., 2005. Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study - a large observational study of the determinants of fracture in older men. Contemp. Clin. Trials 26, 569–585. https://doi.org/10.1016/j.cct.2005.05.006.

Penzel, T., Kantelhardt, J.W., Grote, L., Peter, J., Bunde, A., 2003. Comparison of detrended fluctuation analysis and spectral analysis for heart rate variability in sleep and sleep apnea. IEEE Trans. Biomed. Eng. 50, 1143–1151. https://doi.org/10.1109/TBME.2003.817636.

Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P.J., Igel, C., 2021. U-Sleep: resilient high-frequency sleep staging. npj Digit. Med. 4, 72. https://doi.org/10.1038/s41746-021-00440-5.

Phan, H., Mikkelsen, K.B., Chen, O., Koch, P., Mertins, A., De Vos, M., 2022. SleepTransformer: automatic sleep staging with interpretability and uncertainty quantification. IEEE Trans. Biomed. Eng. 69, 2456–2467. https://doi.org/10.1109/TBME.2022.3147187.

Qin, H., Steenbergen, N., Glos, M., Wessel, N., Kraemer, J.F., Vaquerizo-Villar, F., Penzel, T., 2021. The different facets of heart rate variability in obstructive sleep apnea. Front. Psychiatr. 12, 1–20. https://doi.org/10.3389/fpsyt.2021.642333.

Quan, S.F., Howard, B.V., Iber, C., Kiley, J.P., Nieto, F.J., O'Connor, G.T., Rapoport, D. M., Redline, S., Robbins, J., Samet, J.M., Wahl, P.W., 1997. The sleep heart health study: design, rationale, and methods. Sleep 20, 1077–1085. https://doi.org/10.1093/sleep/20.12.1077.

Radha, M., Fonseca, P., Moreau, A., Ross, M., Cerny, A., Anderer, P., Long, X., Aarts, R. M., 2021. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. npj Digit. Med. 4, 1–11. https://doi.org/10.1038/s41746-021-00510-8.

Redline, S., Tishler, P.V., Tosteson, T.D., Williamson, J., Kump, K., Browner, I., Ferrette, V., Krejci, P., 1995. The familial aggregation of obstructive sleep apnea. Am. J. Respir. Crit. Care Med. 151, 682–687. https://doi.org/10.1164/ajrccm/151.3_Pt_1.682.

Rioul, O., Vetterli, M., 1991. Wavelets and signal processing. IEEE Signal Process. Mag. 8, 14–38. https://doi.org/10.1109/79.91217.

Ronneberger, O., Fischer, P., Brox, T., 2015. In: U-net: convolutional networks for biomedical image segmentation, 18, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

Rosen, C.L., Larkin, E.K., Kirchner, H.L., Emancipator, J.L., Bivins, S.F., Surovec, S.A., Martin, R.J., Redline, S., 2003. Prevalence and risk factors for sleep-disordered breathing in 8- to 11-year-old children: association with race and prematurity. J. Pediatr. 142, 383–389. https://doi.org/10.1067/mpd.2003.28.

Sateia, M.J., 2014. International classification of sleep disorders-third edition. Chest 146, 1387–1394. https://doi.org/10.1378/chest.14-0970.

Shaffer, F., Ginsberg, J.P., 2017. An overview of heart rate variability metrics and norms. Front. Public Heal. 5, 1–17. https://doi.org/10.3389/fpubh.2017.00258.

Sridhar, N., Shoeb, A., Stephens, P., Kharbouch, A., Shimol, D. Ben, Burkart, J., Ghoreyshi, A., Myers, L., 2020. Deep learning for automated sleep staging using instantaneous heart rate. npj Digit. Med. 3, 106. https://doi.org/10.1038/s41746-020-0291-x.

Stephansen, J.B., Olesen, A.N., Olsen, M., Ambati, A., Leary, E.B., Moore, H.E., Carrillo, O., Lin, L., Han, F., Yan, H., Sun, Y.L., Dauvilliers, Y., Scholz, S., Barateau, L., Hogl, B., Stefani, A., Hong, S.C., Kim, T.W., Pizza, F., Plazzi, G., Vandi, S., Antelmi, E., Perrin, D., Kuna, S.T., Schweitzer, P.K., Kushida, C., Peppard, P.E., Sorensen, H.B.D., Jennum, P., Mignot, E., 2018. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. Nat. Commun. 9, 1–15. https://doi.org/10.1038/s41467-018-07229-3.

Tripathi, P., Ansari, M.A., Gandhi, T.K., Mehrotra, R., Heyat, M.B. Bin, Akhtar, F., Ukwuoma, C.C., Muaad, A.Y., Kadah, Y.M., Al-Antari, M.A., Li, J.P., 2022. Ensemble computational intelligent for insomnia sleep stage detection via the sleep ECG signal. IEEE Access 10, 108710–108721. https://doi.org/10.1109/ACCESS.2022.3212120.

Vaquerizo-Villar, F., Álvarez, D., Gutiérrez-Tobal, G.C., Mart\'in-Montero, A., Gozal, D., Tamayo, E., Hornero, R., 2024. Accurate and interpretable deep learning model for sleep staging in children with sleep apnea from pulse oximetry. In: European Medical and Biological Engineering Conference, pp. 38–47. https://doi.org/10.1007/978-3-031-61625-9_5.

Vaquerizo-Villar, F., Álvarez, D., Kheirandish-Gozal, L., Gutiérrez-Tobal, G.C., Barroso-García, V., Crespo, A., del Campo, F., Gozal, D., Hornero, R., 2018. Utility of bispectrum in the screening of pediatric sleep apnea-hypopnea syndrome using oximetry recordings. Comput. Methods Progr. Biomed. 156, 141–149. https://doi.org/10.1016/j.cmpb.2017.12.020.

Vaquerizo-Villar, F., Gutiérrez-Tobal, G.C., Calvo, E., Álvarez, D., Kheirandish-Gozal, L., del Campo, F., Gozal, D., Hornero, R., 2023. An explainable deep-learning model to stage sleep states in children and propose novel EEG-related patterns in sleep apnea. Comput. Biol. Med. 165, 107419. https://doi.org/10.1016/j.compbiomed.2023.107419.

Varga, A.W., Mokhlesi, B., 2019. REM obstructive sleep apnea: risk for adverse health outcomes and novel treatments. Sleep Breath. 23, 413–423. https://doi.org/10.1007/s11325-018-1727-2.

Vinogradova, K., Dibrov, A., Myers, G., 2020. Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 13943–13944. https://doi.org/10.1609/aaai.v34i10.7244.

Wachowiak, M.P., Hay, D.C., Johnson, M.J., 2016. Assessing heart rate variability through wavelet-based statistical measures. Comput. Biol. Med. 77, 222–230. https://doi.org/10.1016/j.compbiomed.2016.07.008.

Wang, C., Jiang, X., Lv, C., Meng, Q., Zhao, P., Yan, D., Feng, C., Xu, F., Lu, S., Jung, T.-P., Leng, J., 2025. GraphSleepFormer: a multi-modal graph neural network for sleep staging in OSA patients. J. Neural. Eng. 22, 026011. https://doi.org/10.1088/1741-2552/adb996.

Wulterkens, B.M., Fonseca, P., Hermans, L.W.A., Ross, M., Cerny, A., Anderer, P., Long, X., van Dijk, J.P., Vandenbussche, N., Pillen, S., van Gilst, M.M., Overeem, S., 2021. It is all in the wrist: wearable sleep staging in a clinical population versus reference polysomnography. Nat. Sci. Sleep 13, 885–897. https://doi.org/10.2147/NSS.S306808.

Yang, G., Ye, Q., Xia, J., 2022. 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, 29–52. https://doi.org/10.1016/j.inffus.2021.07.016.

You, J., Jiang, D., Ma, Y., Wang, Y., 2021. SpindleU-Net: an adaptive U-Net framework for sleep spindle detection in single-channel EEG. IEEE Trans. Neural Syst. Rehabil. Eng. 29, 1614–1623. https://doi.org/10.1109/TNSRE.2021.3105443.

Zhang, G.-Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., Redline, S., 2018. The national sleep research resource: towards a sleep data commons. J. Am. Med. Inf. Assoc. 25, 1351–1358. https://doi.org/10.1093/jamia/ocy064.