

RESEARCH ARTICLE

Apnea Burden-Guided Framework: Enhancing Out-of-Distribution Generalization in PPG-Based Sleep Apnea Characterization

MANTAS RINKEVIČIUS¹, OSKAR PFEFFER², AMAL ALISSA^{2,3}, NANDO HEGEMANN²,
AND VAIDOTAS MAROZAS^{1,4}, (Member, IEEE)

¹Biomedical Engineering Institute, Kaunas University of Technology, 51423 Kaunas, Lithuania

²Physikalisch-Technische Bundesanstalt, 38116 Brunswick, Germany

³Freie Universität Berlin, 14195 Berlin, Germany

⁴Faculty of Electrical and Electronics Engineering, Kaunas University of Technology, 51368 Kaunas, Lithuania

Corresponding author: Mantas Rinkevičius (mantas.rinkevicius@ktu.lt)

This work was supported by the European Partnership on Metrology, co-financed by the European Union Horizon Europe Research and Innovation Program and the Participating States under Project 22HLT01 QUMPHY.

ABSTRACT Background: Sleep apnea is a common sleep-related breathing disorder associated with substantial cardiovascular and metabolic risk. Although overnight polysomnography remains the reference standard for diagnosis, its complexity and cost limit its suitability for long-term preventive monitoring at home. In wearable systems, arterial blood oxygen saturation is commonly used as the main predictor, whereas additional morphological features of the photoplethysmographic pulse wave are usually underexplored. Objective: This study proposes a novel apnea burden prediction-based framework for sleep apnea severity assessment and investigates the influence of photoplethysmographic features on model performance and out-of-distribution generalization. Methods: The proposed framework first predicts apnea burden as a continuous measure, which is subsequently converted into the clinically relevant apnea-hypopnea index for subject-level classification into four severity groups in both in-distribution and out-of-distribution data. Three artificial neural network architectures were evaluated, and the performance metrics were averaged over five independent runs with different fixed random seeds. Results: During out-of-distribution testing, the combination of photoplethysmographic features and arterial blood oxygen saturation led to increases of approximately 15.72% in macro-sensitivity, 9.22% in macro-accuracy, 16.01% in macro-F1-score, 11.08% in Cohen's kappa, and 13.22% in Matthews correlation coefficient, compared with using arterial blood oxygen saturation alone. The low-complexity convolutional-recurrent models achieved the highest overall performance. Conclusion: The results indicated that the proposed apnea burden-guided framework, combined with photoplethysmographic features and arterial blood oxygen saturation, improves sleep apnea characterization while showing encouraging out-of-distribution performance on an independent external dataset. Moreover, simpler hybrid architectures demonstrated strong potential for robust home-based preventive monitoring.

INDEX TERMS Apnea burden, apnea-hypopnea index, arterial blood oxygen saturation, convolutional-recurrent networks, home-based preventive monitoring, out-of-distribution generalization, photoplethysmographic features, sleep apnea characterization.

I. INTRODUCTION

Sleep apnea is a common sleep-related breathing disorder characterized by recurrent episodes of partial or complete

The associate editor coordinating the review of this manuscript and approving it for publication was Muammar Muhammad Kabir¹.

upper airway obstruction during sleep, leading to intermittent hypoxia, sleep fragmentation, daytime sleepiness, and increased cardiovascular risk [1], [2], [3], [4]. Epidemiological studies indicate that a substantial proportion of the adult population is affected (from 3% to 28%), with prevalence increasing due to aging and obesity [5], [6].

The “gold standard” for sleep apnea diagnosis is polysomnography (PSG), a comprehensive overnight examination that records multiple physiological signals, including respiratory airflow, electrocardiography (ECG), electroencephalography (EEG), and arterial blood oxygen saturation (SpO₂). During a PSG test, sleep apnea severity is clinically quantified using the apnea-hypopnea index (AHI), defined as the average number of apnea (complete cessation) and hypopnea (partial reduction in airflow) events lasting ≥ 10 s per hour of sleep, associated with $\geq 3\%$ oxygen desaturation [7], [8]. According to [8], sleep apnea is classified into four categories: normal ($\text{AHI} < 5$), mild ($5 \leq \text{AHI} < 15$), moderate ($15 \leq \text{AHI} < 30$), and severe ($\text{AHI} \geq 30$). Despite the high diagnostic accuracy of PSG in estimating the AHI, it remains costly, complex, and impractical for long-term monitoring in home environments.

To overcome these limitations, wearable-based approaches using photoplethysmography (PPG), which reflects peripheral blood volume variations, and SpO₂ have gained increasing attention [9], [10], [11]. In current practice, SpO₂, derived from dual-wavelength PPG signals (typically red and infrared), is the most commonly used predictor for sleep apnea characterization in wearable devices designed for home-based preventive monitoring. In some cases, heart/pulse rate, derived from a single channel PPG, is also considered as an additional feature. However, the morphological characteristics of the PPG waveform are typically not exploited, and the valuable physiological information contained in pulse wave dynamics remains largely underutilized. Figure 1 illustrates typical changes in respiratory airflow, PPG signal, SpO₂, and heart rate during apnea/hypopnea episodes, highlighting that sleep apnea affects multiple physiological factors beyond oxygen desaturation alone. Due to reduced airflow, SpO₂ decreases, whereas heart rate typically exhibits a transient reduction during respiratory events, followed by a sudden post-event increase as part of the arousal response, resulting in a U-shaped pattern.

Early studies have demonstrated that PPG signals contain relevant patterns related to cardiovascular and autonomic responses to apnea events, including pulse rate variability and amplitude modulation [12], [13], [14], [15] (see Figure 1b). For instance, a study [13] employed a combination of PPG and SpO₂ signals for sleep apnea event detection, showing promising capability to distinguish between central and obstructive apnea types, as well as apnea and hypopnea events. Their findings suggest that incorporating PPG-derived features enhances sleep apnea severity assessment beyond conventional SpO₂-based detectors.

Recent advances in deep learning have significantly improved the analysis of physiological signals by enabling automatic extraction of complex temporal patterns [16], [17], [18]. Artificial neural networks, including convolutional and recurrent architectures, have been successfully applied to sleep apnea detection using pulse rate and SpO₂ time series [19], [20]. In particular, hybrid convolutional-recurrent

models have demonstrated strong capability in learning both local and long-term temporal features. For instance, the DeepSleepNet model was proposed for automatic sleep stage scoring based on raw single-channel EEG [21], and the ApneaNet architecture was introduced to detect obstructive sleep apnea using ECG signal analysis [22]. More recent works have explored advanced architectures, including transformer-based models [23], [24], [25], [26] for processing EEG, ECG or SpO₂ raw signals.

However, a critical limitation of most proposed home-oriented approaches based on PPG is their reliance on SpO₂ as the primary predictor of sleep apnea [26], [27], [28], [29], [30], [31], while overlooking pulse wave morphological features. This may limit the ability to capture more subtle physiological changes in PPG morphology and reduce generalizability when applied to data from different distributions as studies [13], [26], [27], [29] performed with only one dataset. In this context, in-distribution (ID) data refers to samples drawn from the same distribution as the training data, whereas out-of-distribution (OOD) data corresponds to samples originating from different populations, devices, or acquisition conditions. Models often perform well on ID data but exhibit performance degradation on OOD datasets. Recent studies have highlighted the importance of addressing this issue. For example, the solely PPG-based ApSense study emphasized the need for multi-dataset validation [32], whereas the SleepPPG-Net2 explicitly addressed cross-dataset generalization for sleep staging from raw PPG time series [33].

In this study, we hypothesize that combining PPG pulse wave features with SpO₂ in artificial neural networks improves classification accuracy and enhances OOD generalization. In addition, we propose a couple of novel and underexplored features for sleep apnea assessment. Therefore, this work investigates the influence of PPG-derived features on model performance under distributional shifts. We employ a novel sleep apnea burden (AB) prediction framework and compare three network architectures of varying complexity with different predictor combinations. The AB metric is defined as the proportion of time affected by respiratory events, thus providing a continuous representation of disease severity. While AHI is widely used in clinical practice, AB offers a complementary perspective by capturing the temporal extent of respiratory disturbances. Our proposed approach enables continuous estimation of AB, which is subsequently transformed into the clinically relevant AHI metric, allowing subject-level severity classification, consistent with clinical practice.

The main contributions of this study are summarized as follows:

- 1) A model-agnostic AB regression-based framework is proposed for sleep apnea severity assessment using PPG-derived features and SpO₂.
- 2) The contribution of different PPG feature combinations, including the novel and underexplored features, is systematically evaluated.

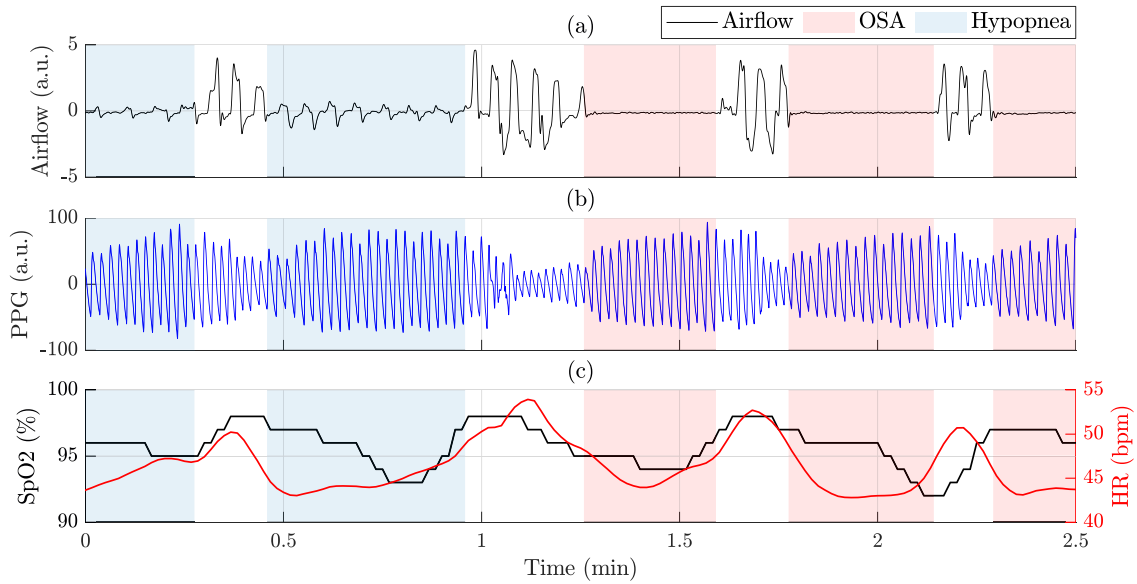


FIGURE 1. The changes of signals during obstructive sleep apnea–OSA (complete cessation) and hypopnea (partial reduction in airflow) episodes: (a) respiratory airflow, (b) finger PPG signal, (c) arterial blood oxygen saturation, SpO2, and heart rate, HR.

- 3) The proposed framework is assessed under both ID and OOD scenarios using three representative deep learning architectures.
- 4) The potential advantages of regression-based sleep apnea characterization over conventional classification-based approaches are discussed.

The remainder of this paper is organized as follows. Section II presents the materials and methods, including the dataset description, data partitioning and preprocessing, PPG feature extraction, and the investigated model architectures. Section III reports the experimental results, including parameter settings, performance of different model architectures, and comparison of predictor combinations. Section IV discusses the main findings, highlighting the novelty of the proposed method, the impact of model architectures and features, as well as limitations of the study and directions for future work. Finally, Section V concludes the obtained results.

II. MATERIALS AND METHODS

This section describes the datasets used and the research workflow for the proposed AB-guided sleep apnea severity assessment framework. The proposed research workflow consists of (see Figure 2): (i) signal preprocessing, (ii) predictor preparation, (iii) model training, (iv) severity assessment, and (v) performance evaluation. A pseudocode of the proposed AB-guided framework is provided in Algorithm 1.

A. DATASET DESCRIPTION

1) MESA DATA

In this work, the Multi-Ethnic Study of Atherosclerosis (MESA) dataset [34] was used for model training, validation, and ID testing. The MESA sleep study comprises data from 2055 subjects (53.63% female), aged between 54 and

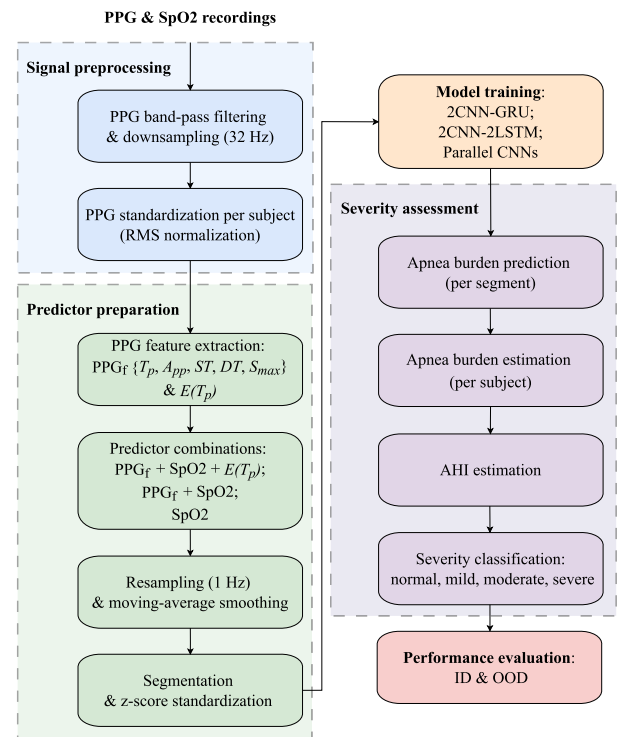


FIGURE 2. The proposed workflow for the AB-guided framework used to process PPG and SpO2 time series for sleep apnea severity assessment. PPG_f corresponds to five pulse wave features, whereas $E(T_p)$ is a new proposed higher-level feature.

95 years. Participants included in the study had not used treatments for sleep apnea, such as continuous positive airway pressure or oxygen devices, for more than a month prior to the study, or only used it less than once a week.

Algorithm 1 Pseudocode of the Proposed AB-Guided Framework

Require: M test subjects; N_j segments for subject j
 Segmented predictor sequences: $\mathcal{X}_{tr}$, $\mathcal{X}_{val}$, $\mathcal{X}_{test}$
 Reference segment-level AB values: $\mathcal{Y}_{tr}$, $\mathcal{Y}_{val}$, $\mathcal{Y}_{test}$
 Training data-derived mean event duration D_{mean}
 Reference severity class c_j

Ensure: Predicted subject-level $\widehat{AB}_j$
 Estimated apnea-hypopnea index $\widehat{AHI}_j$
 Predicted severity class $\widehat{c}_j$

- 1: Estimate the feature-wise z-score normalization parameters μ_{tr} and σ_{tr} from $\mathcal{X}_{tr}$
- 2: Standardize all subsets ($\mathcal{X}_{tr}$, $\mathcal{X}_{val}$, $\mathcal{X}_{test}$) using μ_{tr} and σ_{tr} values
- 3: Initialize the selected model architecture network f_θ
- 4: Train f_θ using $\mathcal{X}_{tr}$ and $\mathcal{Y}_{tr}$
- 5: Select the model parameters θ^* according to the validation loss using $\mathcal{X}_{val}$ and $\mathcal{Y}_{val}$
- 6: **for** $j = 1$ to M **do**
- 7: **for** $s = 1$ to N_j **do**
- 8: $\widehat{AB}_{j,s} \leftarrow f_{\theta^*}(\mathcal{X}_{test}^{j,s})$
- 9: **end for**
- 10: $\widehat{AB}_j \leftarrow \text{Aggregate}(\widehat{AB}_{j,1}, \dots, \widehat{AB}_{j,N_j})$
- 11: $\widehat{AHI}_j \leftarrow \text{Convert}(\widehat{AB}_j, D_{mean})$ according to Equation II-E
- 12: $\widehat{c}_j \leftarrow \text{Classify}(\widehat{AHI}_j)$
- 13: **end for**
- 14: **return** $\{\widehat{AB}_j, \widehat{AHI}_j, \widehat{c}_j\}_{j=1}^M$
- 15: Evaluate differences between c_j and $\widehat{c}_j$

The overnight PSG recordings obtained at home include PPG and SpO2 time series. The sampling rate of PPG signals is 256 Hz, whereas SpO2 is sampled at 1 Hz. Time labels and durations of apnea and hypopnea episodes in the respiratory flow signal were annotated by trained technicians. Due to potential PPG signal quality issues, we included only the 276 subjects whose PSG studies are labeled with the highest quality grade of 7, as in [32].

For the MESA dataset, the PSG study quality grades and corresponding AHI values were obtained from the `mesa-sleep-dataset-0.4.0.csv` file, specifically from the `overall15` and `ahi-a0h3a` columns, respectively. While the AB labels for each subject were computed based on the annotated apnea/hypopnea event onset time and durations provided in the corresponding XML files.

The MESA dataset has already been used in the following recent studies [32], [35], [36], [37].

2) OSASUD DATA

To assess OOD generalization, we tested models on the open-access Obstructive Sleep Apnea Stroke Unit Dataset (OSASUD) [38]. This PSG dataset involves 30 after-stroke patients (36.67% female) admitted due to suspected cerebrovascular events, such as ischemic and hemorrhagic

stroke, or transient ischemic attack. Patients aged <18 years, those unable to comply with standard requirements for PSG monitoring, individuals with severe aphasia impacting understanding or consent, and those at high risk of alcohol or drug withdrawal syndrome were not included. Notably, conditions such as obesity, diabetes mellitus, atrial fibrillation and other cardiac diseases did not serve as exclusion criteria.

The dataset includes PSG signals such as finger PPG and SpO2. The sampling rate of PPG signals is 80 Hz, whereas SpO2 time series are sampled at 1 Hz. In addition, SpO2 values <50% or >100% were identified as artifacts and marked as null.

The OSASUD data were annotated by a trained sleep medicine physician based on the sleep scoring guidelines established by the American Academy of Sleep Medicine, identifying occurrences of central, obstructive, mixed apnea, and hypopnea events (1 s granularity).

For the OSASUD dataset, the AHI labels were obtained from the original study [38]. Meanwhile, the AB values were computed based on the annotated apnea and hypopnea events with a temporal resolution of 1 s.

The OSASUD dataset has already been used in the following recent studies [26], [39], [40], [41]. A comparison of the two datasets used is provided in Table 1.

B. DATASET PARTITIONING AND PREPROCESSING

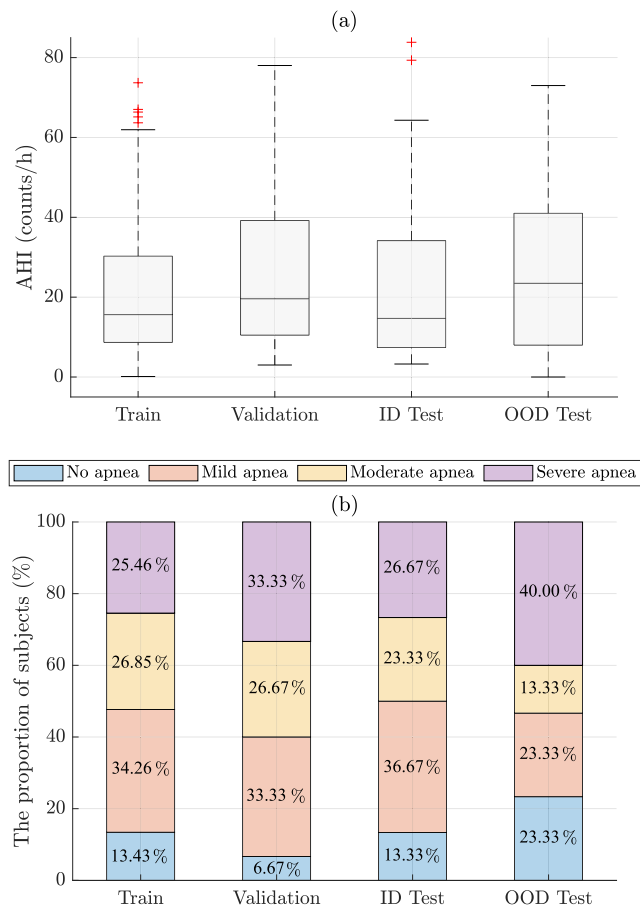
This study included a total of 306 subjects, consisting of 276 individuals from the MESA cohort and 30 patients from the OSASUD cohort. The MESA dataset was used for model training and ID evaluation, and partitioned into training (216 subjects), validation (30 subjects), and ID testing (30 subjects) subsets. The OSASUD dataset, containing 30 subjects from a clinically distinct population, was utilized for OOD evaluation to assess model robustness under domain shift.

The distributions of AHI values in prepared training, validation, ID testing, and OOD testing data are provided in Figure 3. As shown in Figure 3b, the ID testing data is dominated by mild apnea cases, followed by severe, moderate, and no apnea cases. In contrast, the OOD testing data exhibits a higher proportion of severe apnea cases, whereas moderate apnea represents the smallest group.

In this study, the PPG signals were preprocessed using a zero-phase, fourth-order Butterworth band-pass filter (0.4–9 Hz) to mitigate noise and enhance fiducial point detection. These PPG signals were then resampled to $f_s = 32$ Hz using a low-pass finite impulse response filter to prevent aliasing. This sampling rate is sufficient to capture relevant cardiovascular dynamics [42] and PPG morphology changes while reducing computational load. In addition, our experiments have shown that the PPG sampling frequency, provided that the Nyquist criterion ($f_{max} \leq f_s/2$, where $f_{max} \approx 9$ –10 Hz) is satisfied, does not have a significant impact on the performance of PPG feature-based models. To mitigate amplitude-related distributional

TABLE 1. Differences between MESA and OSASUD datasets. PSG–polysomnography, ECG–electrocardiography, EEG–electroencephalography, EOG–electrooculography, EMG–electromyography.

Dataset differences	#1 MESA	#2 OSASUD
Number of subjects (female %)	2055 (53.63%)	30 (36.67%)
Subject age	69.37 ± 9.12 years	68.97 ± 11.22 years
Collecting time	2010–2012	2019–2020
PSG signals	PPG, ECG, EEG, EOG, EMG, SpO2, heart rate, nasal airflow, snoring, 3-axis accelerometer, abdominal and thoracic movements	PPG, ECG, SpO2, respiratory rate, perfusion index, heart rate, nasal airflow, snoring, 3-axis accelerometer, abdominal and thoracic movements
Ethnicity	Diverse population: White, African American, Hispanic, and Chinese American participants	Not considered
Clinical status	General population	After-stroke patients
Environmental conditions	Home-based PSG: Participants from 6 clinical centres	Hospital-based PSG: Clinical Neurology Unit of Udine University Hospital, Italy
Device	Compumedics Somte monitoring system (Compumedics Ltd., Australia)	Embla RemLogic (Natus Medical Inc., Pleasanton, USA)
PPG sensor	Nonin 8000 sensor (finger PPG, sampling rate – 256 Hz)	Embletta polysomnograph + Mindray monitoring system (finger PPG, sampling rate – 80 Hz)

**FIGURE 3.** The distributions of AHI across dataset splits: (a) boxplots of AHI values for training, validation, ID test, and OOD test sets; (b) class distributions of subjects according to sleep apnea severity levels (no apnea, mild, moderate, and severe).

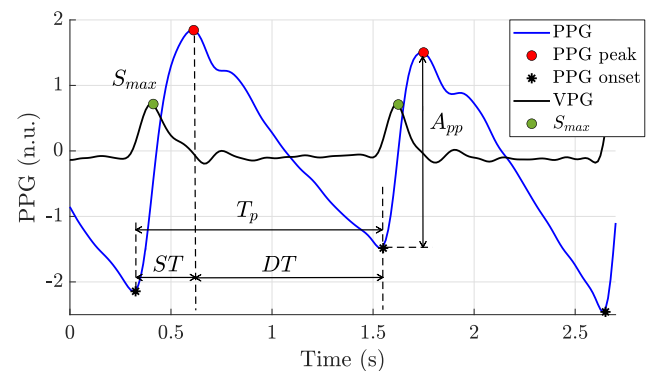
differences between datasets, overnight PPG recordings from each subject were independently standardized using root-mean-square (RMS) normalization prior to feature extraction.

C. MODEL PREDICTORS AND SEGMENTATION

To calculate the PPG-derived features, PPG peaks and onset points were identified using an automatic beat detection algorithm proposed in [43]. We extracted five pulse wave features (see Figure 4) and one proposed higher-level feature (see Figure 5):

- 1) pulse wave interval, T_p ;
- 2) peak-to-peak amplitude, A_{pp} ;
- 3) systolic time, ST ;
- 4) diastolic time, DT ;
- 5) maximum slope of PPG, S_{max} ;
- 6) mean envelope of pulse wave interval, $E(T_p)$.

The listed features were inspired by previous studies [12], [14], [32]. To evaluate the contribution of PPG-derived features, three different input combinations were investigated: (i) SpO2 combined with all six PPG features, including the proposed mean envelope feature $E(T_p)$, (ii) SpO2 combined with five PPG pulse wave features (T_p , ST , DT , A_{pp} , S_{max}), and (iii) SpO2 only.

**FIGURE 4.** PPG pulse wave feature delineation: pulse wave interval, T_p , systolic time, ST , diastolic time, DT , peak-to-peak amplitude, A_{pp} , and maximum slope of PPG, S_{max} , calculated from the first derivative of PPG signal, VPG (velocity photoplethysmogram).

These feature sequences were resampled to 1 Hz and then smoothed with a fifth-order moving-average filter.

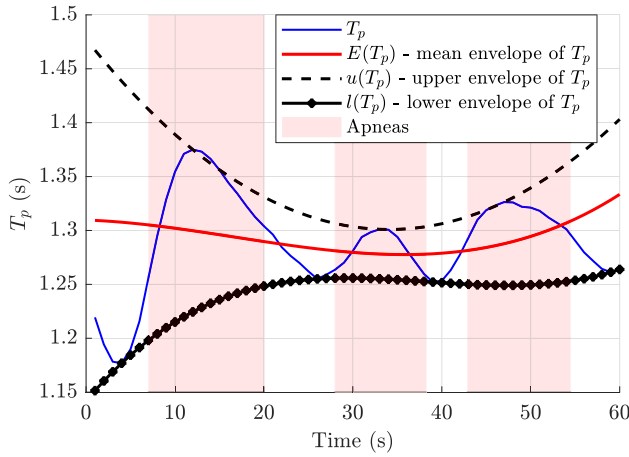


FIGURE 5. The example of the proposed PPG feature, the mean envelope of pulse wave interval time series, $E(T_p)$, with marked apnea episodes.

The typical variability of PPG feature sequences and SpO2 during sleep apnea episodes is illustrated in Figure 6.

Furthermore, we assume that the maximum slope of PPG, S_{max} , defined as the maximum value of its first derivative, may provide complementary information for evaluating respiratory disturbances. This feature reflects the maximum rate of blood volume changes during the systolic upstroke and may capture hemodynamic variations influenced by autonomic nervous system activity associated with sleep apnea. To the best of our knowledge, S_{max} has not been previously explored in this context, and it is therefore investigated in this study.

This study also proposes a novel PPG feature, the mean envelope of the pulse wave interval time series, $E(T_p)$, defined as the average of the upper and lower envelopes of the sequence within a 1-minute epoch (see Figure 5). The upper $u(T_p)$ and lower $l(T_p)$ envelopes are obtained by identifying local maxima and minima of the T_p sequence and interpolating between these extrema. Specifically, a peak-based method is used, where the envelopes are constructed from the detected peaks and troughs using a shape-preserving spline interpolation scheme.

The mean envelope is then computed as:

$$E(T_p) = \frac{u(T_p) + l(T_p)}{2}. \quad (1)$$

The mean envelope of the PPG pulse-to-pulse interval, $E(T_p)$, reflects low-frequency, slowly varying autonomic modulation of heart rate. Because pulse-to-pulse intervals shorten after apnea-induced sympathetic activation with a characteristic delay during arousal and lengthen during recovery (see Figure 6d), this proposed feature captures the minute-scale cardiovascular response to sleep apnea events and overall autonomic balance.

The analyzed six PPG features and SpO2 were divided into 5-minute segments. Each feature was then separately standardized according to the mean and standard deviation of the training data. We opted for a trade-off 5-minute

segmentation because 1-minute and smaller windows are likely too short to capture meaningful associations among PPG-derived features and SpO2 changes during sleep apnea events. While using much longer segments increases the computational load during model training.

D. MODEL ARCHITECTURES

Three artificial neural network architectures were investigated to evaluate different strategies for capturing temporal dependencies in PPG-derived features and SpO2 signals for AB prediction, and to assess the influence of model complexity on performance and generalization. The selected architectures represent different modeling paradigms: hybrid convolutional-recurrent models, namely 2CNN-GRU and 2CNN-2LSTM, with two sequential convolutional layers, followed by gated recurrent unit (GRU) and two stacked long short-term memory (LSTM) layers, respectively, and a purely convolutional multi-scale model (parallel CNNs).

1) 2CNN-GRU ARCHITECTURE

The proposed 2CNN-GRU architecture combines convolutional feature extraction with gated recurrent temporal modeling (see Figure 7a). Convolutional layers capture local patterns within short segments of the input features, whereas the GRU layer models temporal dependencies by learning how these patterns evolve over time in a computationally efficient manner. Compared to LSTM-based models, GRU offers reduced complexity and faster training, which is advantageous for home-based monitoring applications. This architecture was designed to provide a balance between performance and model simplicity.

2) 2CNN-2LSTM ARCHITECTURE

The 2CNN-2LSTM architecture follows a commonly used hybrid approach that integrates convolutional feature extraction with stacked LSTM layers for modeling long-range temporal dependencies (see Figure 7b). This type of architecture has been widely applied in physiological signal analysis and is inspired by previously proposed frameworks for PPG-based blood pressure and heart rate estimation [44]. It serves as a reference model to evaluate the impact of capturing more complex sequential dependencies compared to the proposed GRU-based architecture.

3) PARALLEL CNNs ARCHITECTURE

The parallel CNNs architecture is an Inception-inspired one-dimensional convolutional model (see Figure 7c) designed to capture multi-scale temporal features through parallel processing paths [45]. By extracting features at multiple temporal resolutions simultaneously, this architecture is particularly suitable for physiological signals, where relevant patterns may occur at different time scales. This model was included to assess whether convolutional multi-scale representations can compete with hybrid convolutional-recurrent approaches.

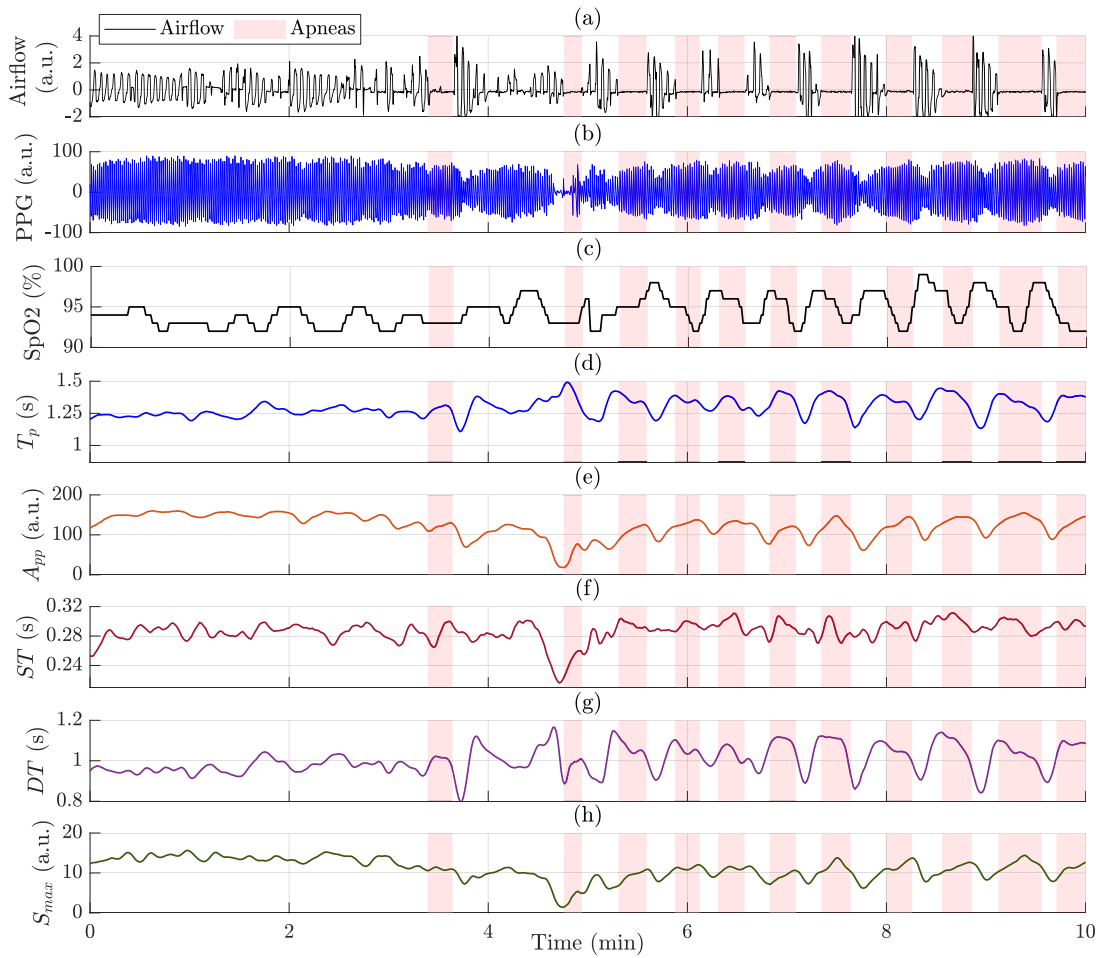


FIGURE 6. The changes of PPG feature sequences and SpO2 during apnea events: (a) respiratory airflow with marked apnea episodes, (b) PPG signal, (c) SpO2, (d) pulse wave interval, T_p , (e) peak-to-peak amplitude, A_{pp} , (f) systolic time, ST , (g) diastolic time, DT , and (h) maximum slope of PPG, S_{max} .

4) PARAMETER SETTINGS

The architectural hyperparameters were selected empirically based on preliminary experiments, with the final configuration chosen according to validation set performance.

The implemented models were trained using identical optimization settings, which were chosen after some minor tuning. The selected model training parameters are provided in Table 2. The models were trained using the mean squared error (MSE) loss function to predict continuous AB values.

TABLE 2. Model training settings.

Parameter	Value
Optimizer	Adam
Learning rate	0.001
Mini-batch size	1024
Maximum epochs	1000
Early stopping (patience)	15 epochs
Loss function	Mean squared error (MSE)

The mini-batch size was selected based on performed experiments, which indicated that this parameter had a

negligible effect on the final model performance. Therefore, a larger mini-batch size of 1024 was chosen to reduce computational time and improve training efficiency.

The learning rate was empirically optimized by comparing values of 0.1, 0.01, 0.001, and 0.0001. Among these, a learning rate of 0.001 consistently provided the best performance across implemented models and was therefore used in our experiments.

E. SEVERITY ASSESSMENT AND PERFORMANCE EVALUATION

To define the labels, the AB value was calculated for each 5-minute segment and then for the entire signal for each subject separately. In our study, the segment-level AB is defined as:

$$AB_{\text{segment}} = \frac{\sum d_i}{T_{\text{seg}}}, \quad (2)$$

where $\sum d_i$ denotes the total duration of apnea and hypopnea events within the segment, and T_{seg} is the segment length (in seconds).

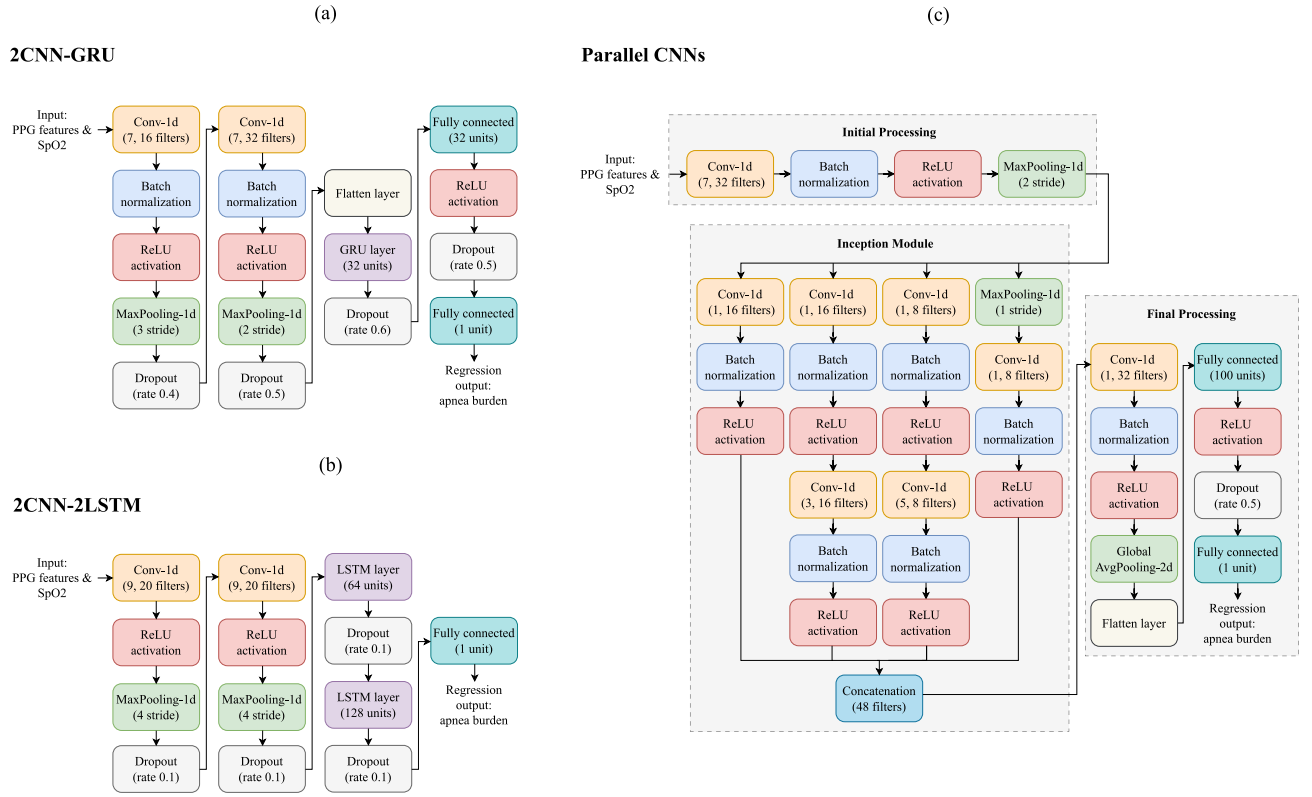


FIGURE 7. The block diagrams of the implemented model architectures: (a) 2CNN-GRU, (b) 2CNN-2LSTM, and (c) parallel CNNs.

The subject-level AB is defined by averaging the segment-wise values:

$$AB = \frac{1}{N} \sum_{k=1}^N AB_{\text{segment}}^k, \quad (3)$$

where N is the total number of segments in the recording.

The implemented models first predict AB for each 5-minute segment based on the extracted features. These predictions are then aggregated over the entire recording for each subject to estimate the subject-level AB (see Equation II-E), which is then converted into the AHI for severity assessment. The predicted AHI is computed as:

$$AHI_{AB} = \frac{AB \cdot 3600}{D_{\text{mean}}}, \quad (4)$$

where AB denotes the overall predicted AB, and D_{mean} represents the mean duration of apnea/hypopnea events (in seconds) calculated from overnight recordings of the training data.

This conversion is justified by the approximately linear relationship between AB and AHI observed in the datasets, as shown in Figure 8. Specifically, the scatter plots show that AHI_{AB} closely follows the reference AHI values along the identity line ($y = x$), indicating that AB provides a consistent surrogate for event frequency when normalized by the average event duration.

The mean absolute error with standard deviation between the estimated AHI_{AB} and the ground-truth AHI is 6.79 ± 6.67 counts/h for MESA data, and 4.99 ± 6.66 counts/h for OSASUD data, indicating the error introduced by the fixed-duration AB-to-AHI conversion.

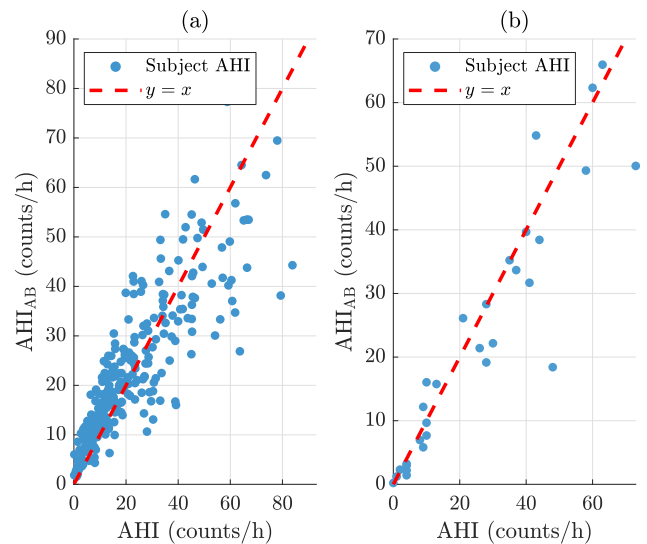


FIGURE 8. Scatter plots of reference AHI versus AB-derived AHI (AHI_{AB}): (a) MESA data (276 subjects); (b) OSASUD data (30 subjects). The dashed red line ($y = x$) indicates ideal agreement between the two measures.

Based on the estimated AHI_{AB} , the subjects were categorized into four sleep apnea severity groups (see Subsection II-A), enabling a multi-classification framework. This evaluation was performed for both ID and OOD test-subsets.

To assess model prediction performance, regression metrics were evaluated. These include mean absolute error (MAE) and mean absolute scaled error (MASE) for both predicted AB and estimated AHI.

The MAE is defined as:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|, \quad (5)$$

where y_i is the ground-truth value, $\hat{y}_i$ is the predicted value, and N is the number of samples.

The MASE was computed as:

$$MASE = \frac{MAE}{MAE_{baseline}}, \quad (6)$$

where $MAE_{baseline}$ denotes the MAE obtained by a constant predictor equal to the median AB or AHI value from the testing data.

To assess multi-classification performance [46], macro-averaged metrics were computed, including sensitivity, specificity, positive predictive value (PPV), accuracy, and F1-score. Macro-averaging computes each metric independently per class and then averages them, assigning equal importance to all classes. Additionally, micro-accuracy was calculated as the total number of correct predictions divided by the total number of samples, such that classes with a larger number of samples contribute more to the overall score.

In addition, agreement metrics including Matthews correlation coefficient (MCC) and Cohen's kappa were evaluated. The extended MCC for multi-classification was computed using the confusion matrix formulation proposed by Gorodkin [47], [48]:

$$MCC = \frac{\text{trace}(\mathbf{C}) \cdot S - \sum_k r_k c_k}{\sqrt{(S^2 - \sum_k c_k^2)(S^2 - \sum_k r_k^2)}}, \quad (7)$$

where $\mathbf{C}$ is the confusion matrix, $S = \sum_{i,j} C_{ij}$ is the total number of samples, r_k and c_k denote the row and column sums corresponding to class k , representing the total number of true and predicted samples for that class, respectively, and $\text{trace}(\mathbf{C})$ is the sum of diagonal elements, i.e., the total number of correct predictions across all classes.

Cohen's kappa coefficient, originally introduced in [49], can be used to measure agreement between predicted and true labels while accounting for chance agreement. Cohen's kappa is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad (8)$$

where $p_o = \frac{\text{trace}(\mathbf{C})}{N}$ is the observed agreement, and

$$p_e = \frac{\sum_k r_k c_k}{N^2} \quad (9)$$

is the expected agreement by chance, with N being the total number of samples.

All performance metrics were computed over five independent training runs with different fixed random seeds to ensure reproducibility. The final results are reported as mean $\pm$ standard deviation, providing an estimate of model stability.

III. RESULTS

A. DATA-EXTRACTED STATISTICS AND BASELINE ERRORS

From the training subset, the mean duration of apnea/hypopnea events was calculated as $D_{\text{mean}} = 21.39$ s. This value was subsequently used to convert predicted AB into the AHI metric. Additionally, the median values of the ID testing set were $\text{median}(AB_{ID}) = 11.22\%$ and $\text{median}(AHI_{ID}) = 14.70$ counts/h. While the median values of the OOD testing set were $\text{median}(AB_{OOD}) = 11.16\%$ and $\text{median}(AHI_{OOD}) = 23.50$ counts/h. These median values were used as baseline predictors for calculating the MASE.

Table 3 summarizes the baseline mean absolute errors for AB and AHI estimation obtained when always predicting the median values from the ID and OOD testing data, respectively. These baseline errors were used for normalization in MASE calculation.

TABLE 3. Mean absolute errors (baseline) for ID and OOD testing cases.

Mean absolute errors (baseline)	ID testing	OOD testing
MAE(AB) (%)	8.05	9.35
MAE(AHI) (counts/h)	15.50	18.17

B. MODEL PREDICTION ERRORS

In terms of regression performance, the lowest MAE(AB) in ID testing was achieved by the 2CNN-2LSTM model regardless of the features used (see Figure 9), whereas the lowest OOD MAE(AB) was obtained by the 2CNN-GRU model using PPG pulse wave features and SpO₂, with a value of $6.59 \pm 0.25\%$ ($MASE = 0.70$). For AHI estimation (see Figure 10), the lowest ID MAE(AHI_{AB}) and OOD MAE(AHI_{AB}) were also achieved by the 2CNN-2LSTM model. In contrast, the parallel CNNs architecture produced the largest OOD regression errors, particularly when PPG features were included. Regarding the comparison of estimating AB and AHI (see Figures 9 and 10), it can be observed that in all cases $MASE(AHI_{AB}) < MASE(AB)$, indicating that the relative error of AHI estimation is lower than that of AB.

C. SUBJECT-LEVEL CLASSIFICATION PERFORMANCE

As illustrated in Figure 11, the confusion matrices provide a detailed view of class-wise prediction behavior across different model architectures and predictor combinations for both ID and OOD testing scenarios. In general, the ID results exhibited a more pronounced diagonal structure, indicating more accurate classification, whereas the OOD results showed increased dispersion of predictions,

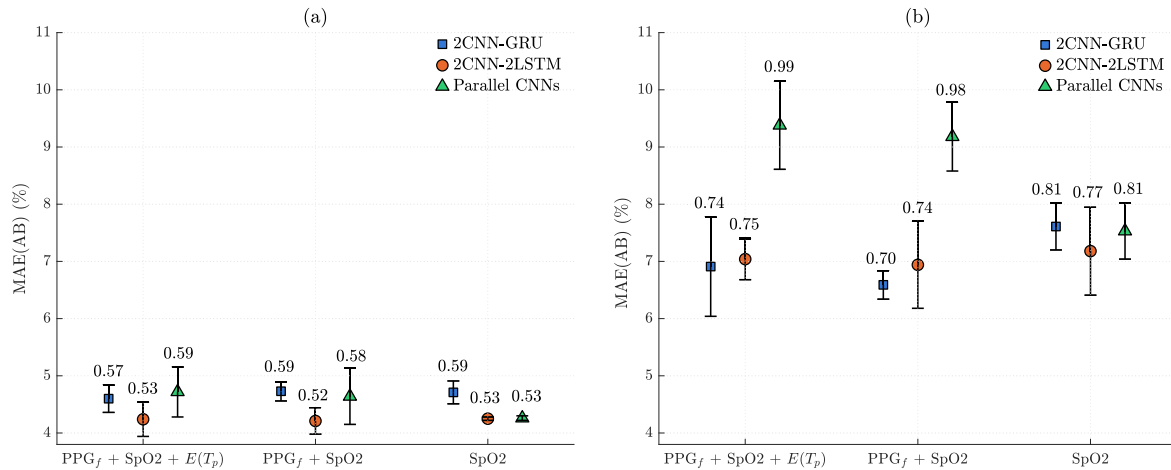


FIGURE 9. Mean absolute errors of predicted AB when using different model architectures and predictor combinations for (a) ID testing and (b) OOD testing data. Error bars (mean $\pm$ standard deviation) obtained from five iterations when different fixed random seeds were used. PPG_f corresponds to five pulse wave features, whereas E(T_p) – the mean envelope feature of the pulse wave interval. Mean absolute scaled errors are shown above error bars.

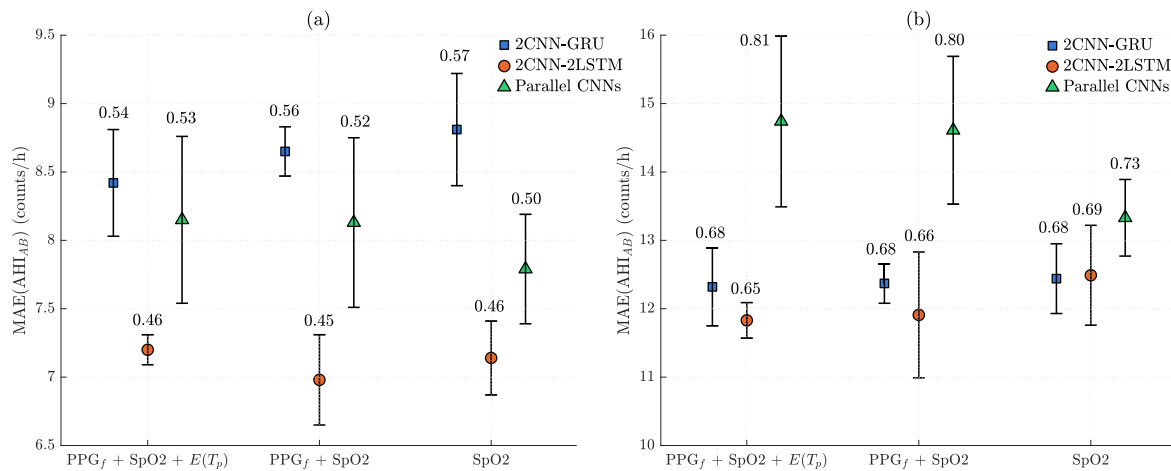


FIGURE 10. Mean absolute errors of predicted AHI when using different model architectures and predictor combinations for (a) ID testing and (b) OOD testing data. Error bars (mean $\pm$ standard deviation) obtained from five iterations when different fixed random seeds were used. PPG_f corresponds to five pulse wave features, whereas E(T_p) – the mean envelope feature of the pulse wave interval. Mean absolute scaled errors are shown above error bars.

particularly between neighboring severity classes. Misclassifications are most frequently observed between mild and moderate groups, reflecting the inherent difficulty of distinguishing adjacent apnea severity levels. Additionally, the inclusion of PPG-derived features led to more concentrated diagonal patterns compared to using SpO₂ alone, suggesting improved class separability.

The subject-level classification performance obtained using different model architectures and predictor combinations is summarized in Table 4. Among the evaluated architectures, the 2CNN-2LSTM model achieved the highest overall performance, particularly in terms of macro-level sensitivity, PPV, and F1-score metrics. This model consistently outperformed the other architectures across both ID and OOD testing scenarios, indicating its stronger capability to capture long-range temporal dependencies in PPG-derived features and SpO₂. Although this highest performance may also be

partially influenced by hyperparameter tuning. For instance, it reached the highest macro-accuracy values of 76.52% and 65.36%, for ID and OOD respectively, when combining PPG pulse wave features, SpO₂, and the proposed feature E(T_p).

The 2CNN-GRU model also benefited from the inclusion of PPG-derived features and showed competitive performance overall. Notably, in OOD testing, the addition of PPG features improved macro-sensitivity from 30.00% to 46.07%, macro-accuracy from 55.15% to 64.52%, and macro-F1-score from 23.63% to 36.70% compared to using SpO₂ alone. This highlights the importance of incorporating PPG-based information, particularly under distribution shifts.

In contrast, the parallel CNNs architecture exhibited different behavior. While it achieved competitive ID performance, particularly when only SpO₂ was used, its overall performance was lower compared to hybrid architectures. Moreover, this model showed the weakest generalization to

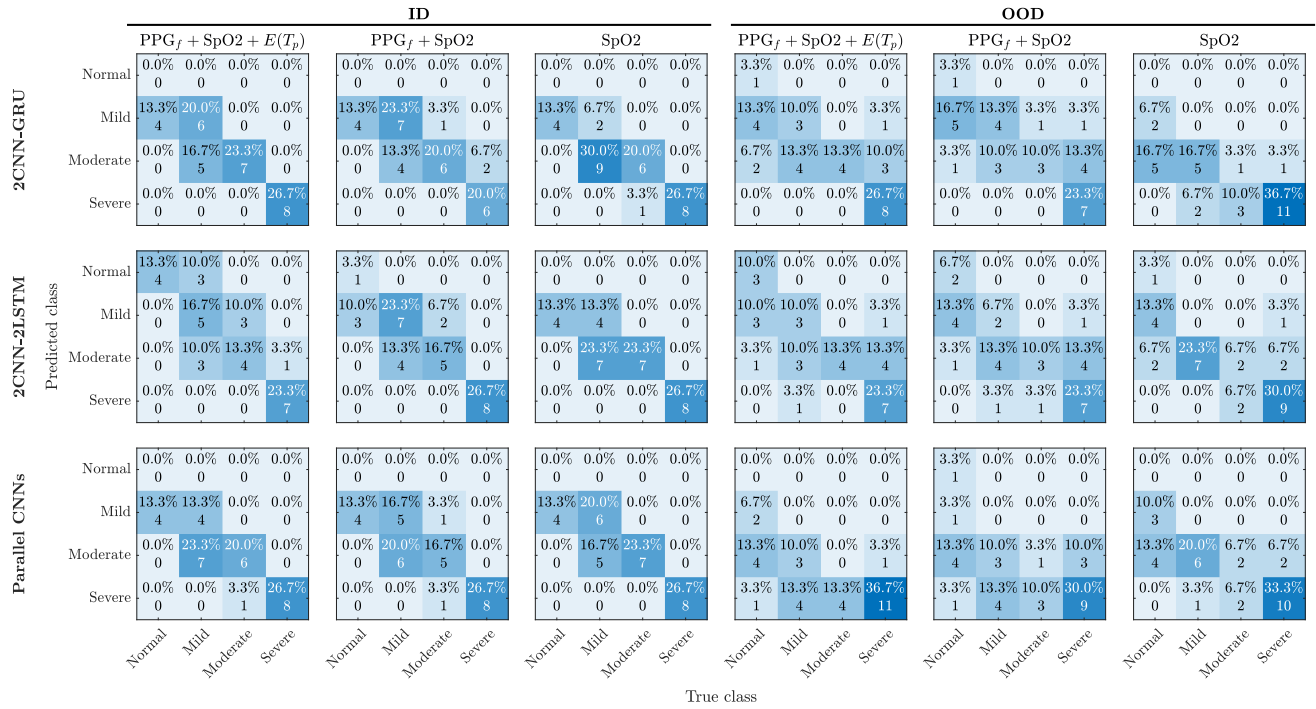


FIGURE 11. Confusion matrices for classifying sleep apnea patients into four severity groups (normal, mild, moderate, and severe) using three model architectures and different predictor combinations. The left panel shows ID, whereas the right panel presents OOD testing results. Each matrix reports class-wise prediction percentages (top) and the corresponding number of subjects (bottom). Predictor combinations include five PPG pulse wave features (PPG_f) with SpO2 and the proposed mean envelope feature $E(T_p)$, PPG_f with SpO2, and SpO2 alone.

OOD data, with a noticeable degradation in sensitivity and F1-score. This suggests that purely convolutional approaches, even when incorporating multi-scale analysis, may be less effective in capturing relations between SpO2 and PPG features compared to hybrid convolutional-recurrent models.

Across all models, combining PPG-derived features with SpO2 consistently led to improved OOD performance compared to using SpO2 alone. The proposed feature $E(T_p)$ provided an additional, albeit modest, benefit, especially for hybrid architectures such as the 2CNN-GRU. These findings demonstrate that PPG-derived features provide complementary physiological information that can be particularly valuable under distribution changes.

A general decrease in performance was observed when transitioning from ID to OOD testing across all models and feature combinations, highlighting the challenges of robust generalization to unseen datasets. Finally, it is worth noting that micro-accuracy values were consistently higher than macro-based metrics, reflecting the class imbalance present in the data and indicating that macro-level evaluation provides a more balanced assessment of model performance across severity groups.

D. AGREEMENT METRICS

The agreement metrics calculated between predicted and true AHI values are shown in Figures 12 and 13. The highest OOD Cohen's kappa was obtained by the 2CNN-2LSTM model using PPG pulse wave features, SpO2, and $E(T_p)$,

reaching $32.59 \pm 7.59\%$, while the highest OOD MCC was achieved by the same model and predictor combination, reaching $31.85 \pm 9.20\%$. The 2CNN-GRU model performed slightly worse, with OOD Cohen's kappa and MCC values of $30.91 \pm 6.36\%$ and $31.34 \pm 10.93\%$, respectively, when PPG pulse wave features, SpO2, and $E(T_p)$ were used. The weakest OOD agreement was observed for the parallel CNNs architecture, when MCC dropped to $15.46 \pm 9.60\%$.

IV. DISCUSSION

This study proposed the novel AB prediction-based framework for characterizing sleep apnea severity and investigated the impact of PPG features on model performance and OOD generalization. The obtained results support the hypothesis that combining PPG-derived features with SpO2 enhances classification performance, particularly under the evaluated OOD scenarios.

A. METHOD NOVELTY AND POTENTIAL ADVANTAGES

The method novelty of this study lies in the use of regression-based model architectures for sleep apnea characterization, in contrast to the majority of existing approaches that rely on direct classification frameworks [19], [26], [27], [28], [29], [30], [50]. Specifically, the proposed models employ a regression output layer to estimate AB, which is subsequently transformed into the clinically relevant AHI metric (see Figure 8 and Equation II-E). This formulation enables a two-stage interpretation: continuous estimation

TABLE 4. Performance metrics obtained for classifying sleep apnea patients into four severity groups using different model architectures and predictor combinations. Averaged results (mean $\pm$ standard deviation) were obtained from five iterations with different fixed random seeds. PPG_f corresponds to five pulse wave features, whereas $E(T_p)$ denotes the mean envelope feature of the pulse wave interval. The highest values for each metric in ID and OOD testing cases within each model are highlighted in bold.

Model	Metric (%)	$PPG_f + SpO_2 + E(T_p)$		$PPG_f + SpO_2$		SpO ₂	
		ID	OOD	ID	OOD	ID	OOD
2CNN-GRU	Macro-Sensitivity	58.97 $\pm$ 4.03	46.07 $\pm$ 11.12	57.46 $\pm$ 1.74	45.00 $\pm$ 5.16	56.56 $\pm$ 4.03	30.00 $\pm$ 1.86
	Macro-Specificity	87.51 $\pm$ 1.69	82.97 $\pm$ 2.54	86.99 $\pm$ 0.78	82.57 $\pm$ 1.26	86.03 $\pm$ 1.73	80.31 $\pm$ 0.63
	Macro-PPV	51.22 $\pm$ 4.47	40.64 $\pm$ 17.55	51.04 $\pm$ 2.26	42.66 $\pm$ 13.27	46.91 $\pm$ 4.66	21.04 $\pm$ 2.02
	Macro-Accuracy	73.24 $\pm$ 2.85	64.52 $\pm$ 6.81	72.23 $\pm$ 1.22	63.79 $\pm$ 3.10	71.29 $\pm$ 2.88	55.15 $\pm$ 1.13
	Macro-F1-score	53.29 $\pm$ 4.36	36.70 $\pm$ 9.89	52.46 $\pm$ 2.10	36.05 $\pm$ 5.83	49.12 $\pm$ 4.74	23.63 $\pm$ 1.29
	Micro-Accuracy	82.33 $\pm$ 2.53	72.67 $\pm$ 3.03	81.67 $\pm$ 1.18	71.33 $\pm$ 2.17	80.00 $\pm$ 2.64	69.33 $\pm$ 1.49
2CNN-2LSTM	Macro-Sensitivity	64.37 $\pm$ 6.78	47.62 $\pm$ 8.45	64.37 $\pm$ 3.76	46.07 $\pm$ 4.28	61.82 $\pm$ 2.96	31.90 $\pm$ 3.56
	Macro-Specificity	88.67 $\pm$ 1.82	83.11 $\pm$ 2.21	88.62 $\pm$ 1.22	83.07 $\pm$ 1.19	88.43 $\pm$ 1.42	80.37 $\pm$ 0.98
	Macro-PPV	63.46 $\pm$ 9.03	55.93 $\pm$ 5.95	63.87 $\pm$ 11.50	57.38 $\pm$ 5.23	52.75 $\pm$ 2.97	38.08 $\pm$ 15.76
	Macro-Accuracy	76.52 $\pm$ 4.13	65.36 $\pm$ 5.32	76.50 $\pm$ 2.31	64.57 $\pm$ 2.72	75.13 $\pm$ 2.19	56.14 $\pm$ 2.24
	Macro-F1-score	61.70 $\pm$ 6.26	44.83 $\pm$ 6.94	62.35 $\pm$ 6.53	43.33 $\pm$ 3.70	55.49 $\pm$ 3.55	28.82 $\pm$ 6.18
	Micro-Accuracy	84.00 $\pm$ 2.53	74.00 $\pm$ 3.03	84.00 $\pm$ 1.90	73.67 $\pm$ 1.39	83.67 $\pm$ 2.17	69.00 $\pm$ 1.49
Parallel CNNs	Macro-Sensitivity	56.19 $\pm$ 1.37	28.69 $\pm$ 6.93	58.44 $\pm$ 5.73	33.69 $\pm$ 5.09	60.91 $\pm$ 2.49	30.42 $\pm$ 3.49
	Macro-Specificity	86.31 $\pm$ 0.59	79.16 $\pm$ 2.06	87.46 $\pm$ 2.33	80.27 $\pm$ 1.56	88.00 $\pm$ 1.19	80.25 $\pm$ 0.78
	Macro-PPV	51.29 $\pm$ 10.40	30.93 $\pm$ 23.72	49.33 $\pm$ 5.65	35.24 $\pm$ 11.56	51.80 $\pm$ 2.65	21.24 $\pm$ 1.59
	Macro-Accuracy	71.25 $\pm$ 0.89	53.92 $\pm$ 4.48	72.95 $\pm$ 4.03	56.98 $\pm$ 3.31	74.45 $\pm$ 1.84	55.33 $\pm$ 2.08
	Macro-F1-score	51.55 $\pm$ 3.50	24.30 $\pm$ 9.23	52.82 $\pm$ 5.84	27.66 $\pm$ 4.29	54.39 $\pm$ 3.11	23.76 $\pm$ 1.77
	Micro-Accuracy	80.67 $\pm$ 0.91	70.00 $\pm$ 2.89	82.33 $\pm$ 3.46	71.33 $\pm$ 2.17	83.00 $\pm$ 1.83	69.00 $\pm$ 1.49

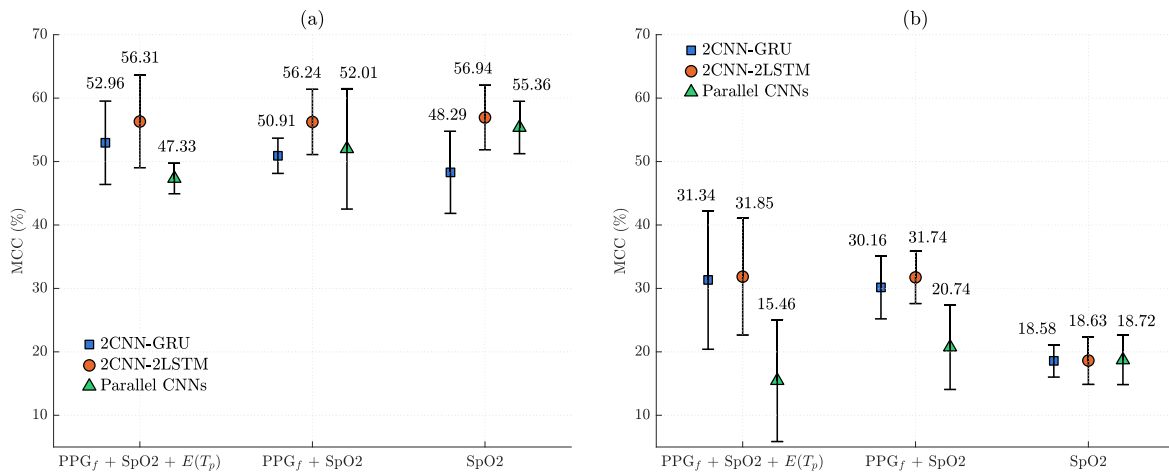


FIGURE 12. Matthews correlation coefficient, MCC, of classifying sleep apnea patients when using different model architectures and predictor combinations for (a) ID testing and (b) OOD testing data. Error bars (mean $\pm$ standard deviation) obtained from five iterations when different fixed random seeds were used. PPG_f corresponds to five pulse wave features, whereas $E(T_p)$ – the mean envelope feature of the pulse wave interval. Mean values are shown above error bars.

followed by classification-per-subject. An important advantage of this approach is its reduced dependence on pre-defined decision thresholds, which are typically required in classification-based models and may introduce bias or sensitivity to class imbalance. Furthermore, predicting a continuous physiological quantity provides a more flexible and clinically meaningful representation of disease severity, enabling post hoc categorization according to established

AHI thresholds. As a result, the proposed framework better aligns with current clinical practice, where subject-level assessment of sleep apnea is essential for diagnosis, monitoring, and treatment planning.

From a clinical perspective, the proposed framework could be integrated into sleep apnea screening workflows as a decision-support tool for subject-level severity assessment based on overnight PPG and SpO₂ recordings. Owing to

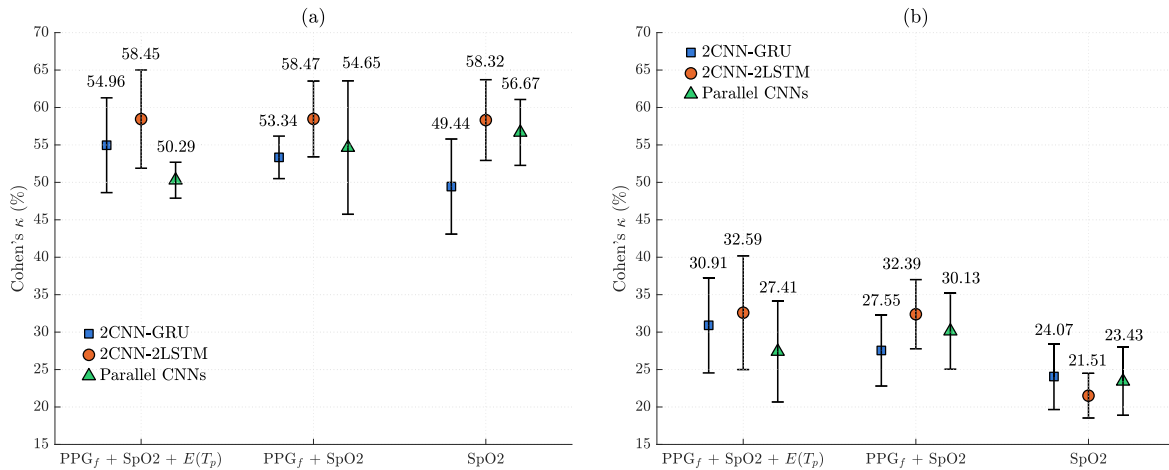


FIGURE 13. Cohen's kappa of classifying sleep apnea patients when using different model architectures and predictor combinations for (a) ID testing and (b) OOD testing data. Error bars (mean $\pm$ standard deviation) obtained from five iterations when different fixed random seeds were used. PPG_f corresponds to five pulse wave features, whereas E(T_p) – the mean envelope feature of the pulse wave interval. Mean values are shown above error bars.

the non-invasive nature and widespread availability of these signals in wearable devices, it could facilitate longitudinal monitoring of disease severity and treatment response, for example during continuous positive airway pressure therapy. Furthermore, because the proposed framework relies only on PPG and SpO₂ analysis rather than full PSG, it may provide a more accessible and cost-effective solution for long-term follow-up in both clinical and home settings.

Although the primary focus of this study is on OOD generalization, the obtained ID performance results are also of significant practical importance. High ID accuracy indicates that the proposed models are capable of capturing subject group-specific physiological patterns, which is particularly relevant in personalized or patient-adaptive systems as proposed in [51]. In such scenarios, periodic model recalibration using newly acquired patient data could further improve performance over time. This adaptive approach may reduce inter-subject variability and enhance prediction reliability in longitudinal monitoring settings.

B. MODEL ARCHITECTURES AND FEATURES

Within the evaluated experimental setting, the 2CNN-2LSTM architecture achieved the most favorable balance between classification performance and OOD robustness. A comparison of predictor combinations showed that adding PPG-derived features generally can improve OOD generalization for low-complexity hybrid architectures. In the case of both 2CNN-GRU and 2CNN-2LSTM, PPG-included predictor combinations outperformed SpO₂ alone across OOD performance and agreement metrics. In contrast, for the parallel CNNs model, the best ID results were obtained using only SpO₂.

Since each architecture was individually optimized, the reported comparison should be interpreted as exploratory rather than as a definitive benchmark of architectural superiority. The primary objective of this study was to

evaluate the proposed AB-guided framework and the contribution of PPG-derived features, rather than to identify the optimal neural network architecture. The most substantial OOD generalization improvements were observed for the hybrid convolutional-recurrent architectures, particularly the 2CNN-2LSTM model, which achieved the highest gains in macro-F1-score (+16.01%), Cohen's kappa (+11.08%), and MCC (+13.22%). Similarly, the 2CNN-GRU model exhibited the strongest improvement in macro-sensitivity (+16.07%) and macro-PPV (+19.60%), indicating enhanced detection capability for apnea severity classes under distribution shift. While the parallel CNNs architecture showed notably smaller improvements across most metrics (the largest increases were +14.00% and +6.70% in PPV and Cohen's kappa, respectively), suggesting a lower benefit from additional PPG features in this multi-scale convolutional-only design. Therefore, these findings suggest that, within the scope of the evaluated models, datasets, and training setup, PPG features tend to provide the greatest benefit when combined with low-complexity hybrid architectures, leading to improved robustness in OOD scenarios.

Regarding the ID testing, the contribution of PPG features was relatively modest, which is consistent with findings reported in [52], where accuracy of detecting apnea events reached 68.70%, 82.20%, and 83.40% when using PPG features, SpO₂, and their combined analysis, respectively.

From a practical perspective, our study supports using 2CNN-2LSTM and 2CNN-GRU models for sleep apnea characterization. In addition to the conventional SpO₂ predictor, the inclusion of PPG-derived biomarkers appears valuable, especially when robustness to distribution shifts is required in real-world applications.

C. LIMITATIONS

Despite the promising results, several limitations should be acknowledged. The evaluation was conducted on a

limited number of datasets, which may not fully represent real-world clinical variability. A noticeable performance gap between ID and OOD results remains, indicating sensitivity to distribution shifts. In some cases including the proposed feature $E(T_p)$ exhibited increased variability across training runs, suggesting limited performance stability.

Although the external OSASUD dataset provided an opportunity to evaluate the proposed framework under a clinically meaningful domain shift, it includes only 30 subjects. While this cohort differs substantially from the training population in terms of clinical status, acquisition conditions, and recording equipment (see Table 1), its limited size may affect the generalizability of the OOD results. Therefore, the presented OOD performance should be interpreted as preliminary evidence supporting the findings of the proposed framework to an independent external cohort.

Another important limitation is related to the estimation of the AHI. The predicted AHI was derived from the predicted AB using a fixed mean event duration D_{mean} , calculated from the training data. However, in real-world conditions, the duration of apnea/hypopnea events is highly variable and can range approximately from 10 seconds to up to 2 minutes. Therefore, the use of a fixed average duration may introduce estimation bias and limit the accuracy of the AHI, particularly across different populations and sleep conditions. Due to potential conversion errors (see Section II-E and Figure 8), borderline severity cases should be examined more thoroughly. Nevertheless, this approach was adopted as a practical simplification to enable a consistent transformation from AB to a clinically interpretable metric within the proposed framework.

D. FUTURE WORK

Further studies should evaluate the proposed framework using larger and more diverse external datasets to strengthen the evidence for its generalizability. Future work will also focus on improving model robustness and interpretability, including deeper analysis of feature contributions and behavior. For instance, additional PPG-related biomarkers could be investigated, as we did with the derivative-based feature, S_{max} , to further enhance model performance.

Moreover, the combination of PPG features and SpO2 could be explored using alternative deep learning architectures, such as transformer-based models, which have recently shown strong potential for modeling long-range dependencies in physiological signals [23], [24], [25], [26]. Since the proposed AB-guided framework is model-agnostic and independent of the underlying learning paradigm, it can also be integrated with self-supervised and foundation-model approaches. Although the evaluation of these approaches was beyond the scope of the present study, they represent promising directions for future research.

In addition, advanced techniques such as domain adaptation [53] and transfer learning could be explored to improve generalization under distribution variance.

Another promising direction is the development of adaptive, patient-specific models with periodic recalibration.

Future research could investigate the determination of clinically meaningful AB cut-off thresholds for direct classification into severity categories. Such an approach would eliminate the need for transformation from AB to AHI and could simplify the overall framework while preserving clinical interpretability. In addition, future studies should investigate patient-specific AB-to-AHI conversion strategies and perform dedicated sensitivity or error propagation analyses to further quantify the contribution of the conversion model to the overall prediction error.

V. CONCLUSION

We hypothesized that combining PPG pulse wave features with SpO2 in artificial neural networks would improve sleep apnea severity classification and enhance OOD generalization. Our results support this hypothesis: fusing PPG-derived features with SpO2 and employing hybrid convolutional-recurrent architectures improved classification performance, particularly under the evaluated distribution shifts. Among the implemented models, the 2CNN-2LSTM architecture achieved the highest overall performance.

The findings highlight the importance of jointly optimizing model architecture and feature representation for robust and generalizable sleep apnea assessment. The new proposed approach provides a clinically relevant AB-guided framework for subject-level severity estimation and has strong potential for deployment in non-invasive, real-world, and home-based sleep monitoring applications.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this manuscript.

DATA AVAILABILITY

The MATLAB scripts for preparing and partitioning the MESA and OSASUD datasets are available online.

REFERENCES

- [1] V. K. Somers, D. P. White, R. Amin, W. T. Abraham, F. Costa, A. Culebras, S. Daniels, J. S. Floras, C. E. Hunt, L. J. Olson, T. G. Pickering, R. Russell, M. Woo, and T. Young, "Sleep apnea and cardiovascular disease," *J. Amer. College Cardiol.*, vol. 52, no. 8, pp. 686–717, Aug. 2008, doi: 10.1016/j.jacc.2008.05.002.
- [2] A. Kadish and J. Jacobson, "Sleep patterns and arrhythmias," *J. Amer. College Cardiol.*, vol. 78, no. 12, pp. 1208–1209, Sep. 2021, doi: 10.1016/j.jacc.2021.07.024.
- [3] X. Li, T. Zhou, H. Ma, T. Huang, X. Gao, J. E. Manson, and L. Qi, "Healthy sleep patterns and risk of incident arrhythmias," *J. Amer. College Cardiol.*, vol. 78, no. 12, pp. 1197–1207, Sep. 2021, doi: 10.1016/j.jacc.2021.07.023.
- [4] A. Malhotra and D. P. White, "Obstructive sleep apnoea," *Lancet*, vol. 360, no. 9328, pp. 237–245, Jul. 2002, doi: 10.1016/s0140-6736(02)09464-3.
- [5] T. Young, P. E. Peppard, and D. J. Gottlieb, "Epidemiology of obstructive sleep apnea: A population health perspective," *Amer. J. Respiratory Crit. Care Med.*, vol. 165, no. 9, pp. 1217–1239, May 2002, doi: 10.1164/rccm.2109080.

- [6] N. R. C. Wilson, O. J. Veatch, and S. M. Johnson, "On the relationship between diabetes and obstructive sleep apnea: Evolution and epigenetics," *Biomedicines*, vol. 10, no. 3, p. 668, Mar. 2022, doi: [10.3390/biomedicines10030668](https://doi.org/10.3390/biomedicines10030668).
- [7] R. G. Smallwood, M. V. Vitiello, E. C. Giblin, and P. N. Prinz, "Sleep apnea: Relationship to age, sex, and Alzheimer's dementia," *Sleep*, vol. 6, no. 1, pp. 16–22, Sep. 1983, doi: [10.1093/sleep/6.1.16](https://doi.org/10.1093/sleep/6.1.16).
- [8] American Academy of Sleep Medicine, "Sleep-related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research," *Sleep*, vol. 22, no. 5, pp. 667–689, Aug. 1999, doi: [10.1093/sleep/22.5.667](https://doi.org/10.1093/sleep/22.5.667).
- [9] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Meas.*, vol. 28, no. 3, pp. R1–R39, Feb. 2007, doi: [10.1088/0967-3334/28/3/r01](https://doi.org/10.1088/0967-3334/28/3/r01).
- [10] M. Elgendy, "On the analysis of fingertip photoplethysmogram signals," *Current Cardiol. Rev.*, vol. 8, no. 1, pp. 14–25, Jun. 2012, doi: [10.2174/157340312801215782](https://doi.org/10.2174/157340312801215782).
- [11] A. Romem, A. Romem, D. Koldobskiy, and S. M. Scharf, "Diagnosis of obstructive sleep apnea using pulse oximeter derived photoplethysmographic signals," *J. Clin. Sleep Med.*, vol. 10, no. 3, pp. 285–290, Mar. 2014, doi: [10.5664/jcsm.3530](https://doi.org/10.5664/jcsm.3530).
- [12] J. Lázaro, E. Gil, J. M. Vergara, and P. Laguna, "Pulse rate variability analysis for discrimination of sleep-apnea-related decreases in the amplitude fluctuations of pulse photoplethysmographic signal in children," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 1, pp. 240–246, Jan. 2014, doi: [10.1109/JBHI.2013.2267096](https://doi.org/10.1109/JBHI.2013.2267096).
- [13] R. Lazazzera, M. Deviaene, C. Varon, B. Buyse, D. Testelmans, P. Laguna, E. Gil, and G. Carrault, "Detection and classification of sleep apnea and hypopnea using PPG and SpO₂ signals," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 5, pp. 1496–1506, May 2021, doi: [10.1109/TBME.2020.3028041](https://doi.org/10.1109/TBME.2020.3028041).
- [14] E. Gil, J. María Vergara, and P. Laguna, "Detection of decreases in the amplitude fluctuation of pulse photoplethysmography signal as indication of obstructive sleep apnea syndrome in children," *Biomed. Signal Process. Control*, vol. 3, no. 3, pp. 267–277, Jul. 2008, doi: [10.1016/j.bspc.2007.12.002](https://doi.org/10.1016/j.bspc.2007.12.002).
- [15] M. Á. Goda, A. Oksenberg, A. Azarbarzin, and J. A. Behar, "Analysis of differential photoplethysmography signal patterns in apnea and hypopnea," *Physiol. Meas.*, vol. 47, no. 2, Feb. 2026, Art. no. 025006, doi: [10.1088/1361-6579/ae3ef0](https://doi.org/10.1088/1361-6579/ae3ef0).
- [16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: [10.1038/nature14539](https://doi.org/10.1038/nature14539).
- [17] O. Faust, Y. Hagiwara, T. J. Hong, O. S. Lih, and U. R. Acharya, "Deep learning for healthcare applications based on physiological signals: A review," *Comput. Methods Programs Biomed.*, vol. 161, pp. 1–13, Jul. 2018, doi: [10.1016/j.cmpb.2018.04.005](https://doi.org/10.1016/j.cmpb.2018.04.005).
- [18] S. S. Mostafa, F. Mendonça, A. G. Ravelo-García, and F. Morgado-Dias, "A systematic review of detecting sleep apnea using deep learning," *Sensors*, vol. 19, no. 22, p. 4934, Nov. 2019, doi: [10.3390/s19224934](https://doi.org/10.3390/s19224934).
- [19] P. Sharma, A. Jalali, M. Majmudar, K. S. Rajput, and N. Selvaraj, "Deep-learning based sleep apnea detection using SpO₂ and pulse rate," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2022, pp. 2611–2614, doi: [10.1109/EMBC48229.2022.9871295](https://doi.org/10.1109/EMBC48229.2022.9871295).
- [20] C. Singthong and T. Siribornvornratanakul, "Deep-learning based sleep apnea detection using sleep sound, SpO₂, and pulse rate," *Int. J. Inf. Technol.*, vol. 16, no. 8, pp. 4869–4874, May 2024, doi: [10.1007/s41870-024-01906-x](https://doi.org/10.1007/s41870-024-01906-x).
- [21] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, Nov. 2017, doi: [10.1109/TNSRE.2017.2721116](https://doi.org/10.1109/TNSRE.2017.2721116).
- [22] G. Srivastava, A. Chauhan, N. Kargeti, N. Pradhan, and V. S. Dhaka, "ApneaNet: A hybrid 1DCNN-LSTM architecture for detection of obstructive sleep apnea using digitized ECG signals," *Biomed. Signal Process. Control*, vol. 84, Jul. 2023, Art. no. 104754, doi: [10.1016/j.bspc.2023.104754](https://doi.org/10.1016/j.bspc.2023.104754).
- [23] M. A. Almarshad, S. Al-Ahmadi, S. Islam, A. Soudani, and A. S. BaHammam, "Transformer-based deep learning approach for obstructive sleep apnea detection using single-lead ECG," *Front. Artif. Intell.*, vol. 9, pp. 1–14, Feb. 2026, doi: [10.3389/frai.2026.1727091](https://doi.org/10.3389/frai.2026.1727091).
- [24] C. Wan, M. C. Nnamdi, W. Shi, B. Smith, C. Purnell, and M. D. Wang, "Advancing sleep disorder diagnostics: A transformer-based EEG model for sleep stage classification and OSA prediction," *IEEE J. Biomed. Health Informat.*, vol. 29, no. 2, pp. 878–886, Feb. 2025, doi: [10.1109/JBHI.2024.3512616](https://doi.org/10.1109/JBHI.2024.3512616).
- [25] Y. Zhang, L. Zhou, S. Zhu, Y. Zhou, Z. Wang, L. Ma, Y. Yuan, Y. Xie, X. Niu, Y. Su, H. Liu, X. Hei, Z. Shi, X. Ren, and Y. Shi, "Deep learning for obstructive sleep apnea detection and severity assessment: A multimodal signals fusion multiscale transformer model," *Nature Sci. Sleep*, vol. 17, pp. 1–15, Jan. 2025, doi: [10.2147/nss.s492806](https://doi.org/10.2147/nss.s492806).
- [26] M. A. Almarshad, S. Al-Ahmadi, M. S. Islam, A. S. BaHammam, and A. Soudani, "Adoption of transformer neural network to improve the diagnostic performance of oximetry for obstructive sleep apnea," *Sensors*, vol. 23, no. 18, p. 7924, Sep. 2023, doi: [10.3390/s23187924](https://doi.org/10.3390/s23187924).
- [27] A. John, K. K. Nundy, B. Cardiff, and D. John, "SomnNET: An SpO₂ based deep learning network for sleep apnea detection in smartwatches," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 1961–1964, doi: [10.1109/EMBC46164.2021.9631037](https://doi.org/10.1109/EMBC46164.2021.9631037).
- [28] S. S. Mostafa, F. Mendonça, F. Morgado-Dias, and A. Ravelo-García, "SpO₂ based sleep apnea detection using deep learning," in *Proc. IEEE 21st Int. Conf. Intell. Eng. Syst. (INES)*, Oct. 2017, pp. 91–96, doi: [10.1109/INES.2017.8118534](https://doi.org/10.1109/INES.2017.8118534).
- [29] E. Mortazavi, B. Tarviridzadeh, K. Alipour, and M. Ghamari, "Deep learning approaches for assessing pediatric sleep apnea severity through SpO₂ signals," *Sci. Rep.*, vol. 14, no. 1, pp. 1–21, Oct. 2024, doi: [10.1038/s41598-024-67729-9](https://doi.org/10.1038/s41598-024-67729-9).
- [30] N. H. Hoang and Z. Liang, "AI-driven sleep apnea screening with overnight blood oxygen saturation: Current practices and future directions," *Frontiers Digit. Health*, vol. 7, pp. 1–10, Apr. 2025, doi: [10.3389/fdgth.2025.1510166](https://doi.org/10.3389/fdgth.2025.1510166).
- [31] J.-U. Park, H.-K. Lee, J. Lee, E. Urtnasan, H. Kim, and K.-J. Lee, "Automatic classification of apnea/hypopnea events through sleep/wake states and severity of SDB from a pulse oximeter," *Physiol. Meas.*, vol. 36, no. 9, pp. 2009–2025, Aug. 2015, doi: [10.1088/0967-3334/36/9/2009](https://doi.org/10.1088/0967-3334/36/9/2009).
- [32] T. Choksatchawathi, G. Sawadwuthikul, P. Thuwajit, T. Kaewlee, T. Mateepithaktham, S. Saisaard, T. Sudhawiyanukul, B. Chaitusaney, W. Saengmolee, and T. Wilaiprasitporn, "ApSense: Data-driven algorithm in PPG-based sleep apnea sensing," *IEEE Internet Things J.*, vol. 11, no. 20, pp. 33915–33926, Oct. 2024, doi: [10.1109/JIOT.2024.3433500](https://doi.org/10.1109/JIOT.2024.3433500).
- [33] S. Attia, R. S. Hershkovich, A. Tabakhov, A. Ang, A. Oksenberg, R. Tauman, and J. A. Behar, "SleepPPG-Net2: Deep learning generalization for sleep staging from photoplethysmography," *Physiol. Meas.*, vol. 46, no. 12, Dec. 2025, Art. no. 125001, doi: [10.1088/1361-6579/ae1a34](https://doi.org/10.1088/1361-6579/ae1a34).
- [34] X. Chen, R. Wang, P. Zee, P. L. Lutsey, S. Javaheri, C. Alcántara, C. L. Jackson, M. A. Williams, and S. Redline, "Racial/Ethnic differences in sleep disturbances: The multi-ethnic study of atherosclerosis (MESA)," *SLEEP*, pp. 877–888, Jun. 2015, doi: [10.5665/sleep.4732](https://doi.org/10.5665/sleep.4732).
- [35] Á. S. Alarcón, N. M. Madrid, R. Seepold, and J. A. Ortega, "Obstructive sleep apnea event detection using explainable deep learning models for a portable monitor," *Frontiers Neurosci.*, vol. 17, pp. 1–19, Jul. 2023, doi: [10.3389/fnins.2023.1155900](https://doi.org/10.3389/fnins.2023.1155900).
- [36] M. Rinkevicius, J. Lázaro, E. Gil, P. Laguna, P. H. Charlton, R. Bailón, and V. Marozas, "Obstructive sleep apnea characterization: A multimodal cross-recurrence-based approach for investigating atrial fibrillation," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 10, pp. 6155–6167, Oct. 2024, doi: [10.1109/JBHI.2024.3428845](https://doi.org/10.1109/JBHI.2024.3428845).
- [37] M. Rinkevicius, P. H. Charlton, R. Bailón, and V. Marozas, "Influence of photoplethysmogram signal quality on pulse arrival time during polysomnography," *Sensors*, vol. 23, no. 4, p. 2220, Feb. 2023, doi: [10.3390/s23042220](https://doi.org/10.3390/s23042220).
- [38] A. Bernardini, A. Brunello, G. L. Gigli, A. Montanari, and N. Saccomanno, "OSASUD: A dataset of stroke unit recordings for the detection of obstructive sleep apnea syndrome," *Scientific Data*, vol. 9, no. 1, pp. 1–10, Apr. 2022, doi: [10.1038/s41597-022-01272-y](https://doi.org/10.1038/s41597-022-01272-y).
- [39] A. Bernardini, A. Brunello, G. L. Gigli, A. Montanari, and N. Saccomanno, "AIOA: An approach to the automatic identification of obstructive sleep apnea events based on deep learning," *Artif. Intell. Med.*, vol. 118, Aug. 2021, Art. no. 102133, doi: [10.1016/j.artmed.2021.102133](https://doi.org/10.1016/j.artmed.2021.102133).
- [40] Y. Ji, D. Chen, Y. Zuo, T. Gao, and Y. Tang, "Accurate apnea and hypopnea localization in PSG with multi-scale object detection via dual-modal feature learning," *Biomed. Signal Process. Control*, vol. 89, Mar. 2024, Art. no. 105717, doi: [10.1016/j.bspc.2023.105717](https://doi.org/10.1016/j.bspc.2023.105717).

- [41] P. Kulkarni, "Obstructive sleep apnea syndrome identification using CNN-LSTM hybrid model," *J. Electr. Syst.*, vol. 20, no. 2, pp. 2386–2394, Apr. 2024, doi: [10.52783/jes.2013](https://doi.org/10.52783/jes.2013).
- [42] P. Reali, R. Lolatto, S. Coelli, G. Tartaglia, and A. M. Bianchi, "Information retrieval from photoplethysmographic sensors: A comprehensive comparison of practical interpolation and breath-extraction techniques at different sampling rates," *Sensors*, vol. 22, no. 4, p. 1428, Feb. 2022, doi: [10.3390/s22041428](https://doi.org/10.3390/s22041428).
- [43] M. Aboy, J. McNamara, T. Thong, D. Tsunami, M. S. Ellenby, and B. Goldstein, "An automatic beat detection algorithm for pressure signals," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 10, pp. 1662–1670, Oct. 2005, doi: [10.1109/TBME.2005.855725](https://doi.org/10.1109/TBME.2005.855725).
- [44] M. Panwar, A. Gautam, D. Biswas, and A. Acharyya, "PP-Net: A deep learning framework for PPG-based blood pressure and heart rate estimation," *IEEE Sensors J.*, vol. 20, no. 17, pp. 10000–10011, Sep. 2020, doi: [10.1109/JSEN.2020.2990864](https://doi.org/10.1109/JSEN.2020.2990864).
- [45] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9, doi: [10.1109/CVPR.2015.7298594](https://doi.org/10.1109/CVPR.2015.7298594).
- [46] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: An overview," 2020, *arXiv:2008.05756*.
- [47] J. Gorodkin, "Comparing two K-category assignments by a K-category correlation coefficient," *Comput. Biol. Chem.*, vol. 28, nos. 5–6, pp. 367–374, Dec. 2004, doi: [10.1016/j.compbiolchem.2004.09.006](https://doi.org/10.1016/j.compbiolchem.2004.09.006).
- [48] G. Jurman, S. Riccadonna, and C. Furlanello, "A comparison of MCC and CEN error measures in multi-class prediction," *PLoS ONE*, vol. 7, no. 8, Aug. 2012, Art. no. e41882, doi: [10.1371/journal.pone.0041882](https://doi.org/10.1371/journal.pone.0041882).
- [49] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: [10.1177/001316446002000104](https://doi.org/10.1177/001316446002000104).
- [50] K. Wei, L. Zou, G. Liu, and C. Wang, "MS-Net: Sleep apnea detection in PPG using multi-scale block and shadow module one-dimensional convolutional neural network," *Comput. Biol. Med.*, vol. 155, Mar. 2023, Art. no. 106469, doi: [10.1016/j.combiomed.2022.106469](https://doi.org/10.1016/j.combiomed.2022.106469).
- [51] R. Mukkamala, G. S. Stergiou, and A. P. Avolio, "Cuffless blood pressure measurement," *Annu. Rev. Biomed. Eng.*, vol. 24, no. 1, pp. 203–230, Jun. 2022, doi: [10.1146/annurev-bioeng-110220-014644](https://doi.org/10.1146/annurev-bioeng-110220-014644).
- [52] M. Deviaene, J. Lázaro, D. Huysmans, D. Testelmans, B. Buyse, S. Van Huffel, and C. Varon, "Sleep apnea detection using pulse photoplethysmography," in *Proc. Comput. Cardiol. (CinC)*, Maastricht, The Netherlands, 2018, pp. 1–4, doi: [10.22489/CinC.2018.134](https://doi.org/10.22489/CinC.2018.134).
- [53] M. Moulaeifard, P. H. Charlton, and N. Strodthoff, "Generalizable deep learning for photoplethysmography-based blood pressure estimation—A benchmarking study," *Mach. Learn., Health*, vol. 1, no. 1, Sep. 2025, Art. no. 010501, doi: [10.1088/3049-477x/ae01a8](https://doi.org/10.1088/3049-477x/ae01a8).



MANTAS RINKEVIČIUS received the B.Sc. and M.Sc. degrees in biomedical engineering, and the Ph.D. degree in electrical and electronics engineering from Kaunas University of Technology (KTU), Lithuania, in 2019, 2021, and 2025, respectively.

From 2019 to 2023, he was a Project Engineer with the Biomedical Engineering Institute, KTU, and UAB Gruppo FOS Lithuania. From 2023 to 2025, he was a Junior Researcher with the Biomedical Engineering Institute, KTU.

Since 2026, he has been a Scientific Researcher with the Biomedical Engineering Institute, KTU. His research interests include sleep apnea detection and characterization, photoplethysmogram signal quality analysis, cuffless blood pressure estimation, autonomic nervous system assessment, stress monitoring, heart rate variability and arrhythmia analysis, biomedical signal processing and feature extraction, and machine learning and deep learning methods.

Dr. Rinkevičius received the KTU Talent Scholarship, from 2017 to 2021, the Engineer Vytautas Naginevičius Scholarship, in 2018, the President of the Republic of Lithuania Jonas Žemaitis Scholarship, in 2020 and 2021, the Award for the Most Active KTU Ph.D. Student, in 2024, and the Research Council of Lithuania Scholarship for Ph.D. Studies, in 2025.



quantum algorithms for partial differential equations and tensor networks.

OSKAR PFEFFER received the M.Sc. degree in theoretical physics from the Technical University of Berlin, in 2021. He is currently pursuing the Ph.D. degree with Physikalisch-Technische Bundesanstalt, Berlin. From 2019 to 2021, he was a Student Assistant with the Potsdam Institute for Climate Impact Research, focusing on complex energy networks. His current research interests include uncertainty quantification in machine learning models for medical applications, and



AMAL ALISSA received the B.Sc. degree in mathematics from Lebanese University, Lebanon, in 2021. She is currently pursuing the M.Sc. degree in mathematics with Freie Universität Berlin, Germany. Alongside her studies, she is a Research Assistant with the Department for Mathematical Modeling and Data Analysis, Physikalisch-Technische Bundesanstalt. Her current interests include machine learning and its applications in healthcare.



has focused on uncertainty quantification of machine learning and deep learning for optical and medical applications.

NANDO HEGEMANN received the M.Sc. degree in mathematics from Humboldt Universität zu Berlin, in 2017, and the Ph.D. degree in mathematics from Technical University Berlin, in 2022. Since 2021, he has been a Senior Research Scientist with the Department for Mathematical Modeling and Data Analysis, Physikalisch-Technische Bundesanstalt. He has expertise in mathematical and computational modeling of inverse problems and uncertainty quantification. Recently, his work



VAIDOTAS MAROZAS (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electronics engineering from Kaunas University of Technology (KTU), Lithuania, in 1993, 1995, and 2000, respectively.

He is currently a Professor with the Department of Electronics Engineering and the Director of the Biomedical Engineering Institute, KTU. He has led and participated in numerous national and international research projects, including those funded by Horizon Europe, FP7, H2020, COST Action, and the Swedish Institute Visby Program, and has coordinated multiple research and development projects with industrial partners. He has authored and co-authored numerous scientific publications in leading journals, including *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING* and *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS*, and is a co-author of monographs published by Springer and Academic Press. His research interests include biomedical signal processing, physiological modeling, digital biomarkers, wearable systems for health monitoring, cardiovascular disease assessment, and machine learning methods for biomedical applications.

Dr. Marozas has supervised seven Ph.D. students and has received several awards, including KTU Researcher of the Year, in 2016, and KTU Project Manager of the Year, in 2025.

...