



OPEN Deep learning approaches for assessing pediatric sleep apnea severity through SpO2 signals

Erfan Mortazavi¹, Bahram Tarvirdizadeh^{1✉}, Khalil Alipour¹ & Mohammad Ghamari²

Pediatric Sleep Apnea–Hypopnea (SAH) presents a significant health challenge, particularly in diagnostic contexts, where conventional Polysomnography (PSG) testing, although effective, can be distressing for children. Addressing this, our research proposes a less invasive method to assess pediatric SAH severity by analyzing blood oxygen saturation (SpO2) signals. We adopted two advanced deep learning architectures, namely ResNet-based and attention-augmented hybrid CNN-BiGRU models, to process SpO2 signals in a one-dimensional (1D) format for Apnea–Hypopnea Index (AHI) estimation in pediatric subjects. Employing the CHAT dataset, which includes 844 SpO2 signals, the data was partitioned into training (60%), testing (30%), and validation (10%) sets. A predefined validation subset was randomly selected to ensure the models' robustness via a threefold cross-validation approach. Comparative analysis revealed that while the ResNet model attained an average accuracy of 72.9% across four SAH severity categories with a kappa score of 0.57, the CNN-BiGRU-Attention model demonstrated superior performance, achieving an average accuracy of 75.95% and a kappa score of 0.63. This distinction underscores our method's efficacy in both estimating AHI and categorizing SAH severity levels with notable precision. Further, to evaluate diagnostic capabilities, the models were benchmarked against common AHI thresholds (1, 5, and 10 events/hour) in each test fold, affirming their effectiveness in identifying pediatric SAH. This study marks a significant advance in the field, offering a non-invasive, child-friendly alternative for pediatric SAH diagnosis. Although challenges persist in accurately estimating AHI, particularly in severe cases, our findings represent a critical stride towards improving diagnostic processes in pediatric SAH.

Keywords Pediatric sleep apnea–hypopnea (SAH), Apnea–hypopnea index (AHI), Blood oxygen saturation (SpO2), Convolutional neural networks (CNN), Residual networks (ResNet), Bidirectional gated recurrent units (BiGRU), Attention mechanisms

Pediatric sleep apnea–hypopnea (SAH) is a significant sleep disorder observed in children, characterized by disruptions in breathing patterns during sleep. These disruptions can manifest as reduced airflow (hypopnea) or complete cessation of airflow (apnea), significantly impairing the quality of children's sleep. Such impairments can lead to daytime drowsiness, concentration difficulties, and overall well-being impact¹. Additionally, SAH is linked to cognitive, behavioral, and cardiovascular issues, posing potential long-term health concerns and impeding a child's development¹.

SAH is classified into three main types: obstructive sleep apnea (OSA), central sleep apnea (CSA), and mixed sleep apnea (MSA)². OSA is typically caused by a constricted or blocked airway, making breathing efforts ineffective. In contrast, CSA results from the brain's failure to generate or transmit appropriate signals for breathing initiation, leading to brief pauses in breath. MSA is a condition where both central and obstructive factors contribute to the sleep apnea events².

The conventional diagnostic standard for pediatric SAH is the overnight polysomnography (PSG) test. This test, conducted in specialized sleep laboratories, involves monitoring a range of physiological signals such as the electroencephalogram (EEG), electromyogram (EMG), electrocardiogram (ECG), airflow (AF), chest and abdominal movements, blood oxygen saturation (SpO2), and photoplethysmogram (PPG)^{1,3}. The collected data aids in calculating the apnea–hypopnea index (AHI), a critical clinical metric representing the average number of apnea and hypopnea events per hour of sleep. The AHI is instrumental in evaluating the presence and severity of

¹Advanced Service Robots (ASR) Laboratory, Department of Mechatronics Engineering, School of Intelligent Systems Engineering, College of Interdisciplinary Science and Technology, University of Tehran, Tehran, Iran. ²Department of Electrical Eng., California Polytechnic State University, San Luis Obispo, CA 93407, USA. ✉email: bahram@ut.ac.ir

SAH, with severity categorized into four groups: normal ($AHI < 1$), mild ($1 \leq AHI < 5$), moderate ($5 \leq AHI < 10$), and severe ($AHI \geq 10$)^{4–6}. Despite its effectiveness, PSG is a complex, costly, and time-consuming process, often uncomfortable for children, underscoring the need for simpler, more accessible diagnostic methods⁷.

One such promising diagnostic tool is the measurement of SpO₂ using pulse oximetry. Pulse oximeters record the PPG signal, which is used to derive SpO₂⁸. SpO₂, reflecting the oxygen content in blood hemoglobin, is extensively explored for its convenience in acquisition and compatibility with portable monitoring^{9,14–29}. Oximetry recordings are critical in revealing how apnea and hypopnea events lead to recurrent oxygen desaturation due to compromised airflow, causing irregular fluctuations in SpO₂ signals in individuals with SAH⁹. According to the American Academy of Sleep Medicine (AASM) guidelines, apneas are identified by a decrease of $\geq 90\%$ in the AF signal for at least two respiratory cycles, while hypopneas are defined as a decrease of $\geq 30\%$ in AF, accompanied by at least a 3% reduction in SpO₂ or an electroencephalographic arousal⁴. Given that oxygen desaturation typically begins 20–40 s after the start of an apneic episode, precise correlation between apneic events and subsequent desaturation is crucial for accurate detection¹⁰. Thus, SpO₂ monitoring is invaluable for real-time evaluation of oxygen levels and essential in identifying SAH-related desaturation events.

Numerous studies have focused on feature engineering techniques and Machine Learning (ML) methods to analyze AF and SpO₂ signals for detecting pediatric OSA¹¹. These studies have utilized classical ML models such as logistic regression, support vector machines (SVM), and ensemble-learning adaptive boosting (AdaBoost) for binary classification tasks, distinguishing between OSA-positive and non-OSA patients^{12–14,17}. Additionally, multilayer perceptron (MLP) neural networks have been employed for AHI estimation, with Hornero et al. using an MLP for AHI estimation from SpO₂ recordings, categorizing subjects into four severity classes of OSA¹⁵. Barroso-García et al. explored AF and SpO₂ recordings for AHI estimation, using recurrence plots (RP) and the 3% oxygen desaturation index (ODI3) from SpO₂ signals for their MLP model¹⁶. Jiménez-García et al. addressed a 4-class classification task assessing pediatric OSA severity using AdaBoost, utilizing features from both AF and SpO₂ signals¹⁷. These studies underscore the potential of ML in OSA screening.

However, deep learning (DL) algorithms present an advantage over traditional ML methods due to their ability to automatically extract complex features from raw data, thereby enhancing diagnostic accuracy and robustness¹⁸. The surge of deep learning innovations has significantly advanced the biomedical field, particularly in processing physiological signals. This has led to notable achievements in disease detection through DL applications, including blood pressure estimation^{19,20}, sleep stage classification²¹, and cardiovascular risk assessment²², etc. Unlike traditional methods, DL techniques can uncover deeper physiological information and enabling automated integration of a variety of features. Consequently, there is a growing trend among researchers to explore the detection of SAH through DL techniques. Several studies have explored deep DL methods for detecting OSA in adults using PPG and ECG signals. These studies focus on the segment-level classification of signal segments as either apneic or non-apneic, and incorporate architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)^{23–25}, including Long Short-Term Memory (LSTM) networks in their models. All these studies manually determine the AHI to assess OSA severity. These approaches have shown promising results in enhancing OSA detection efficiency, demonstrating that CNNs can effectively extract deep features and RNNs can capture temporal features and measure temporal dependencies from physiological signals like PPG and ECG, thereby improving classification accuracy. In the domain of pediatric SAH detection, Vaquerizo-Villar et al. have implemented a CNN-based model to classify segments of SpO₂ signals for SAH detection²⁶. To assess the severity of pediatric SAH, recent studies have adopted DL methods with CNN-based models for regression tasks to estimate the AHI from SpO₂ signals^{27,28}, occasionally in combination with AF signals^{29,31}. Vaquerizo-Villar et al. have also utilized a one dimensional (1D) CNN model to enhance the diagnostic capabilities of oximetry for pediatric SAH²⁷ and later for pediatric OSA²⁸. Furthermore, another research initiative has introduced a two dimensional (2D) CNN framework for estimating pediatric OSA severity by analyzing both AF and SpO₂ signals as 2D data²⁹. In subsequent work, they combined a 2D-CNN with a RNN to assess pediatric OSA severity from AF and SpO₂ signals³¹. García-Vicente et al. have employed a 1D-CNN model to process overnight electrocardiogram (ECG) signals for pediatric OSA severity estimation³⁰. These studies underscore the effectiveness of CNN-based models in extracting features from SpO₂, AF, and ECG signals for both regression approaches in assessing SAH severity and classification for SAH detection. The insights derived from these studies are crucial for the development of reliable computer-aided diagnostic systems for managing childhood SAH.

In our research, we have employed a regression method for assessing SAH severity solely from SpO₂ signals by directly estimating the AHI. SpO₂ signals, which can be easily recorded with pulse oximeter sensors, have proven effective for SAH severity assessment in most previous studies. Recognizing the gap in the literature regarding the enhancement of CNN-based model architectures for AHI estimation and SAH severity assessment with SpO₂ signals, we explored the integration of residual block architecture and attention-based RNNs to enhance CNN models for AHI estimation and SAH severity assessment using SpO₂ signals, our research introduces an innovative approach. In this study, we propose a novel method for pediatric SAH detection, employing SpO₂ signals as 1D raw data. Our key contributions are as follows:

- **Development of Two Unique Models:** We have pioneered the implementation of a 1D ResNet-based model with residual architecture and an attention-based hybrid CNN-RNN network. These architectures are novel in the context of pediatric SAH assessment, representing the first application of such advanced neural network structures for AHI estimation.
- **Comprehensive Apnea Detection:** Moving beyond the predominant focus on obstructive events in the literature, our models are capable of detecting all oxygen desaturations related to various types of apneas, including OSA, CSA, and MSA. This broadened detection scope is particularly crucial in pediatric cases, where

distinguishing between different apnea types based solely on desaturation patterns is challenging without additional chest and abdominal movement data⁴⁸.

Through these methodologies, our study aims not only to enhance the accuracy of SAH detection but also to contribute to the development of more effective diagnostic tools for pediatric sleep disorders.

Data source and signal analysis

In this research, we utilized the Childhood Adenotonsillectomy Trial (CHAT) dataset, a comprehensive and publicly accessible database that includes 1638 sleep studies of 1232 pediatric subjects aged between 5 and 9.9 years, all diagnosed with mild to moderate obstructive sleep apnea. These studies, conducted between 2007 and 2012, are registered under the Clinical Trial Number NCT00560859^{32,33}. The CHAT dataset, available through the National Sleep Research Resource (<https://sleepdata.org/datasets/chat>), categorizes these studies into three sub-groups: Baseline (453 subjects), Follow-up (406 subjects), and Non-randomized (779 subjects). The participants in the CHAT dataset are divided into randomized and non-randomized groups. The Baseline group consists of subjects who were randomly selected for early Adenotonsillectomy (eAT), while the Follow-up group includes individuals from the Baseline group who were observed over a 7-month period post-intervention. The PSG data within the CHAT dataset provides detailed annotations on the onset and duration of apneic events, which are crucial for labeling SpO2 signal segments in our study. The accurate linking of oxygen desaturation events to apneic episodes is essential for determining the number of apneic events present in each segment, serving as the foundation for training our algorithms. In accordance with the American Academy of Sleep Medicine (AASM) 2012 guidelines, the AHI for this study was calculated considering all apneas and hypopneas that were accompanied by either an arousal or a minimum of 3% oxygen desaturation³⁴. This computation was based on the original variables included in the dataset. Consequently, clinical variables that provide reference information were vital for validating the number of apnea-hypopnea events associated with $\geq 3\%$ oxygen desaturations, as identified by our labeling algorithm. Due to the absence of these critical variables in the non-randomized group, our analysis exclusively utilized recordings from the Baseline and Follow-up groups of the CHAT dataset.

For the training and evaluation of our models, we divided the dataset into distinct sets: 60% for training, 30% for testing, and 10% for validation. To ensure the robustness and generalizability of our DL models, a threefold cross-validation method was implemented. Initially, 10% of the data was reserved as a fixed validation set. The remaining data were then randomly distributed into three groups, with careful consideration of the proportion of each SAH severity category. This strategic partitioning was designed to facilitate a comprehensive and balanced assessment of the models' performance across various SAH severity levels.

Methodology

The methodology for AHI estimation in our study is outlined in Fig. 1, comprising an end-to-end pipeline with four key stages: (A) Signal Segmentation and Pre-processing; (B) Labeling; (C) Deep Learning Model; and (D) AHI Estimation.

Signal segmentation and pre-processing

The SpO2 signals, acquired from PSG using a pulse oximeter finger probe, varied in sampling rates, ranging from 1 to 512 Hz. Our initial step involved re-sampling the SpO2 recordings to a unified rate of 1 Hz, rounded to the nearest second decimal place. This standardization, inspired by prior research^{28,34}, aimed to reduce computational demands and achieve consistency across signals. Following re-sampling, we divided the SpO2 signals from each subject into non-overlapping segments of 20 min (1200 samples). This segmentation strategy facilitated the detection of sustained desaturation events, consistent with criteria defining desaturation clusters of at least 10 min duration³⁴. To prepare the signals for analysis, we initially addressed motion artifacts and zero-level artifacts, which commonly arise from sensor disconnections. We eliminated abrupt changes exceeding 4% per second between consecutive samples over a one-second interval and disregarded any instances where oxygen saturation fell below 50%, following the guidance of previous studies^{9,35}. We designed an algorithm inspired by these methodologies to effectively locate zero-level artifacts by detecting signal values below 50%, which are

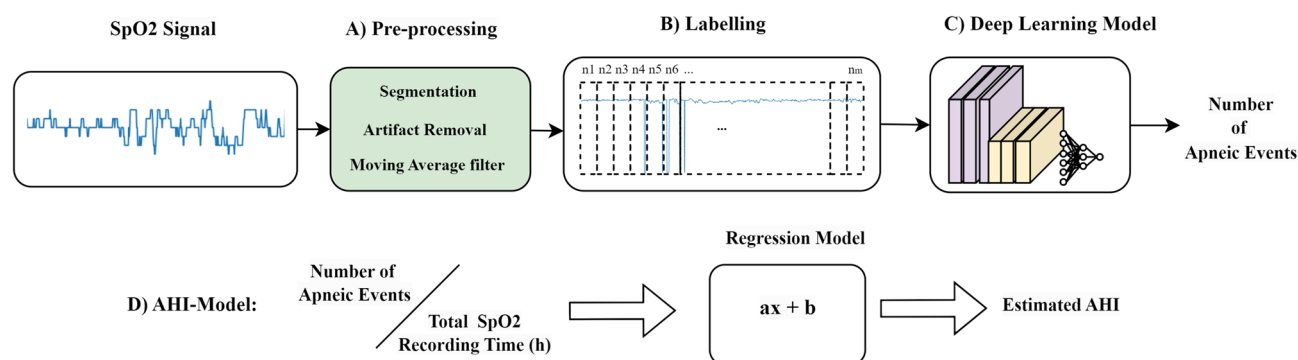


Figure 1. End-to-End Process Flow for AHI Estimation from SpO2 Signals.

not typical for a healthy individual, and identifying abrupt changes by checking for differences greater than 4% between consecutive signal values. These artifacts were then removed and substituted with values derived from linear interpolation between their preceding and following values. This step was crucial for ensuring data integrity, as significant SpO₂ drops below 50% and rapid fluctuations are often indicative of measurement errors or sensor issues. Furthermore, to smooth the signal and reduce short-term variations, we applied a 3-s moving average (MA) filter. This effectively attenuated sharp spikes and ripples in the data³⁵.

Labeling of SpO₂ signal segments

The labeling process for each 20-min SpO₂ signal segment was crucial in our study. Based on annotations provided by sleep technicians, as referenced in³², our labeling algorithm was meticulously designed to accurately identify all desaturation events associated with apneic episodes. This algorithm operates on the principle that desaturations linked to any respiratory event rely on the nadir desaturation (lowest oxygen level during desaturation) reached, typically within a 30-s span following the event's conclusion³². For each segment, the output label was determined based on the number of apnea and hypopnea events associated with a 3% oxygen desaturation occurring within the 20-min window. Figure 2 exemplifies this process, showcasing the correlation between apneic events and their subsequent oxygen desaturations in AF and SpO₂ signals. This labeling was meticulously conducted in accordance with the annotation files provided in the CHAT dataset. To validate the effectiveness of our labeling algorithm, we conducted a comparative analysis. This involved matching the number of detected apneic events, linked with a 3% oxygen desaturation, against the sum of original PSG variables from the dataset. These variables describe the number of each type of apnea event associated with a 3% oxygen desaturation. In our rigorous validation process, only recordings with a labeling error margin below 10% were considered suitable for training and evaluating our models. This criterion led to the selection of 884 SpO₂ recordings for our study. Table 1 in our paper presents the clinical and demographic data of the subjects from these selected recordings.

Deep learning model

ResNet architecture for SpO₂ signal analysis

CNN-based models have demonstrated significant effectiveness in diagnosing the severity of pediatric OSA, as shown in previous research^{27–30}. However, these networks often encounter a major obstacle when increasing the number of convolutional layers: the training loss tends to plateau, a phenomenon largely attributed to the vanishing gradient problem. The Residual Network (ResNet) framework was developed to address this challenge, notably enhancing the accuracy of deep CNNs³⁶. ResNet introduces the concept of residual learning, a paradigm shift from traditional deep network methodologies that attempt a direct mapping from input to output. Instead, ResNets focus on learning residual mappings. This is mathematically expressed as $Y = F(X) + X$, where Y denotes the desired output, $F(X)$ represents the residual function, and X is the input. This approach allows the network to

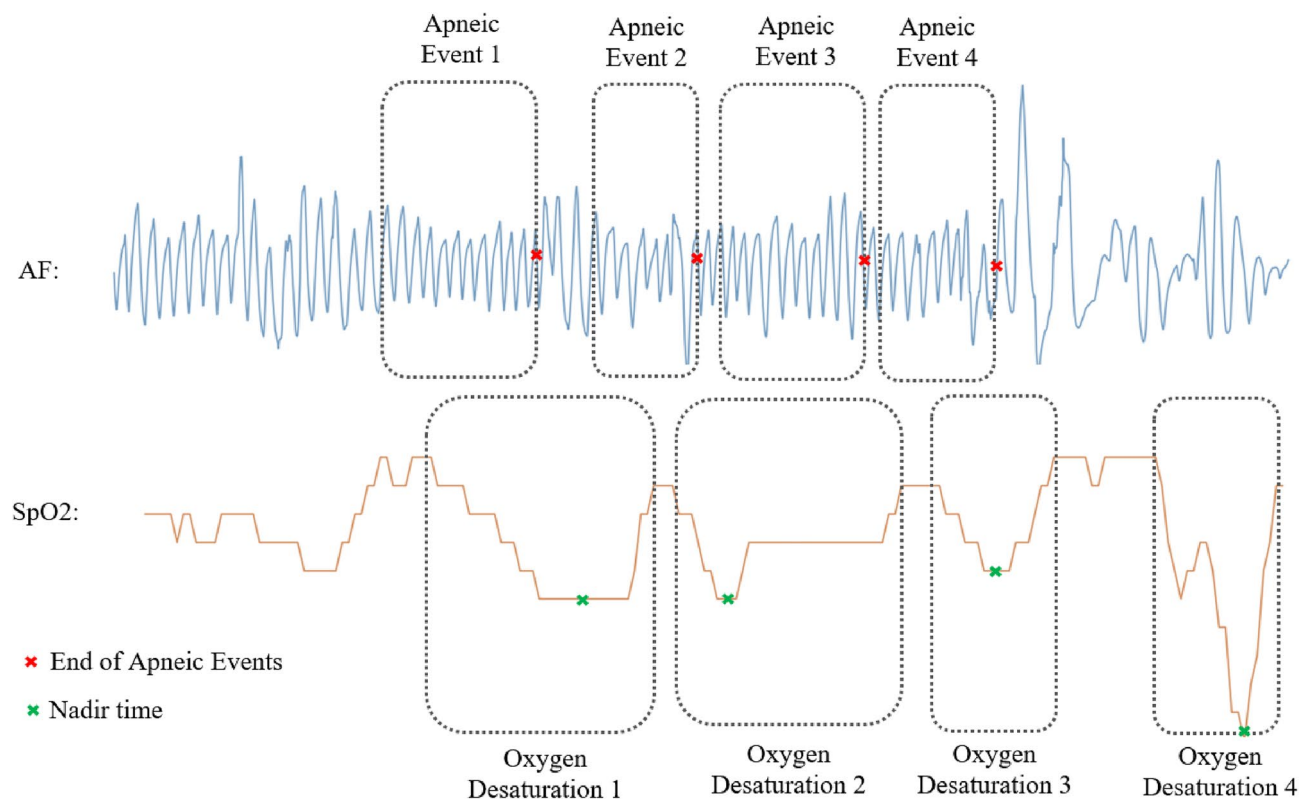


Figure 2. Synchronous events of apnea and oxygen desaturation.

Category	Normal (AHI<1)	Mild (1≤AHI<5)	Moderate (5≤AHI<10)	Severe (10≤AHI)
SpO2 signals (n)	210 (24.8%)	388 (45.9%)	150 (17.7%)	96 (11.3%)
Male	96 (45.7%)	202 (52%)	68 (45.3%)	45 (46.8%)
Age (years)	6.7±1.7	6.7±1.6	6.7±1.5	6.7±1.5
BMI (kg/m²)	17.4±4.8	19.1±5.2	19.9±5.7	20.9±5.6
Total sleep duration	462±45.7	456.4±56	454.4±55.8	462±44.7
Hypopnea duration (min)	1.3±1.7	3.9±3	11.2±5	25.9±14.5
Obstructive Apnea duration (min)	0.7±1.2	1.3±1.9	3.7±5	11.8±12.8
Central Apnea duration (min)	0.5±0.5	1.4±1.2	1.8±1.7	2.7±4.2
Mixed Apnea duration (min)	0.1±0.6	0.0±0.2	0.2±0.4	1±5.2
AHI (e/h)	0.4±0.3	2.5±1	7±1.2	18±9

Table 1. The statistical characteristics of subjects included in this study. Data are presented as number (percentage) or mean ± standard deviation. *SpO2* Blood oxygen saturation signal, *BMI* Body Mass Index, *AHI* Apnea–Hypopnea Index, *OSA* Obstructive Sleep Apnea, *CSA* Central Sleep Apnea, *MSA* Mixed Sleep Apnea.

concentrate on learning the additional information (F(X)) needed to achieve the desired output Y, particularly beneficial when F(X) is close to zero. The benefits of ResNet include easing the training of deep networks by alleviating the gradient vanishing issue and enabling the construction of much deeper networks without sacrificing accuracy. Additionally, ResNet incorporates skip connections, which directly add the input X to the output of the residual function F(X), promoting smoother gradient flow during training. Figure 3 illustrates a typical residual block, showing how input X is transformed into its desired mapping Y.

In our study, we have adapted the ResNet-34 architecture, initially designed for image recognition³⁶, into a 1D format suitable for analyzing SpO2 signals. This adaptation involved replacing 2D convolutional layers (Conv2D) with 1D convolutional layers (Conv1D), which are more apt for processing time-series data like SpO2 signals. This adaptation involved replacing 2D convolutional layers (Conv2D) with 1D convolutional layers (Conv1D), which are more apt for processing time-series data like SpO2 signals. Given the size of our dataset, we started with the relatively shallower ResNet-34 model, consisting of 34 layers. However, to mitigate overfitting, we optimized the model to 16 layers, which provided efficient performance with quicker convergence compared to deeper models.

Figure 4 presents a detailed depiction of our modified ResNet architecture, illustrating the sequence and function of each layer:

- *Input Layer*: Receives the raw 20-min SpO2 signal segment with a size of 1200 × 1.
- *Conv1D Layer*: Applies learnable filters to the input, extracting fundamental patterns and low-level features.
- *Batch Normalization*: Normalizes the activations from the Conv1D layer, stabilizing and expediting the training by reducing internal covariate shift.
- *ReLU Activation*: Introduces non-linearity through the rectified linear unit (ReLU) function.
- *MaxPooling1D*: Reduces the dimensionality of the data, maintaining essential features while lessening computational load.
- *Residual Blocks*: Each block, consisting of two convolutional layers, forms the core component of the ResNet. Our model includes seven such blocks.
- *Flatten Layer*: Transforms the output of the last residual block into a one-dimensional vector, ensuring compatibility with the subsequent fully connected layer.
- *Fully Connected Layer*: Processes the received output to perform the final mapping, producing a numerical prediction.
- *Output Layer*: Generates the final prediction for the number of apneic events (y_{pred}).

Development of the CNN-BiGRU-attention architecture

In our study, we designed a sophisticated model that combines a CNN with a Bidirectional Gated Recurrent Unit (BiGRU) and an attention mechanism. This integrative approach, inspired by its successful application in blood pressure estimation³⁷, is adapted here to process SpO2 signals for AHI estimation. The incorporation of the attention mechanism is driven by its proven effectiveness in focusing on pertinent segments within datasets, a principle extensively utilized in diverse fields such as image captioning, machine translation, and speech recognition³⁸. Our model aims to investigate the synergy of these distinct architectures, assessing how the RNN structure processes features extracted by the CNN and the role of the attention mechanism in augmenting RNN performance for AHI estimation. Figure 5 in our paper illustrates the architecture of the proposed model. The process begins with the CNN layer, which is responsible for extracting pertinent features from the input SpO2 signals. Following feature extraction, the BiGRU layer, recognized for its capability to handle long-term dependencies in sequential data, processes these features. This is achieved by analyzing the data in both forward and backward temporal directions, thereby capturing complex temporal dynamics inherent in the SpO2 signals. Furthermore, we integrate an attention mechanism with the outputs of the BiGRU layer. This mechanism assigns weights to different temporal features, enabling the model to concentrate its predictive capacity on the most crucial segments of the signal. This combination—CNN for initial feature extraction, BiGRU for in-depth

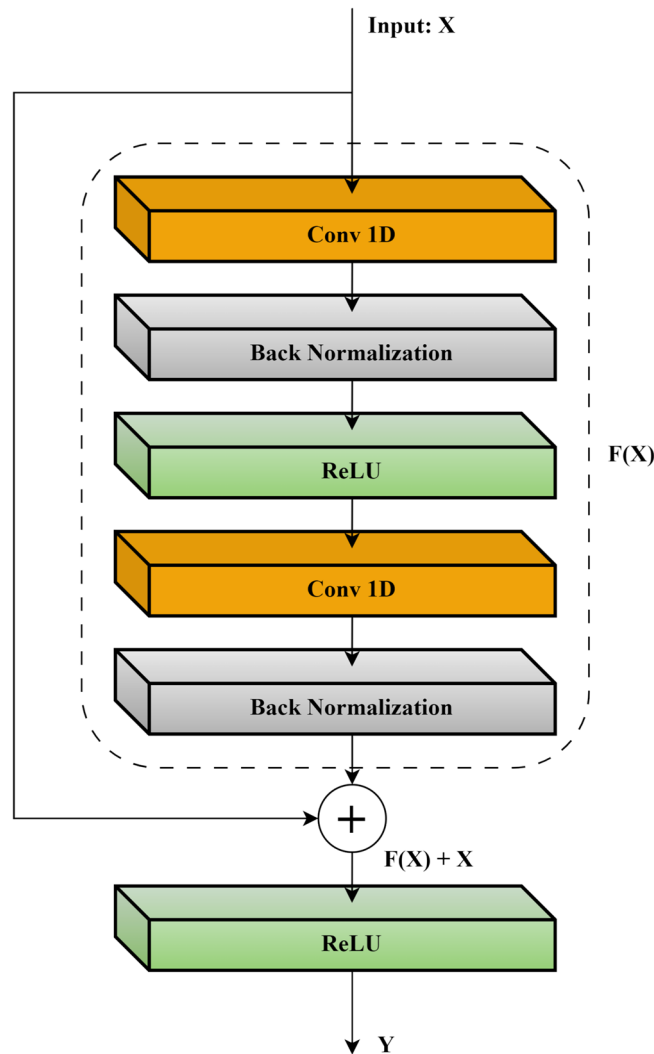


Figure 3. Schematic of a residual learning block.

temporal processing, and attention for targeted focus—constitutes a potent framework designed to enhance the accuracy of AHI estimation from SpO₂ signals.

It is pertinent to note that while RNNs are pivotal in handling sequential data tasks such as speech recognition, they often struggle with long sequences due to the vanishing gradient problem³⁹. LSTM networks, introduced by Hochreiter and Schmidhuber⁴⁰, have been developed to counter this issue, incorporating specialized gates to manage information flow. Bidirectional LSTMs (BiLSTMs) further refine this approach by processing data in both forward and backward directions, thus encompassing past and future contexts in the analysis. BiGRUs, a variant conceived by Cho et al.⁴¹, streamline the design of LSTMs by combining the input and forget gates, thereby reducing the model's complexity while retaining efficiency in processing bidirectional sequence data. The architecture of the GRU cell employed in this analysis is portrayed in Fig. 6. At each time step, the GRU cell encompasses two crucial input vectors: the preceding hidden output value h_{t-1} , which includes feature values from the previous time step across all feature maps extracted by the CNN layer filters, and the current input vector x_t , which contains the current feature values from all feature maps extracted by the CNN layer filters. The computation of the contemporary hidden output value of the cell h_t unfolds through the following equations:

$$\begin{aligned}
 z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]) \\
 r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]) \\
 \tilde{h}_t &= \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t]) \\
 h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t
 \end{aligned} \tag{1}$$

Here, z_t and r_t denote the update and reset gate vectors, respectively. The weight parameters W_z , W_r and W_h are trainable and contribute to the gate operations. The term \tilde{h}_t signifies the candidate state, capturing the extent

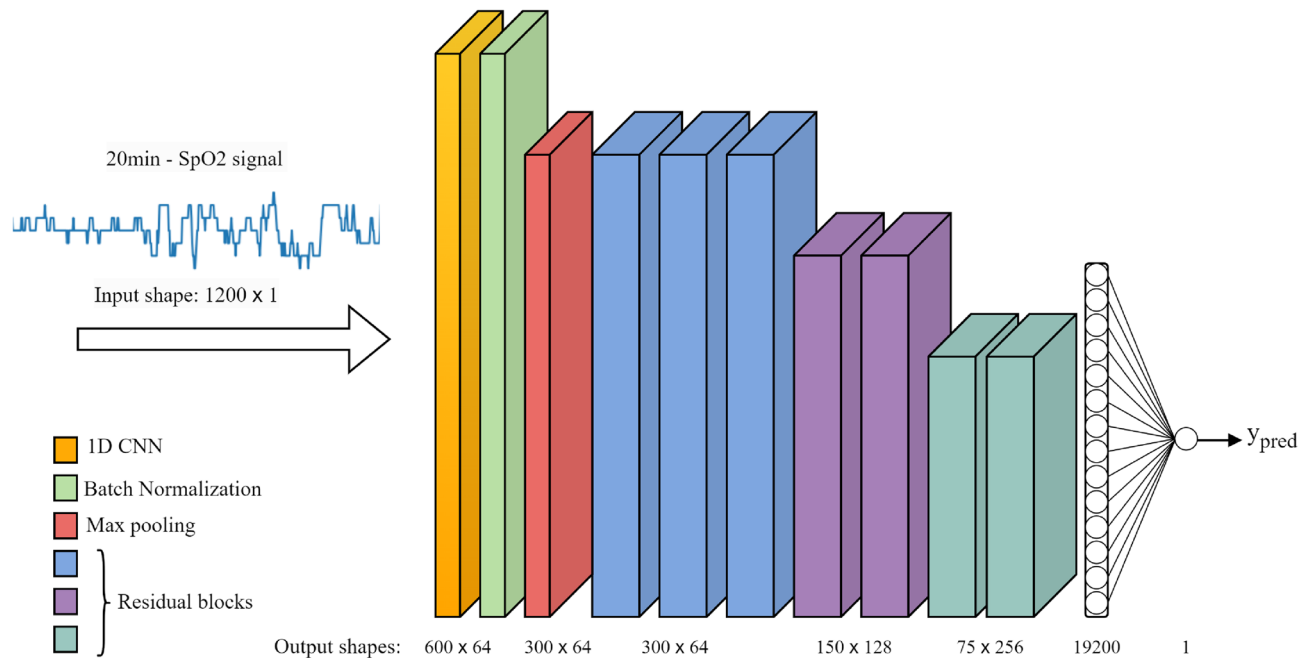


Figure 4. Architecture of the modified ResNet model for SpO2 signal analysis.

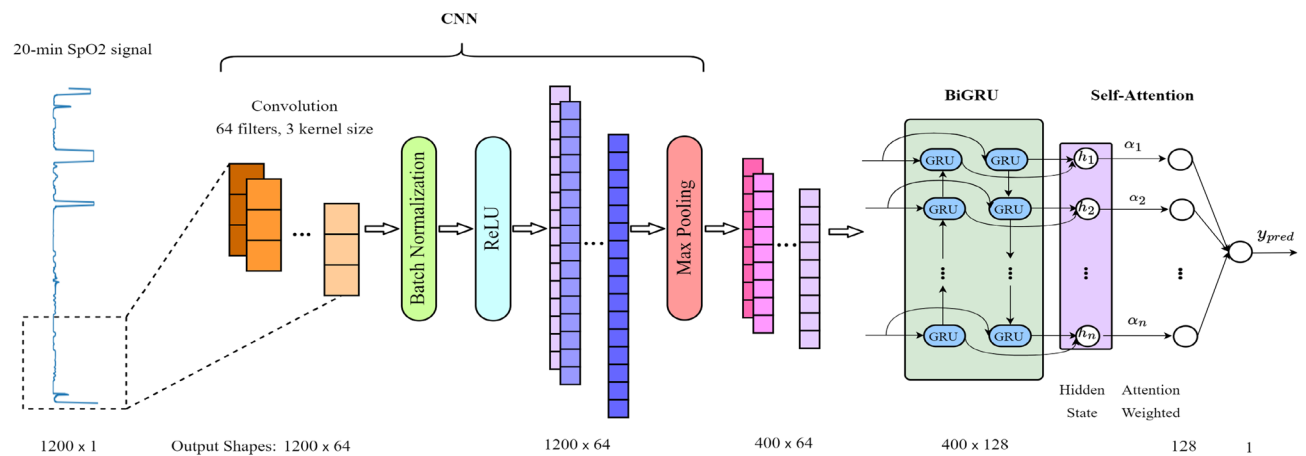


Figure 5. CNN-BiGRU Attention model architecture for SpO2 signal analysis.

of assimilating present information post the reset gate. The activation functions $\sigma(\bullet)$ and $\tanh(\bullet)$ encapsulate the sigmoid and hyperbolic tangent functions, respectively, while \odot signifies element-wise multiplication. Unlike its unidirectional counterpart, the conventional GRU, a bidirectional GRU (BiGRU) is adopted in this study. A BiGRU encompasses both forward and backward layer cells' hidden output values. Figure 7 shows the architecture of a BiGRU layer structure, with one pair of GRU cells at each time step. The final hidden output vector of the BiGRU layer at time step t , \vec{h}_{out_t} is a concatenation of the forward hidden output vector \vec{h}_t (including forward layer's cells hidden output values) and backward hidden output vector \overleftarrow{h}_t (including backward layer's cells hidden output values):

$$\begin{aligned}\vec{h}_t &= [h_{t_{c1}}, h_{t_{c2}}, \dots, h_{t_{c_{n-1}}}, h_{t_{cn}}] \\ \overleftarrow{h}_t &= [h_{t_{cn}}, h_{t_{c_{n-1}}}, \dots, h_{t_{c2}}, h_{t_{c1}}] \\ \vec{h}_{out_t} &= [\vec{h}_t, \overleftarrow{h}_t]\end{aligned}\quad (2)$$

The attention mechanism, initially introduced by Bahdanau et al. in 2014⁴² to address limitations in traditional sequence-to-sequence models, marked a significant breakthrough in natural language processing. This

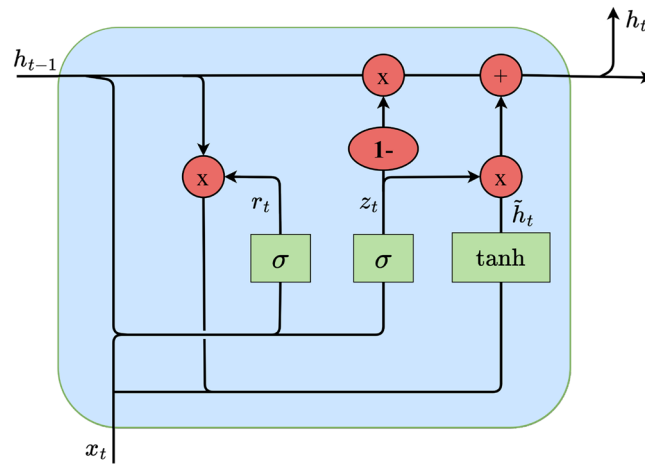


Figure 6. Schematic of a Gated Recurrent Unit (GRU) cell.

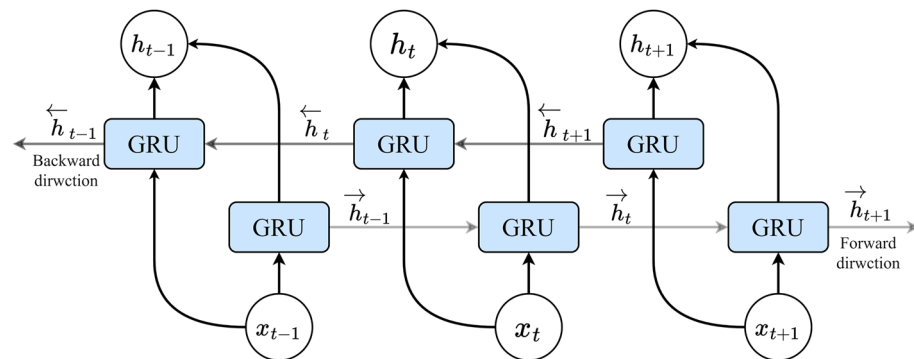


Figure 7. Schematic of a Bidirectional GRU (BiGRU) layer structure with one pair GRU cells at each time step.

mechanism, commonly referred to as attention, has since evolved into self-attention or intra-attention, finding widespread application in diverse DL tasks. In our work, we harness the power of self-attention to enhance the model's ability to capture crucial temporal features within sequential SpO2 signals. By applying self-attention to the outputs of the BiGRU layer, our model dynamically assigns weights to different temporal features, prioritizing the most relevant signal segments for accurate AHI estimation. In developing the self-attention mechanism, we consider the final hidden state matrix H_s from the BiGRU including hidden output vectors of the BiGRU at each time step h_t , where $t \in [1, N]$. The significance score vector \vec{s} is computed using a score function, $score(\cdot)$, before multiplying the hidden state matrix by a randomly initialized weight and bias vectors (\vec{w} and \vec{b}) as outlined below:

$$\vec{s} = score(H_s \vec{w} + \vec{b}), H_s = \begin{bmatrix} \vec{h}_{out1} \\ \vec{h}_{out2} \\ \vdots \\ \vec{h}_{outN} \end{bmatrix} \quad (3)$$

For the score function, we initially experimented with the dot product, tanh, and ReLU functions. Ultimately, we selected ReLU as the score function because it provided better performance and faster convergence during the training process. After obtaining the importance score values for each BiGRU hidden output vectors at each time step (including the hidden output value of all cells) and forming the score vector of \vec{s} , the attention weight α_i for the hidden output vector at time step t is determined by applying a softmax function to the score vector:

$$\vec{s} = [s_1, s_2, \dots, s_N]$$

$$a_i = \frac{\exp(s_i)}{\sum_{j=1}^N \exp(s_j)} \quad (4)$$

This softmax function guarantees that the attention weights collectively sum to 1, effectively normalizing the significance scores across all time step vectors within the hidden state matrix. The attention weights vector \vec{a} includes all attention weights for each cell's hidden output vector. The final output vector \vec{v} is derived by multiplying the attention vector by the hidden state matrix of the BiGRU:

$$\vec{a} = [a_1, a_2, \dots, a_N]$$

$$\vec{v} = \vec{a} H_s \quad (5)$$

This summation offers a comprehensive representation of the input sequence, highlighting the contributions of distinct time steps based on their computed attention weights. To further clarify the effect of the attention layer on each part of the input signal segment, we plotted an attention map (heatmap) of the attention weights vector (attention scores) for a 20-min signal segment from the test set, as represented in Fig. 8. The heatmaps, obtained by graphing the alpha vector resulting from the softmax output in Eq. (4), were scaled using log10 to better represent the distribution of attention score values.

Model training and optimization

For training the ResNet and CNN-BiGRU-Attention models, we employed distinct initialization methods. The ResNet model was initialized using the He-normal method⁴³, while the CNN-BiGRU-Attention model started with random weights. Data were fed into both models in batches and shuffled at each training epoch to enhance convergence. Since the number of non-apneic SpO2 segments exceeded the apneic segments, we balanced the data by oversampling the apneic segments, repeating them before the start of each training process. To facilitate efficient weight updates, we employed the adaptive moment estimation (Adam) optimizer⁴⁴ with an initial learning rate. For the loss function, we chose the Huber loss⁴⁵ due to its robustness, as evidenced by its strong performance in previous AHI estimation studies^{28,29}. The Huber loss strikes a balance between quadratic and linear loss behaviors, as expressed by its formula:

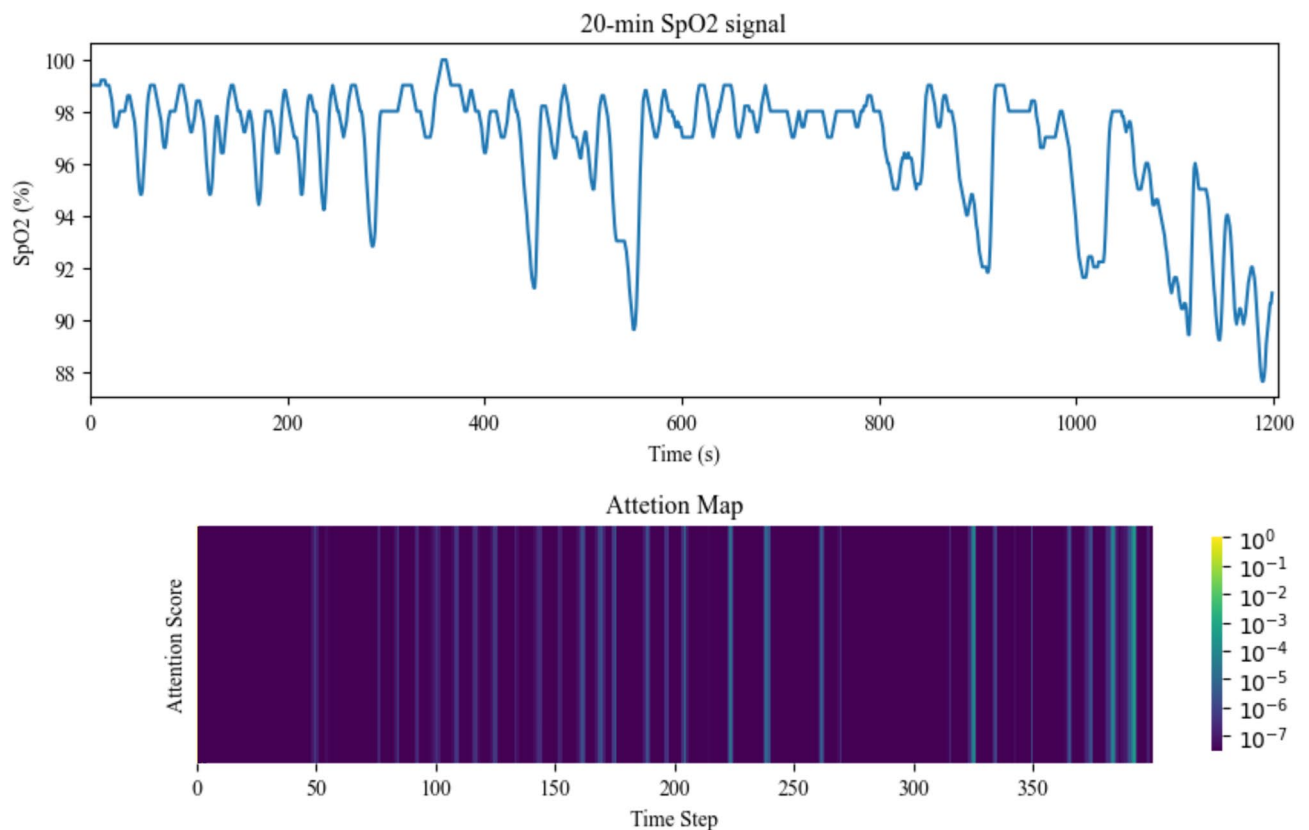


Figure 8. Attention Map of the Attention Scores Derived from the Attention Layer for a 20-Minute Signal Segment Input.

$$L(\hat{y}_n, y_n, \delta) = \begin{cases} \frac{1}{2}(\hat{y}_n - y_n)^2, & |\hat{y}_n - y_n| \leq \delta \\ \hat{y}_n(|\hat{y}_n - y_n| - \frac{\delta}{2}), & \text{otherwise} \end{cases} \quad (6)$$

Here, \hat{y}_n and y_n denote the label and model output for segment n , respectively. The parameter δ acts as a threshold and serves as a tunable hyperparameter crucial for effectively handling data with outliers and noise during the model optimization process. The Huber loss function is particularly effective for datasets with outliers, providing a quadratic loss for smaller errors (inliers) and a linear loss for larger errors (outliers). Two key techniques were implemented to optimize the training process. Firstly, a dynamic learning rate reduction strategy was employed, reducing it by 50% every 10 epochs within the loss function. This strategy promotes training stability, facilitating smoother convergence. Secondly, early stopping was introduced to halt the training process if the validation set loss did not improve for 30 consecutive epochs, ensuring the generalization capability of the models. We employed the Keras deep learning framework with a TensorFlow backend for model training in the Google Colab environment, leveraging the availability of NVIDIA Tesla T4 GPUs.

AHI estimation

After obtaining the estimated number of apneic events in each 20-min segment, the AHI is calculated as the sum of the detected apneic events divided by the total recording duration. Utilizing the DL model's output y_n for each 20-min SpO₂ segment ($n = 1, 2, 3, \dots, N$), the AHI for each patient is calculated using the formula:

$$\text{AHI - Model} = \frac{\sum_{n=1}^N y_n}{\text{SpO}_2 \text{ signal recording time (h)}} \quad (7)$$

where N is the total number of 20-min SpO₂ segments in the SpO₂ signal. To calculate the AHI, knowledge of the total sleep duration for each patient is essential, particularly for sleep staging analysis. In the absence of this information, we made the assumption that the total length of the SpO₂ signal serves as an approximation of the sleep duration. To refine this estimation, a regression model was employed to map the calculated AHI, based on the total SpO₂ recorded time, to the actual AHI, taking sleep duration into consideration. The regression model utilizes an optimization method to determine the optimal coefficients of the linear equation $ax + b$ that minimizes an error function. In this study, we opted for the Huber loss function (Huber regressor) to train the model on the validation set. The Huber loss function adjusts the parameters of the linear equation during training to minimize the loss across all predictions. To strike a balance between sensitivity to outliers and overall model performance, the delta value is crucial. In our case, the delta value was determined by selecting the value that minimized the root mean square error on the validation set. The choice of a delta value significantly influences the behavior of the Huber loss function. A higher delta makes the loss more robust to outliers, as it allows for a larger linear region, while a lower delta increases sensitivity to outliers by making the loss more quadratic. In our study, after experimenting with different delta values, we found that a delta of 6 led to the minimum root mean squared error on the validation set. This particular value was chosen to strike an optimal balance, aiming to maximize the accuracy of our AHI estimates in the presence of data variability and potential outliers. This approach effectively mitigates the lack of explicit sleep duration information, enhancing the accuracy of AHI estimation. The final estimated AHI is obtained through the Huber regression model. Figure 9 displays a scatter plot of total record time-based AHI and actual AHI in the validation set, along with the fitted regression function line on data points.

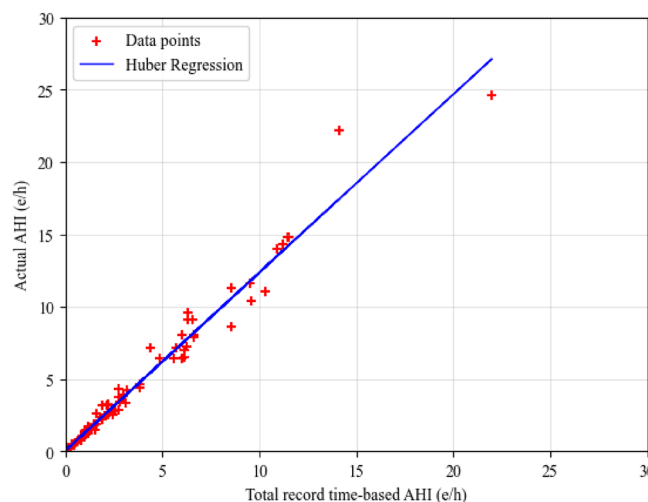


Figure 9. Scatter plot of total record time-based AHI and actual AHI in the validation set.

Performance evaluation

To assess the proficiency of the models in estimation, we employed fundamental regression metrics. These metrics include mean absolute error (MAE), root mean squared error (RMSE) and R-squared (R^2) for per-recording AHI estimation. The formulas for each metric are specified accordingly:

(1) MAE.

$$MAE = \frac{1}{n} \sum_{i=1}^n |AHI_{\text{actual}, i} - AHI_{\text{estimated}, i}| \quad (8)$$

(2) RMSE.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (AHI_{\text{actual}, i} - AHI_{\text{estimated}, i})^2} \quad (9)$$

(3) R^2 .

$$R^2 = 1 - \frac{\sum_{i=1}^n (AHI_{\text{actual}, i} - AHI_{\text{estimated}, i})^2}{\sum_{i=1}^n (AHI_{\text{actual}, i} - \overline{AHI}_{\text{actual}})^2} \quad (10)$$

We utilized scatter and Bland–Altman⁴⁶ plots to compare the predicted AHI by the models with the actual AHI from the PSG test, ensuring proper alignment for a comprehensive evaluation. For apnea severity classification, the overall agreement of predicted AHI in estimating the severity of SAH was evaluated using confusion matrices, four-class Cohen's kappa coefficient (kappa)⁴⁷, and four-class accuracy. To ensure a thorough evaluation, we categorized patients based on common AHI thresholds of 1, 5, and 10 events per hour (e/h), facilitating binary classification into those below and above each specified threshold. For each threshold, we assessed the model's diagnostic performance, covering sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR+), and negative likelihood ratio (LR-):

- Sensitivity (Se): Percentage of SAH positive patients correctly classified.
- Specificity (Sp): Percentage of SAH negative patients correctly classified.
- Positive Predictive Value (PPV): Proportion of actual positive cases among instances predicted as positive by the model.
- Negative Predictive Value (NPV): Proportion of actual negative cases among instances predicted as negative by the model.
- Positive Likelihood Ratio (LR+): $Se/(1-Sp)$.
- Negative Likelihood Ratio (LR-): $(1-Se)/Sp$.

This comprehensive approach provided a robust evaluation of our model's accuracy in effectively categorizing patients according to the severity of SAH.

Results

The study employed a threefold cross-validation approach on the dataset, reserving a fixed 10% for the validation set and selecting a test set in each iteration, leaving the remaining data for model training. Before initiating the cross-validation, an extensive exploration of diverse hyperparameter combinations was conducted to optimize the model's performance. The ultimate hyperparameters were chosen based on achieving the highest four-class kappa value on the validation set. The selected configuration encompassed a learning rate of 0.001, 120 epochs, a Huber loss delta (δ) value of 1.5, and a batch size of 32. Following the threefold cross-validation, Figs. 10 and 11 showcase scatter plots for each ResNet and CNN-BiGRU-Attention model, respectively, on the test set of each fold. These scatter plots display the estimated AHI by each model against the actual AHI for each of the three test sets, along with R^2 and RMSE. Notably, in both figures, the scatter plot points of the Fold 2 test set exhibit a higher concentration near the diagonal line, indicating superior agreement between the actual and estimated AHI compared to the other fold test sets. Figure 12 introduces a Bland–Altman plot for the estimated AHI by the ResNet model compared to the reference AHI. Across all folds, the majority of data points fall within the confidence interval, indicating an acceptable level of agreement (LoA). In Fold-1, the negative mean value suggests a slight underestimation by the proposed model in the test subset. The Fold-2 plot demonstrates a positive mean error close to zero (0.09), with narrower LoA (−3.26 to 3.45), reflecting a more accurate estimation with reduced variability and superior agreement compared to Folds 1 and 3. In Fold-3, the mean bias is essentially negligible (0.03); however, the LoA is the broadest (−5.89 to 5.83) compared to other folds, indicating increased variability in the estimates despite an unbiased mean. Figure 13 exhibits a Bland–Altman plot for the estimated AHI by the CNN-BiGRU-Attention model compared to the reference AHI. Across all folds, the negative mean value indicates a slight tendency of CNN-BiGRU-Attention to underestimate the AHI. The Fold-2 plot showcases narrower LoA (−3.5 to 2.8), reflecting a more accurate estimation with reduced variability and superior agreement compared to other folds, while Fold-3 has the broadest LoA, suggesting increased variability in the estimates.

The confusion matrix for each model in every fold test set is meticulously outlined in Fig. 14. Additionally, Table 2 provides a comprehensive overview of the regression and classification metrics for each model's

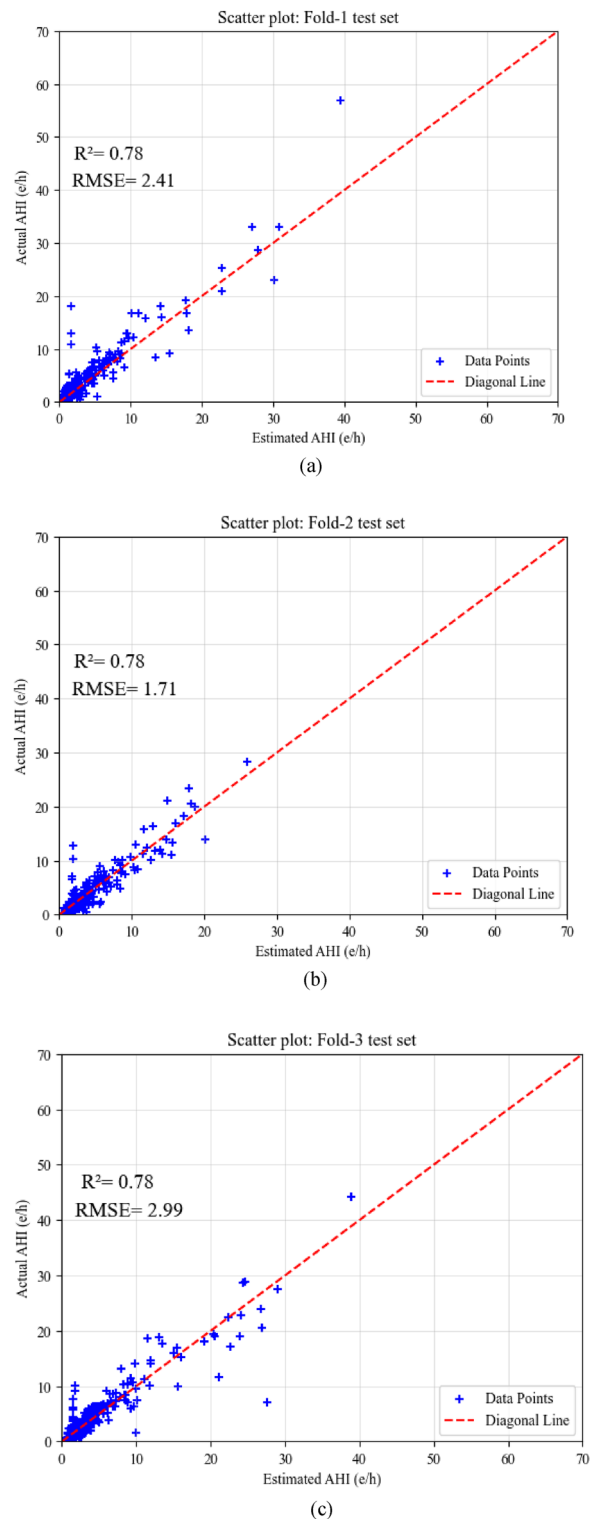


Figure 10. Scatter plots between actual AHI and estimated AHI by ResNet model in fold-1 (a), fold-2 (b) and fold-3 test sets.

performance in each fold test set. The ResNet model exhibited the highest four-class accuracy and four-class kappa value in the fold-3 test set, despite having the highest RMSE value. In contrast, the CNN-BiGRU-Attention model showcased the highest four-class accuracy and four-class kappa value on the fold-2 test set with the minimum RMSE across all folds. For additional insights into the diagnostic capabilities of each model at commonly used AHI thresholds of 1, 5, and 10 e/h, Table 3 presents detailed results for each fold test set.

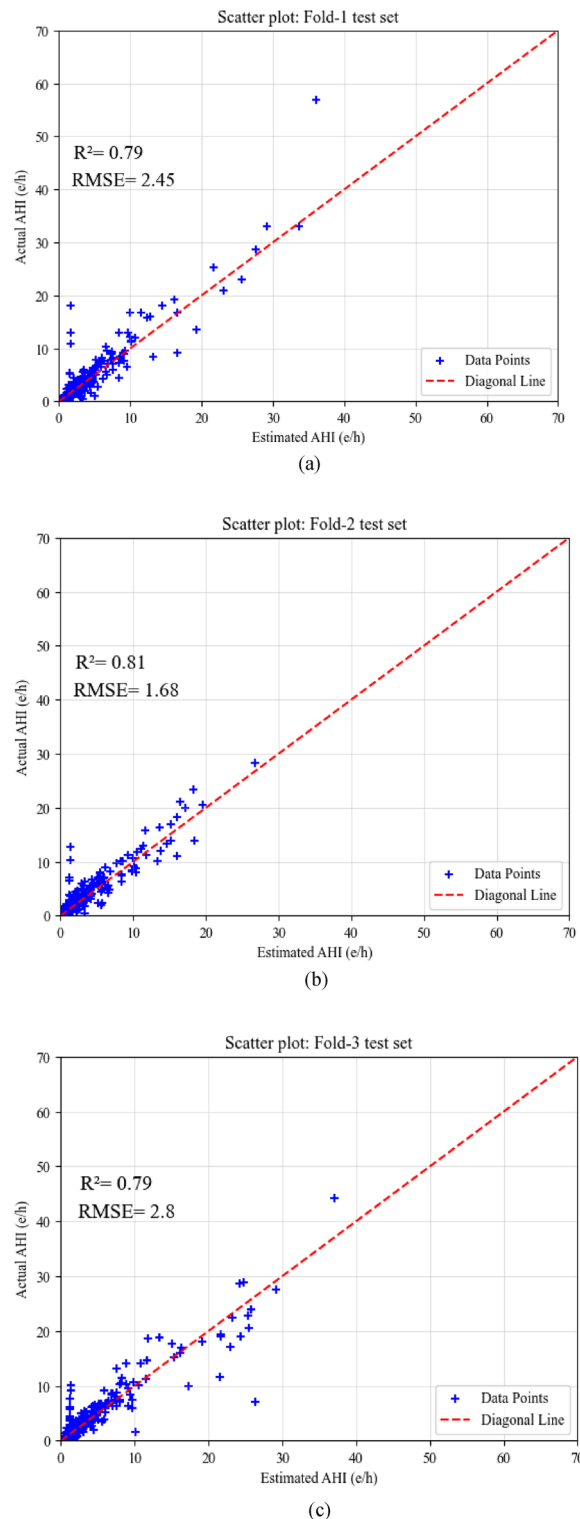


Figure 11. Scatter plots between actual AHI and estimated AHI by CNN-BiGRU-Attention model in fold-1 (a), fold-2 (b) and fold-3 test sets.

It is worth noting that, based on the table results, at the AHI = 1 e/h threshold, both models may tend to misclassify healthy subjects (normal group) as SAH groups. At higher AHI thresholds (AHI = 5e/h and AHI = 10e/h), the models tend to misclassify subjects with higher SAH severity groups (moderate and severe) as lower SAH (mild and normal). Therefore, it is important to ensure that at AHI = 1e/h, the specificity value is not low, along with accuracy, and at high AHI thresholds, sensitivity should not be low alongside accuracy. As seen in the table, despite a reduction in sensitivity with higher AHI thresholds, both models maintained remarkable sensitivity

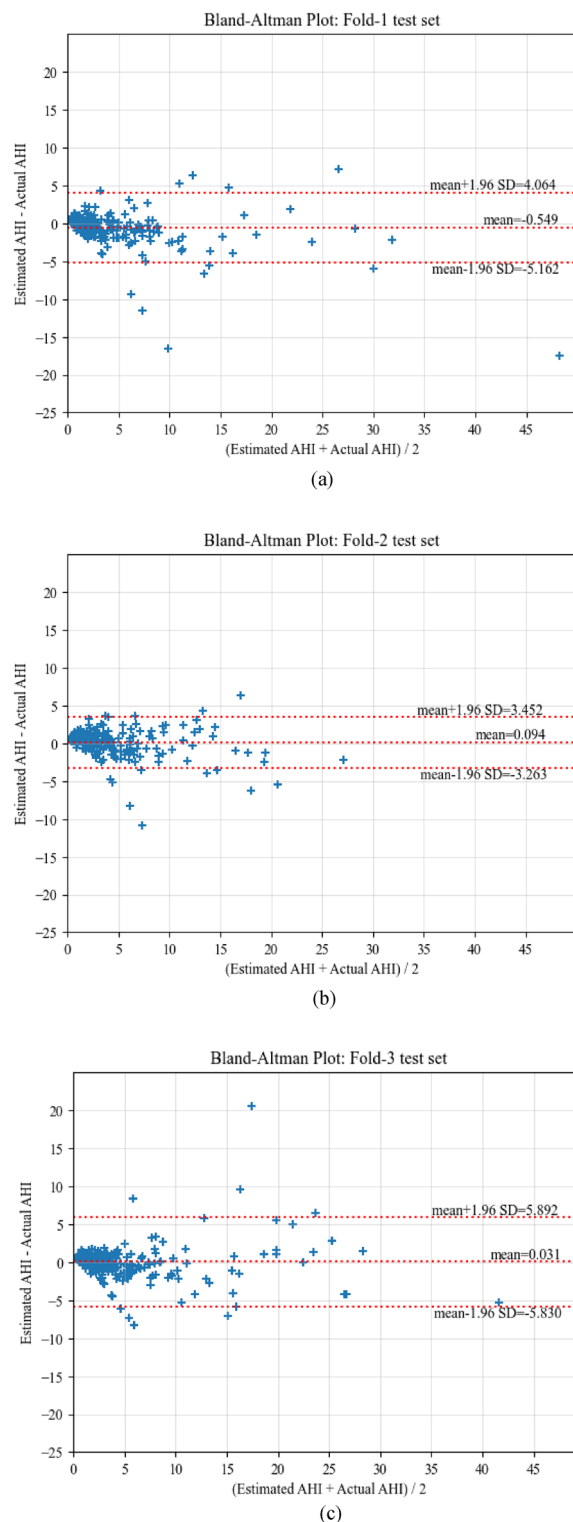


Figure 12. Bland–Altman plots of actual AHI and estimated AHI by ResNet model in fold-1 (a), fold-2 (b) and fold-3 test sets.

and accuracy, indicating their proficiency in accurately detecting patients with high severity SAH, which holds pivotal clinical implications. The ResNet model exhibited overall high accuracy, sensitivity, and specificity on fold-3 test set, particularly with high sensitivity and accuracy at AHI thresholds of 5 e/h and 10 e/h on fold-2. The CNN-BiGRU-Attention model demonstrated overall high accuracy, sensitivity, and specificity on fold-2 test set, particularly with high sensitivity and accuracy at AHI thresholds of 5 e/h on fold-1.

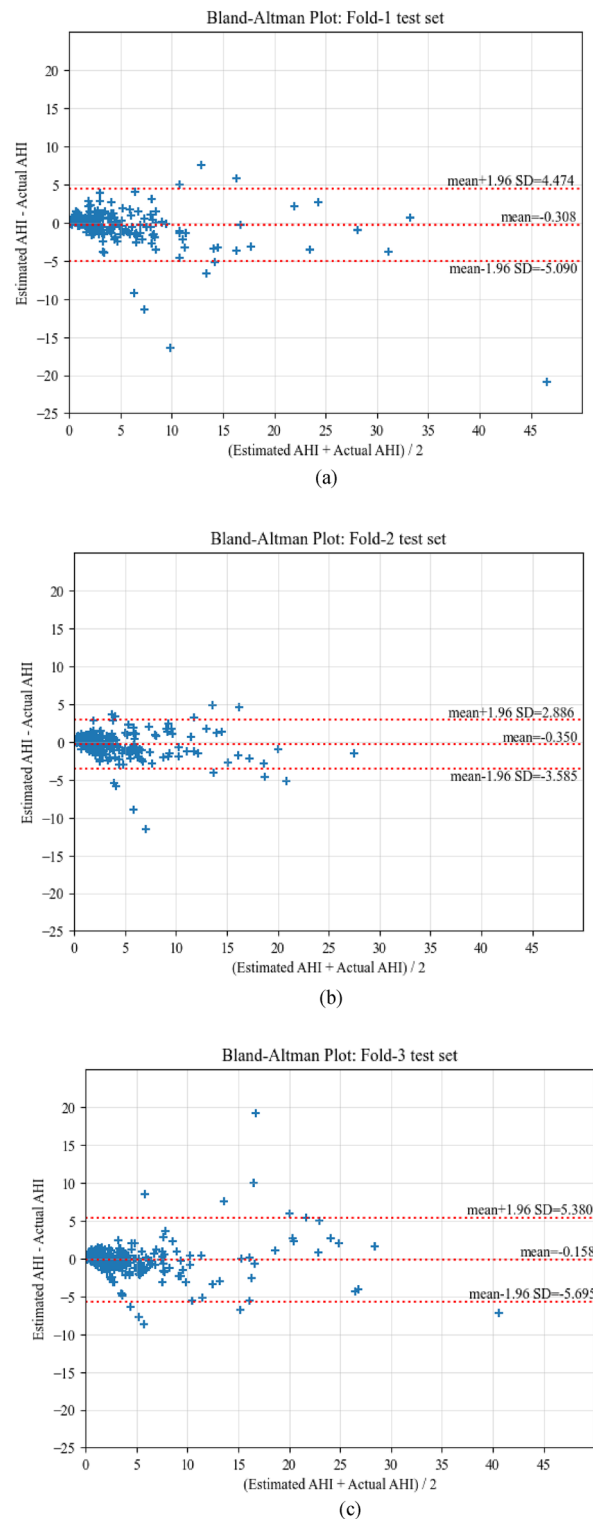


Figure 13. Bland–Altman plots of actual AHI and estimated AHI by CNN-BiGRU-Attention model in fold-1 (a), fold-2 (b) and fold-3 test sets.

Model comparison

To assess the overall performance of the models, we analyzed the average metrics across all folds, including four-class accuracy, four-class kappa, and RMSE. Results in Table 4 demonstrate CNN-BiGRU-Attention outperforming the ResNet. CNN-BiGRU-Attention model exhibited higher average four-class accuracy and kappa values, as well as lower RMSE, across all fold test sets. Moreover, it featured significantly fewer parameters and a smaller size compared to the ResNet model, resulting in reduced training time. The findings in Table 2 consistently reinforce

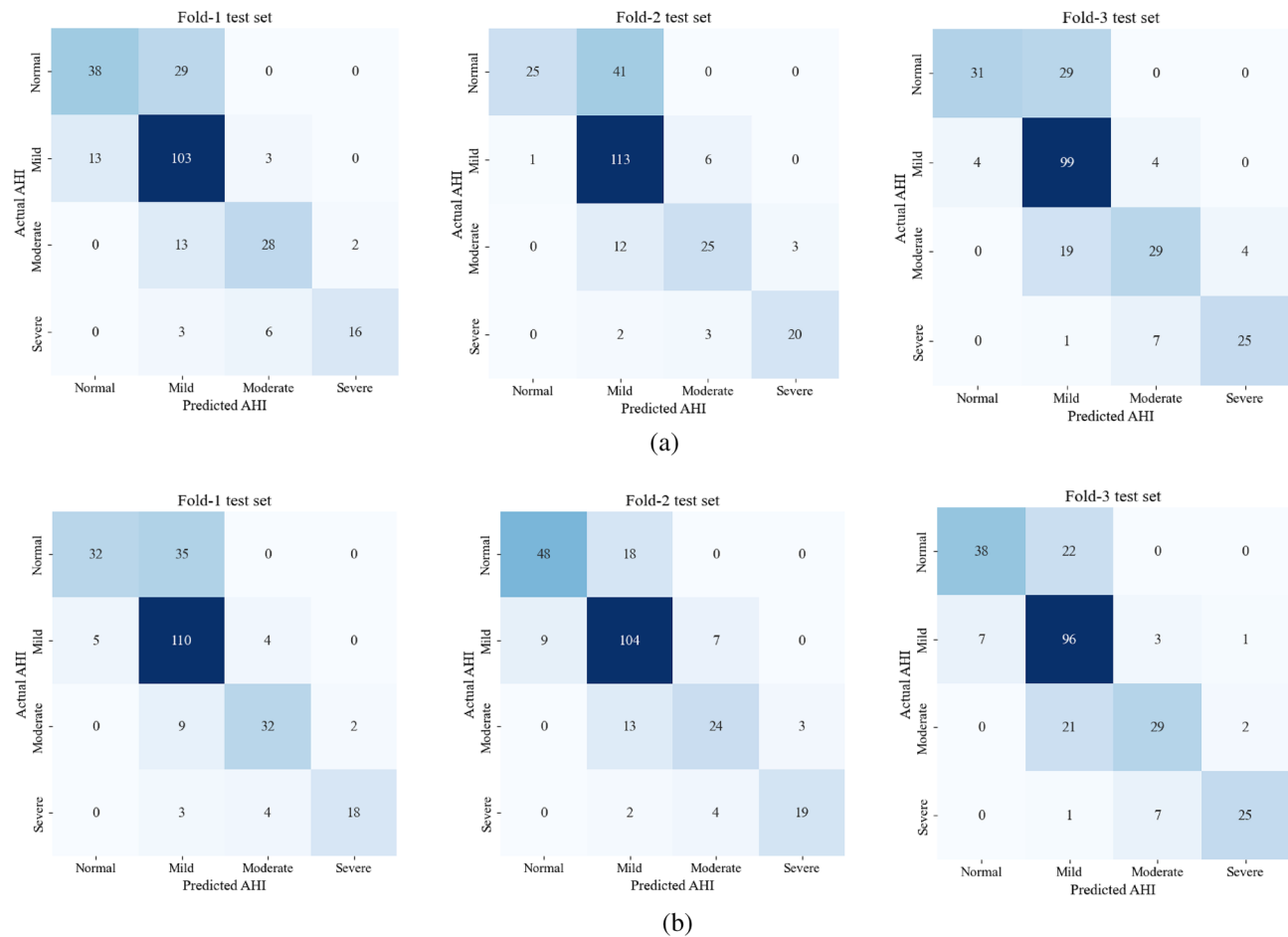


Figure 14. Confusion matrices of the predicted SAH severity group by ReaNet model (a) and CNN-BiGRU-Attention model (b) in each fold test set.

Model	Fold	Test set (n)	MAE	RMSE	R ²	Four-class kappa	Four-class accuracy (%)
ResNet	Fold-1	254	1.31	2.41	0.78	0.57	72.8
	Fold-2	253	1.13	1.71	0.78	0.56	72.90
	Fold-3	253	1.46	2.99	0.78	0.59	73.01
CNN-BiGRU-Attention	Fold-1	254	1.25	2.45	0.79	0.61	75.59
	Fold-2	253	1.06	1.68	0.81	0.65	77.68
	Fold-3	253	1.44	2.8	0.79	0.62	74.60

Table 2. Models performance on each fold test set. MAE Mean Absolute Error, RMSE Root Mean Squared Error, R² R-squared.

the superior performance of CNN-BiGRU-Attention on each fold test set. The ResNet model is equipped with residual connections that help mitigate the vanishing gradient problem, thus facilitating the training of deeper networks. However, its primary focus on spatial feature extraction may not adequately capture the temporal dependencies present in sequential SpO2 data. In contrast, the CNN-BiGRU-Attention model integrates CNN for spatial feature extraction with BiGRU for temporal feature learning. The inclusion of an attention layer further refines the model's focus on the most pertinent aspects of the signal, thereby enhancing overall performance. This hybrid approach enables the CNN-BiGRU-Attention model to more effectively capture both spatial and temporal features compared to ResNet, which is particularly beneficial for time-series data like SpO2 signals that contain critical temporal dependencies for accurate apnea-hypopnea event detection. Despite ResNet's advanced deep learning capabilities, our limited dataset posed overfitting challenges when training deeper layers. This experience highlighted that more streamlined architectures like CNN-BiGRU-Attention can offer better generalization across different folds in cross-validation, whereas ResNet may be more susceptible to overfitting, leading to less consistent performance on validation and test sets. To assess the importance of each component in the proposed CNN-BiGRU-Attention model, we conducted an ablation study examining four configurations:

Model	Fold	AHI threshold	Se (%)	Sp (%)	Acc (%)	PPV (%)	NPV (%)	LR +	LR-
ResNet	Fold-1	1 e/h	93.04	56.71	83.46	85.71	74.50	2.14	0.12
		5 e/h	76.47	98.38	92.51	94.54	91.95	47.41	0.23
		10 e/h	64.00	99.12	95.66	88.88	96.18	73.27	0.36
	Fold-2	1 e/h	99.45	37.87	83.26	81.77	96.15	1.60	0.01
		5 e/h	78.46	96.77	92.03	89.47	92.78	24.32	0.22
		10 e/h	80.00	98.67	96.81	86.95	97.80	60.26	0.20
	Fold-3	1 e/h	97.91	51.66	86.90	86.63	88.57	2.02	0.04
		5 e/h	76.47	97.60	90.47	94.20	89.07	31.92	0.24
		10 e/h	75.75	98.17	95.23	86.20	96.41	41.47	0.24
CNN-BiGRU-Attention	Fold-1	1 e/h	97.33	47.76	84.25	83.87	86.49	1.86	0.05
		5 e/h	82.35	97.85	93.70	93.33	93.81	38.29	0.18
		10 e/h	72.00	99.13	96.46	90.00	97.01	82.44	0.28
	Fold-2	1 e/h	95.14	72.73	89.24	90.72	84.21	3.48	0.06
		5 e/h	76.92	96.24	91.24	87.72	92.27	20.43	0.23
		10 e/h	76.00	98.67	96.41	86.36	97.38	57.25	0.24
	Fold-3	1 e/h	96.35	63.33	86.11	88.49	84.44	2.62	0.05
		5 e/h	74.11	97.60	89.68	94.02	88.10	30.94	0.26
		10 e/h	75.75	98.63	95.63	89.28	96.42	55.30	0.24

Table 3. The diagnostic ability of each model for AHI thresholds of 1, 5 AND 10 E/H. *Se* Sensitivity, *Sp* Specificity, *Acc* Accuracy, *PRV* Positive Predictive Value, *NPV* Negative Predictive Value, *LR +* Positive Likelihood Ratio, *LR -* Negative Likelihood Ratio.

Model	Average four-class accuracy (%)	Average four-class kappa	Model size (MB)	Number of parameters
ResNet	72.90	0.57	85.32	22,366,673
CNN-BiGRU-Attention	75.95	0.63	0.199	51,089

Table 4. Comparative analysis of models' performance across all folds.

without the attention layer, without the CNN layer, without the BiGRU layer, and without both the BiGRU and attention layers. The average metrics across all threefold test sets for each architecture are presented in Table 5. The results indicate that the omission of any component leads to a significant decline in the model's performance. This emphasizes the importance of the CNN layers for extracting spatial features, the BiGRU layers for capturing temporal dependencies, and the attention mechanism for highlighting relevant segments of the signal. These findings validate the effectiveness of the chosen architecture in estimating the AHI from SpO2 signals.

Discussion

In this study, we delved into the efficacy of leveraging a residual architecture and a combination of CNN and RNN architectures, augmented by an attention mechanism, to process SpO2 signals as 1D raw data for the estimation of AHI and the assessment of pediatric SAH severity. To the best of our knowledge, the application of these architectures in evaluating the severity of pediatric SAH using SpO2 signals constitutes a novel contribution. We employed a threefold cross-validation method to showcase the models' generalizability on the dataset. Remarkably, both models demonstrated commendable performance in terms of four-class accuracy and kappa values on each fold test set. Furthermore, the CNN-BiGRU-Attention model showcased a significant average four-class accuracy of 75.95% and an average four-class kappa of 0.63 across all folds, indicating its ability to

Model	Average RMSE	Average R ²	Average Four-class kappa	Average Four-class accuracy (%)
CNN-BiGRU-Attention	2.31	0.79	0.63	75.95
CNN-BiGRU	2.66	0.74	0.56	70.26
BiGRU-Attention	6.31	0.21	0.23	46.26
CNN -Attention	4.47	0.32	0.32	54.14
CNN	3.69	0.44	0.43	64.4

Table 5. Ablation study results for cnn-bigr-attention model layers. *RMSE* Root Mean Squared Error, *R²* R-squared.

accurately classify patients across various levels of SAH severity. As evidenced by the results in Table 4, this model achieved high accuracies of 89.24%, 91.24%, and 96.41% for AHI thresholds of 1, 5, and 10 e/h on the fold-2 test set, underscoring its high diagnostic capability to detect SAH patients, especially those in need of urgent.

treatment or at high health risk. Despite the promising results, both models tended to underestimate the AHI for subjects with severe SAH, leading to low sensitivities for AHI thresholds of 5 and 10 e/h. This observation might be attributed to the imbalance of subjects from each SAH severity group in the dataset; however, we addressed this by oversampling apneic SpO₂ signal segments during the training process. Additionally, it is crucial to acknowledge that the AHI estimated by both models is derived using the length of the SpO₂ signal as the total recording time, while AHI calculated in PSG tests is based on the total sleep time. Although we employed an additional linear regression model to mitigate this error by mapping AHI calculated by record time to AHI calculated by sleep time via the validation set, limitations may still exist in accurately estimating the AHI.

Table 6 provides an overview of previous studies dedicated to the analysis of pediatric SAH and OSA severity assessment. Hornero et al. employed a Multi-Layer Perceptron (MLP) network to estimate AHI from 3602 SpO₂ recordings, categorizing subjects into four OSA severity classes, achieving an overall accuracy of 54.7%¹⁵. Jiménez-García et al. (2020), in their work, utilized the AdaBoost algorithm for a 4-class classification of pediatric OSA, using features from both AF and SpO₂ signals across a dataset of 974 pediatric subjects, attaining a 4-class accuracy of 57.95%¹⁷. These studies, while significant in using ML algorithms, are contrasted by more recent research showing enhanced performance with DL algorithms. Recent studies have used CNN-based models in combination with the CHAT dataset to estimate AHI from ECG signals³⁰, SpO₂ signals^{27,28}, and a combination of SpO₂ and AF signals^{29,31}. In 2020, Vaquerizo-Villar et al. employed a 1D CNN model for assessing SAH severity from 746 SpO₂ signals in the CHAT dataset's Baseline and Follow-up parts, reaching a four-class accuracy of 67.15% and a kappa value of 0.31 in a test set of 246 subjects²⁷. Following this, in 2021, the same team applied a similar model to the CHAT dataset's baseline, follow-up, and nonrandomized parts, as well as the University of Chicago Medicine (UofC) and Burgos University Hospital (BUH) datasets, achieving a four-class accuracy of 72.8% and a kappa of 0.51 in a test set of 312 subjects from the CHAT dataset²⁸. In a different approach, Jiménez-García et al., in 2022, utilized a 2D CNN architecture to estimate pediatric OSA severity from SpO₂ and AF signals as raw 2D data, applying it to all parts of the CHAT dataset and the UofC dataset. They achieved a four-class accuracy of 72.55% and a kappa of 0.60 on the CHAT test set²⁹. In their subsequent work, they improved their results to a four-class accuracy of 74.51% and a kappa of 0.62 on the CHAT test set by utilizing a 2D CNN layer followed by a BiGRU layer³¹. García-Vicente et al., in 2023 focused on a 1D CNN model using ECG signals from the CHAT dataset for a similar purpose, attaining a four-class accuracy of 57.86% and a kappa of 0.37 with 299 test subjects³⁰. Our research, distinct from these studies, considered all types of SAH events, including CSA. As previously stated, the main reason for this consideration is that CSA is always associated with a lack of respiratory effort⁴⁸, posing a challenge in classifying CSA and OSA events based solely on the oxygen desaturation of SpO₂ signals without chest and abdominal movement signals. Furthermore, we opted not to use the nonrandomized part of the CHAT dataset due to its lack of crucial clinical information necessary for

Studies	Signals	Datasets	Train set (Dataset subjects/Train subjects)	Validation set (Dataset subjects/Validation subjects)	Test set (Dataset subjects/Test subjects)	Model	Four-class accuracy (%)	Four-class kappa
Hornero et al. (2017) ¹⁵	SpO ₂	UofC + 12 other pediatric sleep laboratories	UofC (4191/589)	-	UofC + 12 other pediatric sleep laboratories (4191/3602)	MLP	54.7	-
Jiménez-García et al. (2020) ¹⁷	SpO ₂ + AF	UofC	UofC (974/584)	-	UofC (974/390)	AdaBoost	57.95	0.39
Vaquerizo-Villar et al. (2020) ²⁷	SpO ₂	CHAT	CHAT Baseline and Follow up (746/400)	CHAT Baseline and Follow up (746/100)	CHAT Baseline and Follow up (746/246)	CNN	67.15	0.31
Vaquerizo-Villar et al. (2021) ²⁸	SpO ₂	CHAT + BUH + UofC	CHAT Baseline and Follow up (859/859)	CHAT Non-randomized (779/467) + BUH (578/347) + UofC (974/382)	CHAT Nonrandomized (779/312)	CNN	72.8	0.51
					BUH (578/231)		60.2	0.42
					UofC (974/392)		61	0.42
Jiménez-García et al. (2022) ²⁹	SpO ₂ + AF	CHAT + UofC	CHAT (1638/1006)	CHAT (1638/326) + UofC (974/584)	CHAT (1638/306)	CNN	72.55	0.60
					UofC (974/390)		61.79	0.44
García-Vicente et al. (2023) ³⁰	ECG	CHAT	CHAT (1638/988)	CHAT (1638/223)	CHAT (1638/299)	CNN	57.86	0.37
García-Vicente et al. (2023) ³¹	SpO ₂ + AF	CHAT + UofC	CHAT (1638/1006)	CHAT (1638/326) + UofC (974/584)	CHAT (1638/306)	CNN-BiGRU	74.51	0.62
					UofC (974/390)		62.31	0.44
This study	SpO ₂	CHAT	CHAT Baseline and Follow up (844/400)	CHAT Baseline and Follow up (844/84)	CHAT Baseline and Follow up (844/253)	ResNet	72.9	0.57
						CNN-BiGRU-Attention	75.95	0.63

Table 6. Performance comparison with previous studies. *CHAT* Childhood Adenotonsillectomy Trial, *UofC* University of Chicago, *BUH* Burgos University Hospital.

accurate labeling. Unfortunately, we were unable to obtain permission to access other private datasets. Despite these limitations, our research results can be more directly compared with Vaquerizo-Villar et al. (2020)²⁷ as we used the same dataset and the same proportion of data split for training, validation, and testing. Moreover, we implemented threefold cross-validation to demonstrate the generability of our models, and it is evident that our models, especially the CNN-BiGRU model, exhibited higher four-class accuracy and kappa values across all folds. In contrast to Jiménez-García et al.^{29,31}, our research focuses solely on the SpO2 signal, a single-channel source, aligning with practical scenarios and emphasizing cost-effectiveness. Unlike García-Vicente et al.³⁰, our model aimed to utilize only SpO2 signals specifically because they can be recorded by pulse oximeters, which are more comfortable for patients to use than electrocardiographs for assessing pediatric SAH severity assessment. Overall, although previous studies utilizing CNN-based models have demonstrated high performance in both SAH and OSA severity assessment, our approach highlights the potential of hybrid CNN-BiGRU models, incorporating residual and attention-based mechanisms, specifically for pediatric SAH severity assessment. While achieving high diagnostic accuracy across various AHI thresholds, our models underscore the difficulty in accurately estimating AHI for severe SAH cases, primarily due to imbalances in subjects from different severity groups. We also stress the significance of considering sleep duration in AHI estimation, suggesting avenues for further improvement, such as integrating contextual information, including sleep stage analysis.

Despite these challenges, our study offers valuable insights into the use of residual architecture and attention-based hybrid CNN-RNN architecture for pediatric SAH assessment, setting a precedent for future developments in this area. In terms of data augmentation, we replicated apneic signal segments. For future research, additional techniques like the overlapping segmentation of signals, as utilized by García-Vicente et al.³⁰ and Jiménez-García et al.^{29,31}, could be adopted for data augmentation and to balance the dataset. Such strategies may also help mitigate overfitting and enhance the generalization capabilities of sophisticated models like ResNet. For future works, we recommend a two-stage approach using two DL models in parallel: one for sleep staging classification and the other for apneic event number estimation. PPG signals, as used in previous studies for sleep staging and estimating total sleep time^{21,49}, can be suitable for this purpose. Concurrently, SpO2 signals derived from PPG can be used to estimate the number of apneic events through a regression model. This dual-model strategy should also encompass the refinement of AHI estimation by incorporating the quantification of apnea events from SpO2 signals. Moreover, we suggest conducting experiments to ascertain whether demographic features such as age and gender correlate with signal characteristics, which could then inform feature embedding during model training. We utilized a moving average filter for noise reduction and signal smoothing, following common practices in prior studies. However, exploring alternative filtering methods could yield signals of higher quality. Additionally, our algorithm for zero-level artifact removal, inspired by previous researches^{9,35}, could be enhanced through further investigation to more effectively address artifacts. Future research may explore diverse filtering techniques to further improve signal fidelity.

Data availability

The following statement have been added to the paper: The datasets analyzed during the current study are available in the Childhood Adenotonsillectomy Trial repository at [<https://www.sleepdata.org/datasets/chat>] upon request from the National Sleep Research Resource (NSRR).

Received: 7 February 2024; Accepted: 15 July 2024

Published online: 01 October 2024

References

- Marcus, C. L. et al. Diagnosis and management of childhood obstructive sleep apnea syndrome. *Pediatrics* **130**(3), e714–e755 (2012).
- Javaheri, S. et al. Sleep apnea: Types, mechanisms, and clinical cardiovascular consequences. *J. Am. Coll. Cardiol.* **69**(7), 841–858 (2017).
- Shokouejad, M. et al. Sleep apnea: A review of diagnostic sensors, algorithms, and therapies. *Physiol. Meas.* **38**(9), R204 (2017).
- Berry, R. B. et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events: Deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine. *J. Clin. Sleep Med.* **8**(5), 597–619 (2012).
- Church, G. D. "The role of polysomnography in diagnosing and treating obstructive sleep apnea in pediatric patients. *Curr. Problems Pediatr. Adolesc. Health Care* **42**(1), 2–25 (2012).
- Tan, H.-L., Gozal, D., Ramirez, H. M., Bandla, H. P. & Kheirandish-Gozal, L. Overnight polysomnography versus respiratory polygraphy in the diagnosis of pediatric obstructive sleep apnea. *Sleep* **37**(2), 255–260 (2014).
- Nixon, G. M. et al. Planning adenotonsillectomy in children with obstructive sleep apnea: The role of overnight oximetry. *Pediatrics* **113**(1), e19–e25 (2004).
- Chan, E. D., Chan, M. M. & Chan, M. M. Pulse oximetry: Understanding its basic principles facilitates appreciation of its limitations. *Respir. Med.* **107**(6), 789–799 (2013).
- Koley, B. L. & Dey, D. On-line detection of apnea/hypopnea events using SpO2 signal: A rule-based approach employing binary classifier models. *IEEE J. Biomed. Health Inform.* **18**(1), 231–239 (2013).
- Moret-Bonillo, V., Alvarez-Estévez, D., Fernández-Leal, A. & Hernández-Pereira, E. Intelligent approach for analysis of respiratory signals and oxygen saturation in the sleep apnea/hypopnea syndrome. *Open Med. Inform. J.* **8**, 1 (2014).
- Gutiérrez-Tobal, G. C. et al. Reliability of machine learning to diagnose pediatric obstructive sleep apnea: Systematic review and meta-analysis. *Pediatr. Pulmonol.* **57**(8), 1931–1943 (2022).
- Gutiérrez-Tobal, G. C. et al. Diagnosis of pediatric obstructive sleep apnea: Preliminary findings using automatic analysis of airflow and oximetry recordings obtained at patients' home. *Biomed. Signal Process. Control* **18**, 401–407 (2015).
- Calderón, J. M., Álvarez-Pitti, J., Cuenca, I., Ponce, F. & Redon, P. Development of a minimally invasive screening tool to identify obese pediatric population at risk of obstructive sleep apnea/hypopnea syndrome. *Bioengineering* **7**(4), 131 (2020).
- Crespo, A. et al. Multiscale entropy analysis of unattended oximetric recordings to assist in the screening of paediatric sleep apnoea at home. *Entropy* **19**(6), 284 (2017).

15. Hornero, R. *et al.* Nocturnal oximetry-based evaluation of habitually snoring children. *Am. J. Respir. Crit. Care Med.* **196**(12), 1591–1598 (2017).
16. Barroso-García, V. *et al.* Usefulness of recurrence plots from airflow recordings to aid in paediatric sleep apnoea diagnosis. *Comput. Methods Programs Biomed.* **183**, 105083 (2020).
17. Jiménez-García, J. *et al.* Assessment of airflow and oximetry signals to detect pediatric sleep apnea-hypopnea syndrome using AdaBoost. *Entropy* **22**(6), 670 (2020).
18. Mostafa, S. S., Mendonça, F., Ravelo-García, A. G. & Morgado-Dias, F. A systematic review of detecting sleep apnea using deep learning. *Sensors* **19**(22), 4934 (2019).
19. Ma, C. *et al.* PPG-based continuous BP waveform estimation using polarized attention-guided conditional adversarial learning model. *IEEE J. Biomed. Health Inform.* <https://doi.org/10.1109/JBHI.2023.3324319> (2023).
20. Ma, C. *et al.* KD-Informer: Cuff-less continuous blood pressure waveform estimation approach based on single photoplethysmography. *IEEE J. Biomed. Health Inform.* <https://doi.org/10.1109/JBHI.2022.3181328> (2022).
21. Casal, R., Di Persia, L. E. & Schlotthauer, G. Temporal convolutional networks and transformers for classifying the sleep stage in awake or asleep using pulse oximetry signals. *J. Computat. Sci.* **59**, 101544 (2022).
22. Weng, W.-H. *et al.* Predicting cardiovascular disease risk using photoplethysmography and deep learning. *PLOS Glob. Public Health* **4**(6), e0003204 (2024).
23. Zarei, A., Beheshti, H. & Asl, B. M. Detection of sleep apnea using deep neural networks and single-lead ECG signals. *Biomed. Signal Process. Control* **71**, 103125 (2022).
24. Jothi, E. S. J., Anitha, J. & Hemanth, D. J. A photoplethysmography-based diagnostic support system for obstructive sleep apnea using deep learning approaches. *Comput. Electr. Eng.* **102**, 108279 (2022).
25. Wei, K., Zou, L., Liu, G. & Wang, C. MS-Net: Sleep apnea detection in PPG using multi-scale block and shadow module one-dimensional convolutional neural network. *Comput. Biol. Med.* **155**, 106469 (2023).
26. Vaquerizo-Villar, F. *et al.* Convolutional neural networks to detect pediatric apnea-hypopnea events from oximetry. In *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (ed. Vaquerizo-Villar, F.) 3555–3558 (IEEE, 2019).
27. Vaquerizo-Villar, F. *et al.* Automatic assessment of pediatric sleep apnea severity using overnight oximetry and convolutional neural networks. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (ed. Vaquerizo-Villar, F.) 633–636 (IEEE, 2020).
28. Vaquerizo-Villar, F. *et al.* A convolutional neural network architecture to enhance oximetry ability to diagnose pediatric obstructive sleep apnea. *IEEE J. Biomed. Health Inform.* **25**(8), 2906–2916 (2021).
29. Jiménez-García, J. *et al.* A 2D convolutional neural network to detect sleep apnea in children using airflow and oximetry. *Comput. Biol. Med.* **147**, 105784 (2022).
30. García-Vicente, C. *et al.* ECG-based convolutional neural network in pediatric obstructive sleep apnea diagnosis. *Comput. Biol. Med.* **167**, 107628 (2023).
31. Jiménez-García, J. *et al.* An explainable deep-learning architecture for pediatric sleep apnea identification from overnight airflow and oximetry signals. *Biomed. Signal Process. Control* **87**, 105490 (2024).
32. Marcus, C. L. *et al.* A randomized trial of adenotonsillectomy for childhood sleep apnea. *N. Engl. J. Med.* **368**(25), 2366–2376 (2013).
33. Zhang, G.-Q. *et al.* The national sleep research resource: Towards a sleep data commons. *J. Am. Med. Inform. Assoc.* **25**(10), 1351–1358 (2018).
34. Brouillette, R. T. *et al.* Nocturnal pulse oximetry as an abbreviated testing modality for pediatric obstructive sleep apnea. *Pediatrics* **105**(2), 405–412 (2000).
35. Deviaene, M. *et al.* Automatic screening of sleep apnea patients based on the spo 2 signal. *IEEE J. Biomed. Health Inform.* **23**(2), 607–617 (2018).
36. He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016).
37. Eom, H. *et al.* End-to-end deep learning architecture for continuous blood pressure estimation using attention mechanism. *Sensors* **20**(8), 2338 (2020).
38. Niu, Z., Zhong, G. & Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **452**, 48–62 (2021).
39. Bengio, Y., Simard, P. & Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **5**(2), 157–166 (1994).
40. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Computat.* **9**(8), 1735–1780 (1997).
41. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. Preprint at <https://arxiv.org/quant-ph/1409.1259> (2014).
42. Bahdanau, D., Cho, K., Bengio, Y. Neural machine translation by jointly learning to align and translate. Preprint at <https://arxiv.org/quant-ph/1409.0473> (2014).
43. He, K., Zhang, X., Ren, S., Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proc. IEEE International Conference on Computer Vision*, pp. 1026–1034 (2015).
44. Kingma, D. P., Ba, J. Adam: A method for stochastic optimization. Preprint at <https://arxiv.org/quant-ph/1412.6980> (2014).
45. Huber, P. J. Robust estimation of a location parameter. In *Breakthroughs in Statistics: Methodology and Distribution* (ed. Huber, P. J.) 492–518 (Springer, 1992).
46. Bland, J. M. & Altman, D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **327**(8476), 307–310 (1986).
47. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**(1), 37–46 (1960).
48. Randerath, W. *et al.* Definition, discrimination, diagnosis and treatment of central breathing disturbances during sleep. *Eur. Respir. J.* **49**(1), 1600959 (2017).
49. Vaquerizo-Villar, F. *et al.* A deep learning model based on the combination of convolutional and recurrent neural networks to enhance pulse oximetry ability to classify sleep stages in children with sleep apnea. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (ed. Vaquerizo-Villar, F.) 1–4 (IEEE, 2023).

Author contributions

Erfan Mortazavi: Formal analysis, Writing—original draft. Bahram Tarvirdizadeh: Methodology, Project administration. Khalil Alipour: Conceptualization, Project administration. Mohammad Ghamari: Supervision, Project administration.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024