



Published in final edited form as:

Proc Mach Learn Res. 2023 August ; 219: 167–185.

Bringing At-home Pediatric Sleep Apnea Testing Closer to Reality: A Multi-modal Transformer Approach

Hamed Fayyaz,

University of Delaware

Abigail Strang,

Nemours Children's Health

Rahmatollah Beheshti

University of Delaware

Abstract

Sleep apnea in children is a major health problem affecting one to five percent of children (in the US). If not treated in a timely manner, it can also lead to other physical and mental health issues. Pediatric sleep apnea has different clinical causes and characteristics than adults. Despite a large group of studies dedicated to studying adult apnea, pediatric sleep apnea has been studied in a much less limited fashion. Relatedly, at-home sleep apnea testing tools and algorithmic methods for automatic detection of sleep apnea are widely present for adults, but not children. In this study, we target this gap by presenting a machine learning-based model for detecting apnea events from commonly collected sleep signals. We show that our method outperforms state-of-the-art methods across two public datasets, as determined by the F1-score and AUROC measures. Additionally, we show that using two of the signals that are easier to collect at home (ECG and SpO₂) can also achieve very competitive results, potentially addressing the concerns about collecting various sleep signals from children outside the clinic. Therefore, our study can greatly inform ongoing progress toward increasing the accessibility of pediatric sleep apnea testing and improving the timeliness of the treatment interventions¹.

1. Introduction

Obstructive sleep apnea hypopnea syndrome (OSAHS) in pediatric patients are breathing disorders during sleep that is characterized by recurring events of obstruction, usually bringing about sleep fragmentation, sporadic oxygen desaturation (hypoxemia), and excessive carbon dioxide in the bloodstream (hypercapnia) (Loughlin et al., 1996; Vaquerizo-Villar et al., 2020). It is estimated that 1% to 5% of children suffer from OSAHS with the peak prevalence at ages between 2 and 8 years (Kheirandish-Gozal and Gozal, 2012; Bixler et al., 2009; Marcus et al., 2012). While OSAHS affects subjects of all ages, from infants to the elderly, the clinical manifestations, predisposing factors, and patterns of sleep data in children are different from those in adults (Choi et al., 2010). Moreover, the

¹Our code is publicly available at: <https://github.com/healthylaife/Pediatric-Apnea-Detection>.

FAYYAZ@UDEL.EDU .

distinctive symptoms of OSAHS in children are scarce and require more attention (Gipson et al., 2019), making the diagnosis more challenging.

Polysomnography is the gold standard for diagnosing sleep-related breathing disorders, including apnea and hypopnea. Polysomnography refers to the process used to collect biological signals and parameters during sleep, which is generally performed in clinical lab settings and during the night (Somnus represents sleep in Latin). The purpose of polysomnography is to evaluate underlying causes of sleep disturbances (Rundo and Downey III, 2019). A polysomnogram commonly involves collecting signals including brain electrical activity (electroencephalogram or EEG), eye movements during sleep (electrooculogram or EOG), cardiac rate and rhythm (electrocardiogram or ECG), blood oxygen saturation (pulse oximetry or SpO₂), measurement of exhaled air to indirectly measure blood CO₂ (end-tidal carbon dioxide or ET-CO₂), respiratory effort in thorax and abdomen (respiratory inductance plethysmography or RIP), and nasal and oral airflow. Polysomnography is generally considered effective, however, presents many challenges, including complexity, cost, intrusiveness, and the need for intensive involvement of clinical providers (Spielmanns et al., 2019).

Considering polysomnography challenges, a fairly large family of studies has explored ways to offer more accessible ways for diagnosing apnea-hypopnea through home sleep apnea tests (HSATs) (Kirk et al., 2017), and consumer wearable devices (Khor et al., 2023). In this respect, ECG and SpO₂ signals have been frequently used for apnea-hypopnea detection and screening (Ramachandran and Karuppiah, 2021).

Similar to the successful application of artificial intelligence (AI) assistive tools in diagnosing other medical conditions, many recent studies have tried to develop AI tools to diagnose OSAHS without relying on full-blown polysomnography (Chen et al., 2022; John et al., 2021; Bozkurt et al., 2020; Zhao et al., 2021). While some of these studies have reported good performance in adults, very few studies have focused on children (discussed more in Section 2).

This study aims to address the above gaps in the detection of pediatric apnea and hypopnea. We present a new method for detecting OSAHS patterns in pediatric populations and then study the role of various modalities, commonly present in polysomnography, in the presented OSAHS detection. This way, the contributions of our study can be listed as follows:

- We present a customized transformer-based architecture for detecting OSAHS and use a novel data representation technique to handle polysomnography modalities. We show that our model receives state-of-the-art performance when compared to other baselines.
- Using two large public pediatric sleep datasets, we extensively study the role of different combinations of common modalities. We show that using only two easier-to-collect signals (i.e., ECG, SpO₂) can achieve close to maximum performance across different experiments.

Generalizable Insights about Machine Learning in the Context of Healthcare

Similar to many other clinical conditions, advancements in OSAHS diagnosis and treatment have been disproportionately greater in adults than children. At-home sleep testing is common for adults; several (US) FDA-approved testing packages exist; and payers generally cover them. However, for children (especially the younger ones), in-lab sleep testing through polysomnography (PSG) remains the only option available. In-lab sleep testing presents important challenges, such as economics (e.g., insurance deductible and parent time off work), access to care (e.g., long wait, distance to the facility, and the need for multiple study sessions/nights), social challenges (e.g., being a single parent), and handling children with special needs (especially those with developmental issues) in the new environment.

The present study is based on the primary hypothesis that it is possible to achieve adult-level performance in detecting OSAHS. We also study the same hypothesis using subsets of polysomnography data. To test our hypothesis, we design a customized AI model, informed by advances in studying apnea-hypopnea and other biomedical applications. We specifically focus on the feasibility aspects of at-home sleep testing. Through extensive experiments, we show that it is possible to achieve PSG-level performance using two of the easier-to-collect signals (i.e., ECG and SpO₂). While we focus on improving at-home sleep testing, our study can also inform in-lab sleep testing by offering a dedicated method for pediatric settings.

2. Related Work

In the context of studying apnea-hypopnea patterns, two tasks are commonly studied that can be performed manually by an expert or automatically. The first task is to detect the presence of apnea-hypopnea in a certain time interval, and the second task is to identify the severity of apnea-hypopnea, which is determined using the Apnea-Hypopnea Index (the number of apnea-hypopnea events per hour of sleep) (Berry et al., 2012). This work focuses on the detection of apnea-hypopnea, which is also the foundation for the estimation of the severity of apnea-hypopnea (Vaquerizo-Villar et al., 2020, 2022). Due to the challenges of the acquisition and processing of polysomnography signals, simpler alternatives have been utilized for apnea-hypopnea detection. The majority of existing work in the literature investigates the possibility of automatic apnea-hypopnea detection and severity prediction by using ECG and SpO₂ signals (Brouillette et al., 2000; Tan et al., 2014). These two signal types turn out to be the same two signal types that achieve PSG-level performance in our study, too.

Existing studies that use ECG signals generally use band-pass filters to reduce the noise sourcing from the baseline wander, muscle artifacts, power line interference, and other sources (Urtnasan et al., 2018; Bahrami and Forouzanfar, 2022). Since ECG signals contain complicated patterns, many studies have used extensive preprocessing steps in their pipeline for extracting features, including R-R peak intervals, R-wave amplitude, wavelet coefficients, and ECG-derived respiration features (Shen et al., 2021; Erdenebayar et al., 2019; Fatimah et al., 2020; Bozkurt et al., 2020). Using automatic feature extraction approaches, through deep neural networks and similar unsupervised methods, has been also a popular choice (Chang et al., 2020; Chen et al., 2022; Zarei et al., 2022; Feng

et al., 2020). Methods for detecting sleep apnea-hypopnea events from ECG signals are extensively reviewed by Salari et al. (2022).

Similar to ECGs, many studies have used a manual feature extraction process on SpO₂ signals followed by feeding the extracted features to a classification model (Álvarez et al., 2012; Morillo and Gross, 2013; Uçar et al., 2017; Morales et al., 2017; Hwang et al., 2017; Mostafa et al., 2017a). Extracted features mostly are (1) time-based measures, such as the oxygen desaturation index, (2) statistical features, including minimum, maximum, variance, and (3) frequency-domain features based on wavelet transformations. Raw SpO₂ signals fed to deep neural networks are also used for respiratory events detection (Mostafa et al., 2017b; John et al., 2021).

Prior studies have also used both ECG and SpO₂ signals to exploit information from multiple sources and to handle better the signals' imperfection and defects such as missing data or noise (Tuncer et al., 2019; Ravelo-García et al., 2015; Pathinarupothi et al., 2017; Xie and Minn, 2012). Most of these studies focus on adult cohorts, and pediatric apnea-hypopnea is still understudied. Moreover, many existing pipelines are designed to work on a specific set of modalities as input, and therefore, cannot be used in the absence of the input signals.

2.1. Transformers

Transformer-based architectures have been widely adopted in various AI applications, such as health informatics (Nerella et al., 2023; Poulain et al., 2022, 2021), computer vision (Dosovitskiy et al., 2020) and natural language processing (Devlin et al., 2018). They also form the building block of the well-known large language models such as Bloom (Scao et al., 2022) and GPT (Brown et al., 2020).

Among the studies that have used transformers to study sleep patterns, Phan et al. (2022) proposed a hierarchical architecture composed of transformers to perform automatic sleep staging using EEG signals. Also, Lee and Saeed (2022) have used transformers for pediatric sleep staging inspired by the visual transformers approach (Dosovitskiy et al., 2020). The only prior work that has used transformers for apnea-hypopnea detection (as far as the authors could find) relates to the hybrid architecture presented by Hu et al. (2022) that uses transformers, and convolutional neural networks (CNNs) for obstructive apnea detection. We have used this study as our fourth baseline in our experiments. For apnea detection, this study has a rather specific and narrowly defined requirement though. The method needs 6-min intervals and uses 2.5-min before, the middle 60-sec, and 2.5-min after that 60-second for detection. Our proposed method is based on a pure transformer architecture (i.e., no CNN or RNN modules are used) for apnea-hypopnea detection.

3. Problem Formulation

Consider a set of N patients P denoted by $\{P_n\}_{n=1}^N$. Patients can have polysomnography sleep study data S from more than one study session (night) m , denoted by $S_{n,m}$, and collectively shown by $\{S_{n,m}\}_{m=1}^{M_n}$. Here, M_n is the number of sleep studies belonging to the n -th patient. Each study can be divided into equal-length epochs (windows) e :

$$S_{n,m} = \{e_{n,m}^q\}_{q=1}^{Q_{n,m}}, \quad (1)$$

where $Q_{n,m}$ is the total number of epochs in the m -th study of the n -th patient and can be obtained by dividing the study by the desired duration for each epoch (\mathbb{L}):

$$Q_{n,m} = \left\lfloor \frac{\text{length}(S_{n,m})}{\mathbb{L}} \right\rfloor. \quad (2)$$

Each epoch may consist of different polysomnography signals, X . In our supervised setting, each epoch also comes with an integer value, denoted by Y , which shows the number of seconds that an apnea-hypopnea event overlapped with the epoch:

$$e_{n,m}^q = (X_{n,m}^{q,1}, X_{n,m}^{q,2}, \dots, X_{n,m}^{q,K_{n,m}}, Y_{n,m}^q), \quad (3)$$

where, $K_{n,m}$ is the number of different channels that are available in $S_{n,m}$. Similar to the method used in other studies (Mendonca et al., 2018), by considering a threshold for apnea-hypopnea classification, denoted by t_{overlap} , $Y_{n,m}^q$ can be translated into a binary label, y by:

$$y_{n,m}^q = \begin{cases} 0, & Y_{n,m}^q < t_{\text{overlap}}, \\ 1, & Y_{n,m}^q \geq t_{\text{overlap}}. \end{cases}$$

A trained model f (parameterized with θ) can be designed to detect the occurrence of apnea \hat{y} during a given epoch e using the extracted features from the available signals:

$$\hat{y}_{n,m}^q = f_{\theta}(X_{n,m}^{q,1}, X_{n,m}^{q,2}, \dots, X_{n,m}^{q,K_{n,m}}) \quad (4)$$

4. Method

Our proposed method consists of a customized pipeline based on the standard transformer architecture (Vaswani et al., 2017b). Our pipeline (shown in Figure 1) comprises five components: (1) data sources, (2) segmentor, (3) tokenizer, (4) transformer, and (5) multi-layer perceptron (MLP) head.

Segmentor and Tokenizer

The segmentor splits the signals into equal-length epochs and passes them to the tokenizer. The tokenizer receives multi-modal epochs gathered from different sources (e.g., data derived from PSG and patients' electronic health records), and generates a tokenized and

synchronized representation that can be passed to the transformer for the downstream tasks. To do so, we set a signal representation sampling rate, denoted by $f_{sampling}$, and re-sample all data modalities to $f_{sampling}$. As a result, the output of a tokenizer for epochs with \mathbb{L} second length is a set of time series, each comprising $f_{sampling} * \mathbb{L}$ data points. In this work, the tokenizer deal with three types of data: (1) regular time series, (2) irregular time series, and (3) tabular data. Evenly sampled time-series (EEG, ECG, SpO_2 , etc.) are re-sampled with the sampling frequency $f_{sampling}$. For the irregular time series (extracted R-R intervals from ECG and amplitude of R-peaks), interpolation (i.e., constructing new data points based on a discrete set of known data points) is applied to obtain regular time series with sampling frequency $f_{sampling}$. Tabular data (demographic data in our work) is added as a constant signal (repeat the value in every token) to other time series. Finally, the synchronized representation is divided into i equal-length tokens (t_1, t_2, \dots, t_i) and passed to the transformer layers.

Transformer

We use the encoder part of the standard transformer architecture (Vaswani et al., 2017a), as the basis of our transformer (the decoder part is not needed as it is generally used for generative tasks). Our transformer's encoder block comprises multi-head attention and a position-wise feed-forward network. It also includes residual and normalization layers. We stacked five encoder modules to form our transformer unit in the proposed architecture.

The multi-head attention module that we use has the form of scaled dot-product attention. The input to each head consists of the *query*, *keys* of dimension d , and *values*. The output of i -th head H_i is calculated as:

$$Attention(query_i, key, value) = softmax\left(\frac{query_i \times key_i^T}{\sqrt{d}}\right) value_i \quad (5)$$

To enable the model to jointly attend to information from different representation subspaces, attention heads are concatenated, followed by a fully connected layer to form the Multi-Head Attention:

$$Multi\ Head\ Attention(query, key, value) = concat(h_1, \dots, h_n) W^C, \quad (6)$$

where,

$$h_i = Attention(query \times W_i^Q, key \times W_i^K, value \times W_i^V),$$

and W^C , W_i^Q , W_i^K , W_i^V are learnable weights. Each encoder unit also has a position-wise fully connected network (FCN) which is applied to each position separately and identically. This network comprises two fully connected layers with a ReLU activation in between:

$$FCN(x) = ReLU(xW_1 + b_1)W_2 + b_2, \quad (7)$$

where, (W_1, b_1) and (W_2, b_2) are the learnable weights and biases for the first and second layers, respectively.

Multi-layer perceptron

Transformer layers extract features from raw signals. The last part of the architecture is a two-layer fully connected network that acts as a classifier and predicts the probability of an apnea-hypopnea event occurring in the provided epoch using the transformer output. The first and second layer has 256, and 128 neurons, respectively.

Loss function

We use binary cross-entropy as the loss function for our model:

$$\mathcal{L}(\theta) = \frac{1}{T} \sum_{n=1}^N \sum_{m=1}^{M_n} \sum_{q=1}^{Q_{n,m}} \mathbb{H}(y_{n,m}^q, f_{\theta}(y_{n,m}^q | X_{n,m}^q)) \quad (8)$$

where \mathbb{H} is binary cross entropy:

$$\mathbb{H}(y, \hat{y}) = -y \log(\hat{y}) - (1-y)(1 - \log(\hat{y})) \quad (9)$$

and,

$$T = \sum_{n=1}^N \sum_{m=1}^{M_n} Q_{n,m}. \quad (10)$$

In the equations 8 and 10, T denotes the total number of training samples. f_{θ} is the neural network function with learnable parameters θ .

5. Experiments

We study three major research questions in our experiments. These three include the following, Q1: how does the proposed method compare to existing methods in the literature? Q2: is there a subset of PSG signals that achieves comparable performance to the entire PSG signals? Q3: how does the presented model perform across different ages?

Datasets

In our experiments, we use two large public pediatric sleep datasets. These two include the Nationwide Children's Hospital (NCH) Sleep Data Bank (Lee et al., 2022), and Childhood Adenotonsillectomy Trial (CHAT) dataset (Marcus et al., 2013; Redline et al., 2011). Table 1 shows some characteristics of the datasets, and Table 2 shows additional information related to the frequent events and sleep stages of the datasets. We also present additional details for our preprocessing and cohort extraction steps in Appendix A.

NCH –

The first dataset (NCH) offers a large and free source that includes both polysomnography signals linked to patients' EHRs (electronic health records). The linked data includes demographics and longitudinal clinical data such as encounters, medication, measurements, diagnoses, and procedures. Measurements contain body mass index, body mass index percentile, and blood pressure. The dataset was collected between 2017 and 2019 at Nationwide Children's Hospital, Cleveland, Ohio, USA.

CHAT –

We also use recordings from the CHAT study, which is a multi-center, single-blind, randomized, controlled trial designed to analyze the efficacy of early removal of adenoids and tonsils (adenotonsillectomy) on children with mild to moderate obstructive apnea. Physiological measures of sleep were assessed at baseline and at seven months with standardized full PSG with central scoring at the Brigham and Women's Hospital. In total, 1,447 children had screening PSG, and 464 were randomized to treatment. We use the PSGs collected in the baseline in our work.

Baselines

We chose four state-of-the-art studies (on adult apnea detection) with different architectures and components to compare with the proposed model in this work.

CNN –

The first study by Chang et al. (2020) uses a network consisting of 10 CNN layers for feature extraction followed by four fully-connected layers for classification. They also apply batch normalization and dropout for better generalization and to avoid overfitting.

Fusion –

The second study by Chen et al. (2022) uses a lightweight multi-scaled fusion network with multiple CNN and channel-wise attention modules.

CNN+LSTM –

The third model presented by Zarei et al. (2022) uses an automatic feature extraction method developed by combining CNN and long short-term memory (LSTM) modules. A stack of fully connected layers at the end is used to classify the events.

Hybrid Transformer –

The fourth study by Hu et al. (2022) uses a CNN-based attention block followed by three transformer encoders to detect apnea. The attention block has three parallel two-layer CNNs with different kernel sizes to obtain multi-perspective feature representation.

Implementation details

We present detailed technical descriptions for training the models (including hyperparameter tuning and their effects) in Appendix B. We follow a customized procedure for cross-validation (also presented in Appendix B). Our procedure ensures that an equal number of patients are assigned to each fold, and the number of positive samples in each fold is kept similar to the others. We report the performance of the trained models using F1-score (harmonic mean of precision and recall) and AUROC (area under the receiver operating characteristic curve).

5.1. Q1: How does the model compare to baselines?—Table 3 shows the results obtained by our model versus the four baselines introduced above. Our model has achieved superior performance in comparison to the baselines. Adding demographic data to other modalities marginally improved the performance. Beyond studying discrimination power of our model, we also study the learned representations by the model and evaluate its calibration. The results related to these experiments are presented in Section C.1.

5.2. Q2: Is there a subset of PSG working like the whole?—PSG signals typically consist of 6 different types of signals (as defined in Section 1), including ECG, EEG, EOG, respiratory, SpO₂, and CO₂. EEG, EOG, and respiratory signals often include multiple channels. In the case of our two datasets (as well as most pediatric PSGs), the number of these channels is 6, 2, and 3, respectively. To study our research question, we compare the performance of our model when a subset of signal types is used. We start by using only 1 signal type, then 2, and 3. Table 4 shows how the model's performance changes when subsets of size 1 and 2 (versus all) of signal types are used. In Appendix C.2, we also report the results related to the subsets of size 3.

5.3. Q3: How does the performance of the model vary between different ages?—We separate patients according to their age range to study the performance of our model for different ages. The performance of our model and other state-of-the-art models for each age group in the NCH dataset is shown in Figure 2. One can observe that the model's discriminative performance remains consistently over 80%, while the performance is lower in younger ages.

6. Discussion

Working toward our goal of achieving PSG-level OSAHS identification performance using a subset of the PSG signals, we presented a new method to identify apnea-hypopnea events. We showed that our method achieves better classification performance when compared to several recent methods. We attribute the improved performance to the customized preprocessing steps and the customized transformer model that we used in our approach.

We also observe that using only PSG signals achieves comparable performance to PSG plus demographics data. This may offer an interesting area of further research, as prior work has reported a strong role of demographics (such as gender and race) in sleep apnea (Dudley and Patel, 2016; Lumeng and Chervin, 2008). It is also possible that PSG signals act as a proxy for demographic information. A natural future work would be studying the model performance across various demographic subgroups. This step would also help with analyzing the group fairness of our method, besides the discrimination and calibration analysis we reported here.

Through our second series of experiments, we studied the degree to which leaving out some of the PSG signals would necessitate trading off the diagnosis performance. While no single signal achieves PSG-level performance, using pairs of signals seemed to allow achieving this. Specifically, using ECG and SpO₂ achieved performance close to entire PSG signals. Besides the PSG subsets of sizes 1 to 3, we do not report the other combinations (19 others, related to the subsets of size 4 and 5 out of 6), as using large subsets offers a limited practical advantage over using all 6 PSG signals, and we observe competitive results (to using all 6 signals) with the subset of size 2. What is especially noticeable is that ECG and SpO₂ are also the two signals that are easier to collect outside the clinic (Randazzo et al., 2018). Existing consumer wearable technologies, such as wearable bands, often readily record these two signals. While wearable devices are known to be not reliable for clinical decisions (Stehling et al., 2017), they may still help patient-family-provider interactions (Burkart et al., 2021). One of the important limitations of our study, however, relates to the fact that the subset of signals we chose are derived from lab-collected PSG. PSG signals generally have higher reliability than signals collected outside of the clinic using home sleep apnea testing devices or generic wearable devices. In lieu of comparable annotated wearable data to our PSG data, we further study our method by manually injecting noise into the PSG data, to mimic the present noise in wearable devices (Figure 6, in Appendix C.1).

Another noticeable observation relates to the absence of EEG signals from the ideal signal subset. EEG plays a critical role in measuring apnea events in children, and in fact, missing EEG from at-home sleep tests (even level three tools) is considered a main barrier to adopting these tests (Light et al., 2018). In pediatric sleep, arousals are needed to score hypopneas in some cases and central apnea. Our study is the first to demonstrate that diagnosing apnea in children may be doable without EEG signals. Additional studies would be needed to investigate how leaving out EEG from the input signals can achieve competitive results using an opaque-box (deep learning) method. According to prior studies (Thorey et al., 2019), our automated method of apnea-hypopnea detection already surpasses human-level performance.

While PSG acts as a gold standard for diagnosing sleep apnea-hypopnea, given the many challenges that it faces, it is important to work toward improving the feasibility of at-home sleep testing. This is especially important for children, as a critical window is often considered key for maximizing the effects of interventions and preventing various potential sequela, such as poor growth, heart problems, and affecting child's behavior and cognition (Bonsignore et al., 2019). Currently, there exist a few FDA-approved home sleep apnea testing (HSAT) tools for older children (12 years and older) (Pang et al., 2007). The results

related to our third research question (Figure 2) showed that it is possible to achieve competitive results across different ages.

Two major guidelines related to pediatric home sleep apnea testing presently exist. One from the American Academy of Pediatrics (Marcus et al., 2012) states that: “if polysomnography is not available, then alternative diagnostic tests or referral to a specialist for more extensive evaluation may be considered.” The second from the American Academy of Sleep Medicine (Kirk et al., 2017) states that “[u]se of a home sleep apnea test is not recommended for the diagnosis of obstructive sleep apnea in children. The ultimate judgment regarding propriety of any specific care must be made by the clinician, in light of the individual circumstances presented by the patient, available diagnostic tools, accessible treatment options, and resources.” The Covid-19 pandemic especially highlighted the unique role of the patient’s context and clinician’s judgment. We view our study aligned with this latter note. Specifically, we view our method as informing home sleep apnea testing with the goal of being part of an overall approach for disease treatment (not as a standalone solution). Obviously, further studies would be needed to fully examine the utility and safety of such a solution.

Acknowledgments

Our study was supported by NIH awards, P20GM103446 and P20GM113125.

Appendix A.: Datasets and Preprocessing

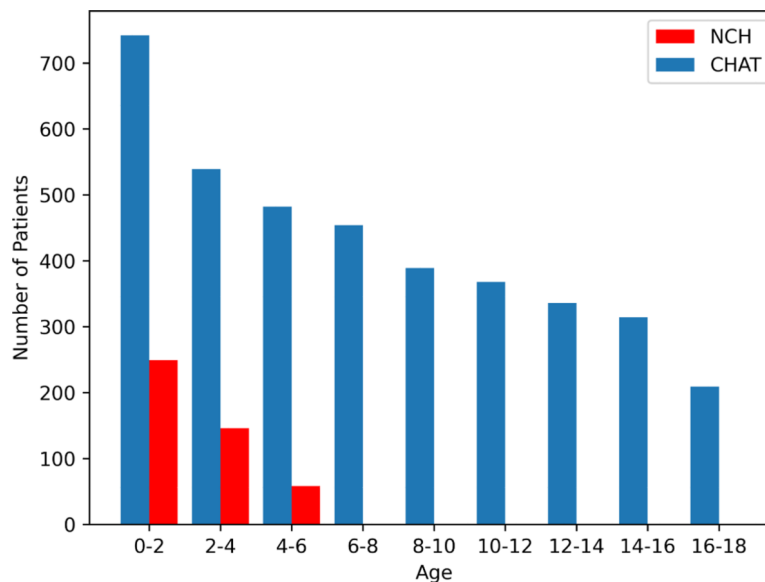


Figure 3:
Patients’ age distribution at the time of the sleep study

The patients’ age distribution is shown in Figure 3. We chose the modalities that are available in at least 70% of studies, namely, EOG, ECG, EEG, SpO₂, End-tidal CO₂, and respiratory signals. Specifically, we use C3-M2, C4-M1 channels from EEG signals. We

omitted studies that do not have all of these modalities to have an identical dataset for all experiments. In CHAT dataset, we used the baseline portion of recordings. Then, we segmented each study into 30 seconds epochs. We also discarded the epochs recorded during the time that the patient was not sleeping. Since studies are recorded with different sampling rates, we re-sampled all studies to $f_{sampling}$. In order to balance the dataset, we under-sampled the minority class. ECG signal was denoised using a band-pass filter with lower and upper cutoff frequencies of 3Hz and 45Hz. Hamilton R-peak detection method was utilized to extract R-R intervals and the amplitude of R-peaks from the ECG signal.

Appendix B.: Additional training and evaluation details

In experiments, we used Adam optimizer (Kingma and Ba, 2017) with a learning rate of 10^{-3} , $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-7}$. A batch size of 256 was used for training. Early stopping was applied and the training was stopped after 20 epochs without improvement in the validation loss. We use L2 weight regularization with $\lambda = 10^{-3}$ and incremental dropout to avoid overfitting. The dropout rate increased by 0.1 after each transformer unit. The model presented in this paper was implemented using Keras (Chollet et al., 2015) inside the TensorFlow (Abadi et al., 2016) framework.

B.1. Hyper-parameters tuning

We investigate the effect of changing hyper-parameters on the model performance. We ran the experiments with models comprised of 4, 6, and 8 layers of transformers with 4, and 6 heads in each layer. Each layer consists of two layers of MLP with (32, 64) and (64, 128) units. Patching may also affect the performance of our model. So, we repeated the experiments by dividing each 60-second epoch into 20, 30, and 60 patches to find the optimal number of patches. Increasing model capability leads to marginal improvement in results. The best performance was obtained with a model comprised of 8 transformers, each has 6 heads and accepting 20 patches for each input.

B.2. Customized Stratified Cross Validation

Applying a common five-fold cross-validation has a few shortcomings. First, more than one fold may have epochs from a specific patient. Since a patient recording has shared features, the model may learn the characteristics of patients' recordings, instead of focusing on fundamental features. So, the cross-validation should be done based on folds that there is no shared patient between them. It means that all the epochs from all sleep studies of a patient should end up in one fold. Second, as patients have different respiratory conditions, the numbers of respiratory event occurrences for each patient vary on a wide range. As a result, if patients are randomly grouped into n folds, the number of respiratory events in the folds is not close to each other. We propose a grouping algorithm, as shown in algorithm 1, tailored to address the drawbacks mentioned above. In this algorithm, we try to not only assign an equal number of patients to each fold but also maintain the number of positive samples in each fold similar to the others. To do this, we calculate the total duration of respiratory events for each patient. Then, we assign patients to folds in a way that the total length of the respiratory event in folds becomes equal to each other as much as possible.

Algorithm 1:

Stratified K-Fold Cross Validation

```

Input:  $\{P_n\}_{n=1}^N$ 
Output:  $\{fold_f\}_{n=1}^K$ 
for  $n \leftarrow 1$  to  $N$  do
   $P_i.score = 0$ 
  for  $m \leftarrow 1$  to  $M_n$  do
    for  $l \leftarrow 1$  to  $L_{n,m}$  do
       $P_i.score += length(E_{i,j}^k)$ 
    end
  end
end
score_sorted_list  $\leftarrow$  sort patients by score
for  $i \leftarrow 1$  to  $N$  do
   $f = i \bmod K$ 
   $fold_f \leftarrow$  score_sorted_list[i]
end

```

Appendix C.: Additional Results**C.1. Model performance**

We project 64-dimensional representations from the last transformer module output of the model, for test epochs, to 2 dimensions using t-SNE (Van der Maaten and Hinton, 2008). Visualization is shown in Figure 4. Each blue and red point represents one 30-second signal epoch with ‘Normal’ and ‘Apnea’ labels, respectively. As shown, the proposed model transforms epochs raw data to a latent space where normal and apnea samples are fairly separable. We also reported the calibration of our method in Figure 5. Besides, the performance of our model trained with ECG, and SpO₂ feeding with noisy signals is shown in Figure 6. Gaussian noise has been added to each signal regarding its power.

C.2. Combination of signal types

Performance of the models trained with every possible combination of three signals is shown in Table 5. In both datasets, among all possible combinations of triple signal groups, the one that includes ECG, and SpO₂ is the top performer. This confirms that the aforementioned signals are the best ones for detecting apnea-hypopnea even when we want to use more than two modalities for apnea-hypopnea detection.

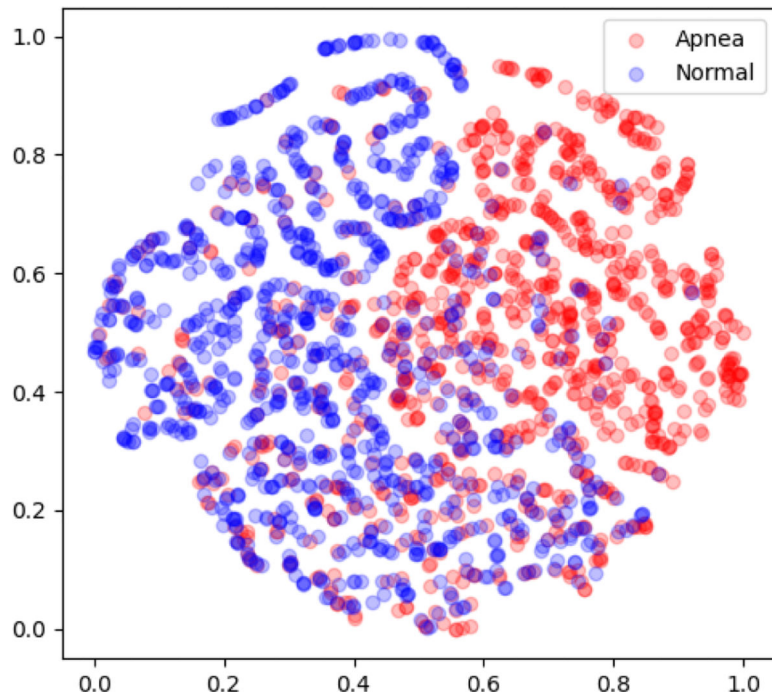


Figure 4: t-SNE visualization of representation learned with the output of the last transformer module in the proposed architecture

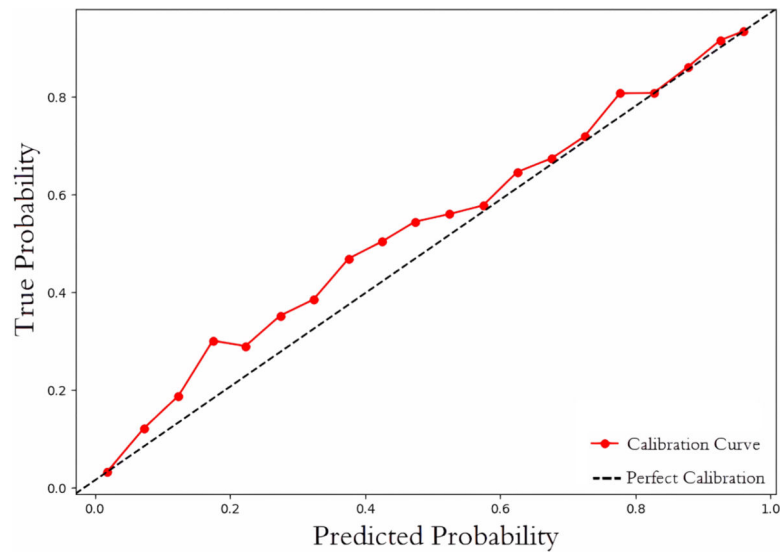


Figure 5: Model Calibration. The proposed model is calibrated in high probability. However, Its performance can be improved for low probabilities.

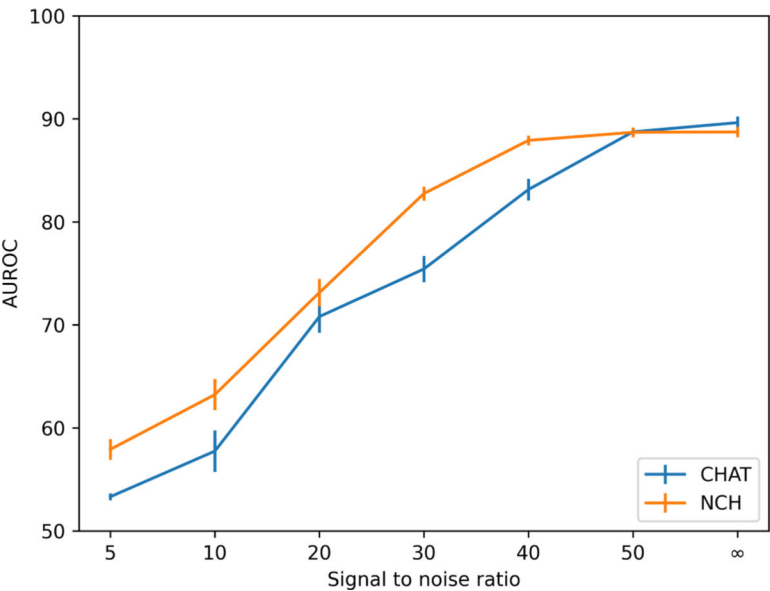


Figure 6:
AUROC of our proposed method dealing with noisy data

Table 5:

Results on trained models with three signals. The mean (standard deviation) values are shown.

EOG	EEG	ECG	Resp	SpO ₂	CO ₂	CHAT		NCH	
						F1	AUC	F1	AUC
✓	✓	✓				79.7(0.9)	85.8(0.7)	78.3(0.8)	84.3(0.5)
✓	✓		✓			80.2(0.7)	86.7(0.7)	80.5(0.9)	88.2(0.7)
✓	✓			✓		80.3(0.3)	87.8(0.6)	80.2(0.7)	88.1(0.5)
✓	✓				✓	78.4(1.3)	85.4(1.3)	77.6(1.3)	84.3(1.0)
✓		✓	✓			80.9(0.7)	87.6(0.5)	79.0(1.0)	86.9(0.7)
✓		✓		✓		83.1(0.5)	89.8(0.3)	80.7(1.0)	88.7(0.7)
✓		✓			✓	79.7(0.9)	86.7(0.9)	76.5(1.2)	83.6(1.0)
✓			✓	✓		82.3(1.0)	89.1(1.1)	80.3(1.3)	88.3(1.3)
✓			✓		✓	80.6(1.1)	87.5(1.0)	78.8(0.5)	86.8(0.7)
✓				✓	✓	81.5(0.4)	88.9(0.8)	80.3(0.7)	88.1(0.8)
	✓	✓	✓			80.4(0.7)	87.4(0.8)	77.8(1.1)	86.0(0.8)
	✓	✓		✓		82.5(0.3)	89.5(0.4)	81.1(0.5)	88.7(0.6)
	✓	✓			✓	79.8(1.3)	86.6(1.2)	75.8(1.1)	82.6(0.8)
	✓		✓	✓		82.0(0.9)	88.9(0.9)	81.1(0.7)	89.2(0.8)
	✓		✓		✓	80.0(1.6)	87.0(1.3)	77.7(1.0)	85.3(0.6)
	✓			✓	✓	80.7(0.6)	88.1(0.9)	80.3(0.9)	87.9(0.7)
		✓	✓	✓		81.7(0.8)	88.6(0.8)	81.7(0.4)	89.5(0.4)
		✓	✓		✓	79.9(1.0)	87.0(1.1)	77.3(0.3)	85.4(0.4)
		✓		✓	✓	82.0(1.4)	88.7(1.5)	81.1(0.6)	88.5(0.6)

EOG	EEG	ECG	Resp	SpO ₂	CO ₂	CHAT		NCH	
						F1	AUC	F1	AUC
			✓	✓	✓	81.1(1.1)	88.2(1.0)	79.7(0.9)	87.5(0.6)

References

- Abadi Martín, Barham Paul, Chen Jianmin, Chen Zhifeng, Davis Andy, Dean Jeffrey, Devin Matthieu, Ghemawat Sanjay, Irving Geoffrey, Isard Michael, et al. Tensorflow: A system for large-scale machine learning. In 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), pages 265–283, Savannah, GA, USA, 2016. ACM.
- Álvarez Daniel, Hornero Roberto, Marcos J Víctor, and Del Campo Félix. Feature selection from nocturnal oximetry using genetic algorithms to assist in obstructive sleep apnoea diagnosis. *Medical engineering & physics*, 34(8):1049–1057, 2012. [PubMed: 22154238]
- Bahrami Mahsa and Forouzanfar Mohamad. Sleep apnea detection from single-lead ecg: A comprehensive analysis of machine learning and deep learning algorithms. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022.
- Berry Richard B, Budhiraja Rohit, Gottlieb Daniel J, David Gozal, Iber Conrad, Kapur Vishesh K, Marcus Carole L, Mehra Reena, Parthasarathy Sairam, Quan Stuart F, 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. *Journal of clinical sleep medicine*, 8(5): 597–619, 2012. [PubMed: 23066376]
- Bixler Edward O, Vgontzas Alexandros N, Lin Hung-Mo, Liao Duanping, Calhoun Susan, Vela-Bueno Antonio, Fedok Fred, Vlasic Vukmir, and Graff Gavin. Sleep disordered breathing in children in a general population sample: prevalence and risk factors. *Sleep*, 32(6):731–736, 2009. [PubMed: 19544748]
- Bonsignore Maria R, Baiaomonte Pierpaolo, Mazzuca Emilia, Castrogiovanni Alessandra, and Marrone Oreste. Obstructive sleep apnea and comorbidities: a dangerous liaison. *Multidisciplinary respiratory medicine*, 14(1):1–12, 2019. [PubMed: 30651988]
- Bozkurt F, Uçar M Kür ad, Bozkurt M Recep, and Bilgin C. Detection of abnormal respiratory events with single channel ecg and hybrid machine learning model in patients with obstructive sleep apnea. *Irbm*, 41(5):241–251, 2020.
- Brouillette Robert T, Morielli Angela, Leimanis Andra, Waters Karen A, Luciano Rina, and Ducharme Francine M. Nocturnal pulse oximetry as an abbreviated testing modality for pediatric obstructive sleep apnea. *Pediatrics*, 105(2):405–412, 2000. [PubMed: 10654964]
- Brown Tom, Mann Benjamin, Ryder Nick, Subbiah Melanie, Kaplan Jared D, Dhariwal Prafulla, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Burkart Sarah, Beets Michael W, Armstrong Bridget, Hunt Ethan T, Dugger Roddrick, von Klinggraeff Lauren, Jones Alexis, Brown David E, and Weaver R Glenn. Comparison of multichannel and single-channel wrist-based devices with polysomnography to measure sleep in children and adolescents. *Journal of Clinical Sleep Medicine*, 17(4):645–652, 2021. [PubMed: 33174529]
- Chang Hung-Yu, Yeh Cheng-Yu, Lee Chung-Te, and Lin Chun-Cheng. A sleep apnea detection system based on a one-dimensional deep convolution neural network model using single-lead electrocardiogram. *Sensors*, 20(15):4157, 2020. [PubMed: 32722630]
- Chen Xianhui, Chen Ying, Ma Wenjun, Fan Xiaomao, and Li Ye. Toward sleep apnea detection with lightweight multi-scaled fusion network. *Knowledge-Based Systems*, 247: 108783, 2022.
- Choi Ji Ho, Kim Eun Joong, Choi June, Kwon Soon Young, Kim Tae Hoon, Lee Sang Hag, Lee Heung Man, Shin Choi, and Lee Seung Hoon. Obstructive sleep apnea syndrome: a child is not just a small adult. *Annals of Otology, Rhinology & Laryngology*, 119(10): 656–661, 2010. [PubMed: 21049849]
- Chollet François et al. Keras. <https://github.com/fchollet/keras>, 2015.

- Devlin Jacob, Chang Ming-Wei, Lee Kenton, and Toutanova Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
- Dosovitskiy Alexey, Beyer Lucas, Kolesnikov Alexander, Weissenborn Dirk, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain, et al. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
- Dudley Katherine A. and Patel Sanjay R.. Disparities and genetic risk factors in obstructive sleep apnea. *Sleep Medicine*, 18:96–102, 2016. ISSN 1389–9457. doi: 10.1016/j.sleep.2015.01.015. URL <https://www.sciencedirect.com/science/article/pii/S1389945715000623>. NHLBI Workshop on Reducing Health Disparities: The Role of Sleep Deficiency and Sleep Disorders. [PubMed: 26428843]
- Erdenebayar Urtnasan, Kim Yoon Ji, Park Jong-Uk, Joo Eun Yeon, and Lee Kyoung-Joung. Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram. *Computer methods and programs in biomedicine*, 180:105001, 2019. [PubMed: 31421606]
- Fatimah Binish, Singh Pushpendra, Singhal Amit, and Pachori Ram Bilas. Detection of apnea events from ecg segments using fourier decomposition method. *Biomedical Signal Processing and Control*, 61:102005, 2020.
- Feng Kaicheng, Qin Hengji, Wu Shan, Pan Weifeng, and Liu Guanzheng. A sleep apnea detection method based on unsupervised feature learning and single-lead electrocardiogram. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12, 2020. [PubMed: 33776080]
- Gipson Kevin, Lu Mengdi, and Kinane T Bernard. Sleep-disordered breathing in children. *Pediatrics in review*, 40(1):3, 2019.
- Hu Shuaicong, Cai Wenjie, Gao Tijie, and Wang Mingjie. A hybrid transformer model for obstructive sleep apnea detection based on self-attention mechanism using single-lead ecg. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022.
- Hwang Su Hwan, Cho Jae Geol, Choi Byung Hun, Baek Hyun Jae, Lee Yu Jin, Jeong Do-Un, Park Kwang Suk, et al. Real-time automatic apneic event detection using nocturnal pulse oximetry. *IEEE Transactions on Biomedical Engineering*, 65(3):706–712, 2017. [PubMed: 28622666]
- John Arlene, Nundy Koushik Kumar, Barry Cardiff, and John Deepu. Somnnet: An spo2 based deep learning network for sleep apnea detection in smartwatches. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pages 1961–1964. IEEE, 2021.
- Kheirandish-Gozal Leila and Gozal David. Sleep disordered breathing in children: a comprehensive clinical guide to evaluation and treatment. 2012.
- Khor Yet H, Khung Su-Wei, Ruehland Warren R, Jiao Yuxin, Lew Jeremy, Munsif Maitri, Ng Yvonne, Ridgers Anna, Schulte Max, Seow Daniel, et al. Portable evaluation of obstructive sleep apnea in adults: A systematic review. *Sleep Medicine Reviews*, page 101743, 2023. [PubMed: 36657366]
- Kingma Diederik P. and Ba Jimmy. Adam: A method for stochastic optimization, 2017.
- Kirk Valerie, Baughn Julie, D’Andrea Lynn, Friedman Norman, Galion Anjalee, Garetz Susan, Hassan Fauziya, Wrede Joanna, Harrod Christopher G, and Malhotra Raman K. American academy of sleep medicine position paper for the use of a home sleep apnea test for the diagnosis of osa in children. *Journal of Clinical Sleep Medicine*, 13(10):1199–1203, 2017. [PubMed: 28877820]
- Lee Harlin and Saeed Aaqib. Pediatric sleep scoring in-the-wild from millions of multichannel eeg signals. arXiv preprint arXiv:2207.06921, 2022.
- Lee Harlin, Li Boyue, DeForte Shelly, Splaingard Mark L, Huang Yungui, Chi Yuejie, and Linwood Simon L. A large collection of real-world pediatric sleep studies. *Scientific Data*, 9(1):1–12, 2022. [PubMed: 35013360]
- Light Matthew P, Casimire Thalia N, Chua Catherine, Koushyk Viachaslau, Burschtin Omar E, Ayappa Indu, and Rapoport David M. Addition of frontal eeg to adult home sleep apnea testing: does a more accurate determination of sleep time make a difference? *Sleep and Breathing*, 22:1179–1188, 2018. [PubMed: 30311183]
- Loughlin GM, Brouillette RT, Brooke LJ, Carroll JL, Chipps BE, England SJ, Ferber P, Ferraro NF, Gaultier C, Givan DC, et al. Standards and indications for cardiopulmonary sleep studies

in children. American journal of respiratory and critical care medicine, 153 (2):866–878, 1996. [PubMed: 8564147]

- Lumeng Julie C. and Chervin Ronald D.. Epidemiology of pediatric obstructive sleep apnea. Proceedings of the American Thoracic Society, 5(2):242–252, 2008. doi: 10.1513/pats.200708-135MG. URL 10.1513/pats.200708-135MG. [PubMed: 18250218]
- Marcus Carole L, Brooks Lee J, Ward Sally Davidson, Draper Kari A, Gozal David, Halbower Ann C, Jones Jacqueline, Lehmann Christopher, Schechter Michael S, Sheldon Stephen, et al. Diagnosis and management of childhood obstructive sleep apnea syndrome. Pediatrics, 130(3):e714–e755, 2012. [PubMed: 22926176]
- Marcus Carole L, Moore René H, Rosen Carol L, Giordani Bruno, Garetz Susan L, Taylor H Gerry, Mitchell Ron B, Amin Raouf, Katz Eliot S, Arens Raanan, et al. A randomized trial of adenotonsillectomy for childhood sleep apnea. N Engl J Med, 368:2366–2376, 2013. [PubMed: 23692173]
- Mendonca Fabio, Mostafa Sheikh Shanawaz, Ravelo-Garcia Antonio G, Morgado-Dias Fernando, and Penzel Thomas. A review of obstructive sleep apnea detection approaches. IEEE journal of biomedical and health informatics, 23(2):825–837, 2018. [PubMed: 29993672]
- Morales John F, Varon Carolina, Deviaene Margot, Borzée Pascal, Testelmans Dries, Buyse Bertien, and Van Huffel Sabine. Sleep apnea hypopnea syndrome classification in spo 2 signals using wavelet decomposition and phase space reconstruction. In 2017 IEEE 14th international conference on wearable and implantable body sensor networks (BSN), pages 43–46. IEEE, 2017.
- Morillo Daniel Sánchez and Gross Nicole. Probabilistic neural network approach for the detection of sahs from overnight pulse oximetry. Medical & biological engineering & computing, 51(3):305–315, 2013. [PubMed: 23160897]
- Mostafa Sheikh Shanawaz, Carvalho Joao Paulo, Morgado-Dias Fernando, and Ravelo-Garcia Antonio. Optimization of sleep apnea detection using spo2 and ann. In 2017 XXVI international conference on information, communication and automation technologies (ICAT), pages 1–6. IEEE, 2017a.
- Mostafa Sheikh Shanawaz, Mendonça Fábio, Morgado-Dias Fernando, and Ravelo-García Antonio. Spo2 based sleep apnea detection using deep learning. In 2017 IEEE 21st international conference on intelligent engineering systems (INES), pages 000091–000096. IEEE, 2017b.
- Nerella Subhash, Bandyopadhyay Sabyasachi, Zhang Jiaqing, Contreras Miguel, Siegel Scott, Bumin Aysegul, Silva Brandon, Sena Jessica, Shickel Benjamin, Bihorac Azra, Khezeli Kia, and Rashidi Parisa. Transformers in healthcare: A survey, 2023.
- Pang Kenny P, Gourin Christine G, and Terris David J. A comparison of polysomnography and the watchpat in the diagnosis of obstructive sleep apnea. Otolaryngology-Head and Neck Surgery, 137(4):665–668, 2007. [PubMed: 17903588]
- Pathinarupothi Rahul Krishnan, Rangan Ekanath Srihari, Gopalakrishnan EA, Vinaykumar R, Soman KP, et al. Single sensor techniques for sleep apnea diagnosis using deep learning. In 2017 IEEE international conference on healthcare informatics (ICHI), pages 524–529. IEEE, 2017.
- Phan Huy, Mikkelsen Kaare, Chén Oliver Y, Koch Philipp, Mertins Alfred, and De Vos Maarten. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. IEEE Transactions on Biomedical Engineering, 69(8):2456–2467, 2022. [PubMed: 35100107]
- Poulain Raphael, Gupta Mehak, Foraker Randi, and Beheshti Rahmatollah. Transformer-based multi-target regression on electronic health records for primordial prevention of cardiovascular disease. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pages 726–731. IEEE, 2021.
- Poulain Raphael, Gupta Mehak, and Beheshti Rahmatollah. Few-shot learning with semi-supervised transformers for electronic health records. In Machine Learning for Healthcare Conference, pages 853–873. PMLR, 2022.
- Ramachandran Anita and Karuppiiah Anupama. A survey on recent advances in machine learning based sleep apnea detection systems. In Healthcare, volume 9, page 914. MDPI, 2021. [PubMed: 34356293]
- Randazzo Vincenzo, Pasero Eros, and Navaretti Silvio. Vital-ecg: A portable wearable hospital. In 2018 IEEE Sensors Applications Symposium (SAS), pages 1–6. IEEE, 2018.

- Ravelo-García Antonio G, Kraemer Jan F, Navarro-Mesa Juan L, Hernández-Pérez Eduardo, Navarro-Esteve Javier, Juliá-Serdá Gabriel, Penzel Thomas, and Wessel Niels. Oxygen saturation and rr intervals feature selection for sleep apnea detection. *Entropy*, 17(5):2932–2957, 2015.
- Redline Susan, Amin Raouf, Beebe Dean, Chervin Ronald D, Garetz Susan L, Giordani Bruno, Marcus Carole L, Moore Renee H, Rosen Carol L, Arens Raanan, et al. The childhood adenotonsillectomy trial (chat): rationale, design, and challenges of a randomized controlled trial evaluating a standard surgical procedure in a pediatric population. *Sleep*, 34(11):1509–1517, 2011. [PubMed: 22043122]
- Rundo Jessica Vensel and Downey Ralph III. Polysomnography. *Handbook of clinical neurology*, 160:381–392, 2019. [PubMed: 31277862]
- Salari Nader, Hosseini-Far Amin, Mohammadi Masoud, Ghasemi Hooman, Khazaie Habibollah, Daneshkhah Alireza, and Ahmadi Arash. Detection of sleep apnea using machine learning algorithms based on ecg signals: A comprehensive systematic review. *Expert Systems with Applications*, 187:115950, 2022.
- Le Scao Teven, Fan Angela, Akiki Christopher, Pavlick Ellie, Ili Suzana, Hesslow Daniel, Castagné Roman, Luccioni Alexandra Sasha, Yvon François, Gallé Matthias, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.
- Shen Qi, Qin Hengji, Wei Keming, and Liu Guanzheng. Multiscale deep neural network for obstructive sleep apnea detection using rr interval from single-lead ecg signal. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021. [PubMed: 33776080]
- Spielmanns Marc, Bost David, Windisch Wolfram, Alter Peter, Greulich Tim, Nell Christoph, Storre Jan Henrik, Koczulla Andreas Rembert, and Boeselt Tobias. Measuring sleep quality and efficiency with an activity monitoring device in comparison to polysomnography. *Journal of clinical medicine research*, 11(12):825, 2019. [PubMed: 31803327]
- Stehling Florian, Keull Judith, Olivier Margarete, Große-Onnebrink Jörg, Mellies Uwe, and Stuck Boris A. Validation of the screening tool apnealink® in comparison to polysomnography for the diagnosis of sleep-disordered breathing in children and adolescents. *Sleep medicine*, 37:13–18, 2017. [PubMed: 28899523]
- Tan Hui-Leng, Gozal David, Ramirez Helena Molero, Bandla Hari P. R., and Kheirandish-Gozal Leila. Overnight Polysomnography versus Respiratory Polygraphy in the Diagnosis of Pediatric Obstructive Sleep Apnea. *Sleep*, 37(2):255–260, 02 2014. ISSN 0161–8105. doi: 10.5665/sleep.3392. [PubMed: 24497654]
- Thorey Valentin, Hernandez Albert Bou, Arnal Pierrick J, and During Emmanuel H. Ai vs humans for the diagnosis of sleep apnea. In 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pages 1596–1600. IEEE, 2019.
- Tuncer Seda Arslan, Akilotu Beyza, and Toraman Suat. A deep learning-based decision support system for diagnosis of osas using ptt signals. *Medical hypotheses*, 127:15–22, 2019. [PubMed: 31088639]
- Kür ad Uçar Muhammed, Bozkurt Mehmet Recep, Bilgin Cahit, and Polat Kemal. Automatic detection of respiratory arrests in osa patients using ppg and machine learning techniques. *Neural Computing and Applications*, 28(10):2931–2945, 2017.
- Urtnasan Erdenebayar, Park Jong-Uk, Joo Eun-Yeon, and Lee Kyoung-Joung. Automated detection of obstructive sleep apnea events from a single-lead electrocardiogram using a convolutional neural network. *Journal of medical systems*, 42(6):1–8, 2018.
- Van der Maaten Laurens and Hinton Geoffrey. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- Vaquerizo-Villar Fernando, Álvarez Daniel, Kheirandish-Gozal Leila, Gutiérrez-Tobal Gonzalo C, Gómez-Pilar Javier, Crespo Andrea, Del Campo Felix, Gozal David, and Hornero Roberto. 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), pages 633–636. IEEE, 2020.
- Vaquerizo-Villar Fernando, Álvarez Daniel, Gutiérrez-Tobal Gonzalo C, Del Campo Félix, Kheirandish-Gozal Leila, Gozal David, Penzel Thomas, and Hornero Roberto. A convolutional neural network to classify sleep stages in pediatric sleep apnea from pulse oximetry signals. In

2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON), pages 108–113. IEEE, 2022.

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Łukasz, and Polosukhin Illia. Attention is all you need. In Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, and Garnett R, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017a.

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Kaiser Łukasz, and Polosukhin Illia. Attention is all you need. *Advances in neural information processing systems*, 30, 2017b.

Xie Baile and Minn Hlaing. Real-time sleep apnea detection by classifier combination. *IEEE Transactions on information technology in biomedicine*, 16(3):469–477, 2012. [PubMed: 22353404]

Zarei Asghar, Beheshti Hossein, and Asl Babak Mohammadzadeh. Detection of sleep apnea using deep neural networks and single-lead ecg signals. *Biomedical Signal Processing and Control*, 71:103125, 2022.

Zhao Xiaoyun, Wang Xiaohong, Yang Tianshun, Ji Siyu, Wang Huiquan, Wang Jinhai, Wang Yao, and Wu Qi. Classification of sleep apnea based on eeg sub-band signal characteristics. *Scientific Reports*, 11(1):1–11, 2021. [PubMed: 33414495]

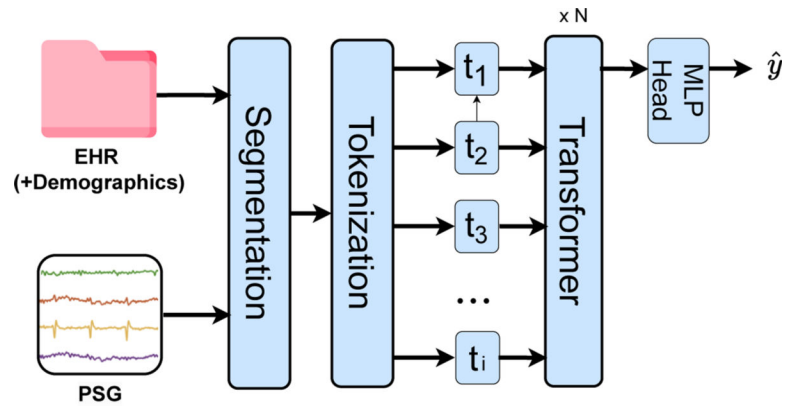


Figure 1:

The model architecture. After segmenting the demographics, from EHRs (electronic health records), and sleep signals, tokenizer synchronizes the data streams from different sources and forms a representation that can be fed to the deep learning model comprised of transformer encoders. The output of the last transformer layer is fed to a multi-layer perception to detect the occurrence of an apnea-hypopnea event.

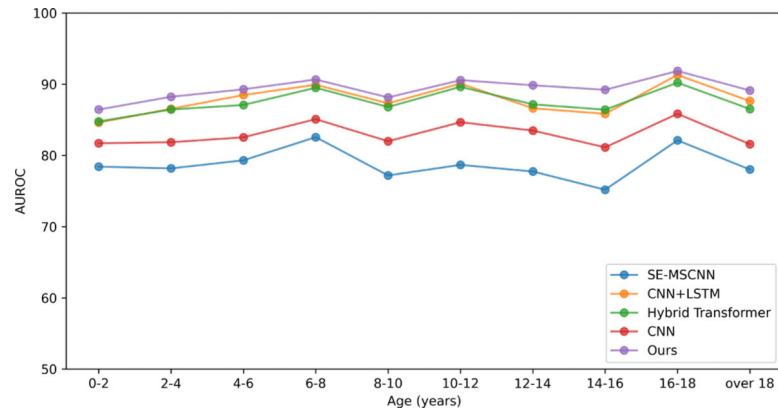


Figure 2: Performance comparison of the proposed and state-of-the-art models on different age groups measured by AUROC on NCH dataset.

Table 1:
NCH and CHAT datasets characteristics

		NCH	CHAT
Number of Patients		3,673	453
Number of Sleep Studies		3,984	453
Sex	Male	2,068	219
	Female	1,604	234
Race	Asian	93	8
	Black	738	252
	White	2,433	161
	Other	409	32
Age (year)		[0–30](mean=8.8)	[5–9](mean=6.5)

Table 2:

The data annotation counts. EEG arousal, Obstructive Hypopnea, and Mixed apnea events are not annotated in the CHAT dataset. There are four sleep stages, including one for rapid eye movement (REM) sleep and three (N1, N2, N3) for non-REM (NREM) sleep. For each dataset, the number of sleep epochs in each stage is shown.

Event	NCH	CHAT (Baseline)
Oxygen Desaturation	215,280	65,006
Oximeter Event	161,644	9,864
EEG arousal	146,052	—
Respiratory Events		
Hypopnea	14,522	15,871
Obstructive Hypopnea	42,179	—
Obstructive apnea	15,782	7,075
Central apnea	6,938	3,656
Mixed apnea	2,650	—
Sleep Stages		
Wake	665,676	10,282
N1	128,410	13,578
N2	1,383,765	19,985
N3	875,486	9,981
REM	611,320	3,283

Table 3:

Comparing our model against four other methods for identifying apnea. The mean (standard deviation) values are shown.

Method	CHAT		NCH Data Bank	
	F1	AUROC	F1	AUROC
CNN (Chang et al., 2020)	77.5(0.8)	86.8(1.0)	77.2(1.1)	86.4(1.2)
SE-MSCNN (Chen et al., 2022)	73.9(2.1)	82.9(1.8)	73.0(2.4)	82.2(1.9)
CNN+LSTM (Zarei et al., 2022)	81.7(0.6)	89.7(0.7)	81.7(0.8)	89.4(0.6)
Hybrid Transformer (Hu et al., 2022)	81.3(1.0)	89.6(0.5)	81.0(0.9)	89.4(0.7)
Ours	83.1 (1.0)	90.0 (0.8)	82.6(0.5)	90.4(0.4)
Ours with demographics	83.9(0.8)	90.6(0.7)	82.9(0.6)	90.7(0.6)

Table 4:

Apnea classification performance using our method when subsets of size 1 and 2 of PSG signals are used. The check-marks indicate the included signal types in each experiment. The mean (standard deviation) values are shown.

EOG	EEG	ECG	Resp	SpO ₂	CO ₂	CHAT		NCH	
						F1	AUROC	F1	AUROC
✓						75.9(0.4)	79.4(0.5)	75.4(1.5)	79.9(1.1)
	✓					73.6(1.1)	78.3(0.7)	72.7(1.3)	77.5(1.0)
		✓				76.8(0.3)	83.7(0.5)	73.0(1.2)	80.1(0.6)
			✓			77.0(0.8)	84.6(1.1)	76.4(0.8)	85.3(0.7)
				✓		75.8(1.1)	84.0(0.6)	78.6(0.9)	87.1(0.7)
					✓	75.2(0.4)	83.9(0.8)	67.4(0.2)	75.9(0.8)
✓	✓					77.6(0.3)	80.3(0.5)	77.0(1.4)	81.2(1.0)
✓		✓				78.6(1.5)	85.2(1.0)	76.6(1.0)	83.4(0.8)
✓			✓			80.0(1.2)	86.5(1.2)	79.9(0.9)	87.6(0.6)
✓				✓		80.3(0.7)	87.5(0.5)	79.6(0.8)	87.8(0.6)
✓					✓	76.7(1.8)	84.9(1.5)	76.1(1.5)	83.1(1.2)
	✓	✓				78.8(0.4)	84.9(0.7)	75.1(0.8)	81.1(0.8)
	✓		✓			78.4(0.7)	85.6(0.7)	79.2(1.0)	86.9(1.3)
	✓			✓		78.0(0.8)	86.7(1.0)	78.9(0.6)	87.4(0.6)
	✓				✓	75.7(2.5)	83.9(2.4)	72.7(1.0)	79.1(0.8)
		✓	✓			79.7(0.7)	86.5(1.0)	77.5(1.1)	85.7(0.9)
		✓		✓		82.5(0.7)	89.4(0.7)	80.7(0.4)	88.4(0.4)
		✓			✓	78.1(1.0)	85.2(1.1)	75.1(0.9)	81.9(0.6)
			✓	✓		81.1(1.3)	88.3(1.5)	78.4(0.9)	87.0(0.8)
			✓		✓	78.3(1.3)	86.0(1.1)	75.9(0.5)	84.7(0.7)
				✓	✓	79.7(1.2)	87.5(0.9)	79.8(0.8)	87.5(0.6)
✓	✓	✓	✓	✓	✓	83.1(1.0)	90.0(0.8)	82.6(0.5)	90.4(0.4)