# Multimodal Sleep Apnea Detection with Missing or Noisy Modalities

**Hamed Fayyaz** FAYYAZ@UDEL.EDU
*University of Delaware*

**Niharika S. D'Souza** NIHARIKA.DSOUZA@IBM.COM
*IBM Research Almaden*

**Rahmatollah Beheshti** RBI@UDEL.EDU
*University of Delaware*

## Abstract

Polysomnography (PSG) is a type of sleep study that records multimodal physiological signals and is widely used for purposes such as sleep staging and respiratory event detection. Conventional machine learning methods assume that each sleep study is associated with a fixed set of observed modalities and that all modalities are available for each sample. However, noisy and missing modalities are a common issue in real-world clinical settings. In this study, we propose a comprehensive pipeline aiming to compensate for the missing or noisy modalities when performing sleep apnea detection. Unlike other existing studies, our proposed model works with any combination of available modalities. Our experiments show that the proposed model outperforms other state-of-the-art approaches in sleep apnea detection using various subsets of available data and different levels of noise, and maintains its high performance (AUROC>0.9) even in the presence of high levels of noise or missingness. This is especially relevant in settings where the level of noise and missingness is high (such as pediatric or outside-of-clinic scenarios). Our code is publicly available at https://github.com/healthylaife/apnea-missing-modality.

## 1. Introduction

Sleep is essential in maintaining and promoting overall health and well-being (Luyster et al., 2012). Insufficient or poor-quality sleep has been linked to a wide range of health problems, including cardiovascular disease (Kasasbeh et al., 2006), obesity (Taheri, 2006), diabetes (Knutson et al., 2006), and mental health disorders (Zimmerman et al., 2006; Schwartz et al., 2005).

Conditions that affect sleep quality, timing, or duration and impact a person's ability to function properly while awake are referred to as sleep disorders. There are various types of sleep disorders, including insomnia, circadian rhythm sleep disorders, sleep-disordered breathing, hypersomnia, parasomnias, and restless legs syndrome (Pavlova and Latreille, 2019).

In particular, sleep apnea and hypopnea syndrome (SAHS) are breathing disorders that are characterized by recurring pauses in breathing during sleep, which usually cause fragmentation of sleep and can lead to oxygen deprivation and disrupt normal physiological processes (Vaquerizo-Villar et al., 2020). The main types of SAHS are obstructive sleep

apnea (the most common type), central sleep apnea, complex (mixed) sleep apnea, and hypopnea. It is estimated that 26% of people between 30 and 70 years (Schwartz et al., 2018) and 1% to 5% of children suffer from SAHS, with the highest prevalence in age 2 to 8 (Kheirandish-Gozal and Gozal, 2012; Bixler et al., 2009; Marcus et al., 2012).

Polysomnography (PSG) is the gold standard for diagnosing sleep-related breathing disorders. It 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 (Rundo and Downey III, 2019). A PSG generally includes: 1) brain electrical activity (electroencephalogram or EEG), 2) eye movements during sleep (electrooculogram or EOG), 3) cardiac rate and rhythm (electrocardiogram or ECG), 4) blood oxygen saturation (pulse oximetry or $SpO_2$), 5) measurement of exhaled air to indirectly measure blood $CO_2$ (end-tidal carbon dioxide or $ETCO_2$), 6) respiratory effort in thorax and abdomen (respiratory inductance plethysmography or RIP), 7) and nasal and oral airflow. PSG is generally considered effective; however, it presents many challenges, including complexity, cost, intrusiveness, and the need for intensive involvement of clinical providers (Spielmanns et al., 2019). The multimodal nature of PSG data provides a diverse and holistic view of the subjects (Muhammad et al., 2021; Kline et al., 2022; Liu et al., 2023).

While a fairly large family of studies aiming to detect SAHS from PSG is present, existing methods have two major limitations. First, conventional methods are built based on the assumption that all modalities are available for all subjects. However, in real-world scenarios, certain modalities may be partially or entirely missing due to technical issues or limitations during data acquisition. Additionally, the PSG signals can be corrupted by noise from various sources, including electrode artifacts, electrical interference, or patient movement. Detecting apnea becomes significantly more challenging when dealing with missing or noisy modalities (especially the primary signals such as $SpO_2$). The second limitation is the applicability of existing methods to a predetermined subset of PSG signals. This issue makes it unclear to what degree a method developed for a certain set of modalities would work (if at all) on applications with different sets of available modalities. To address these limitations, we propose a flexible machine learning-based pipeline for apnea detection with any combination of available signal modalities. Specifically, the contributions of our study are:

- We extensively investigate the effects of missing or partially available modalities in apnea detection.
- We present a novel universal pipeline to predict apnea with any type, length, or quality of available modality.
- Through extensive experiments, we show that our method is robust to noisy and missing modalities and outperforms prior methods in various apnea detection scenarios.

### Generalizable Insights about Machine Learning in the Context of Healthcare

Sleep apnea is a common disorder affecting various individuals including adults and children. Noisy and missing modalities are common in both inside- and outside-clinic sleep studies aiming to diagnose or treat apnea. Our proposed method offers a robust solution to achieve high-quality performance in the absence of complete sleep data.

Similar to sleep apnea, while most of the proposed ML pipelines in healthcare are designed to work with or tested only on complete data, acquired clinical data in the real world are mostly noisy and incomplete. Numerous healthcare-related tasks also face this problem as well. Therefore, our proposed method can offer a generalizable blueprint for similar scenarios as well.

## 2. Related Work

Due to the high relevance of detecting apnea events using machine learning methods, many attempts have been made to automate apnea detection using such methods. A large body of prior work focuses on uni-modal methods. Considering the modalities utilized in prior studies, we highlight four categories in the following (a non-exhaustive list). (1) Most existing work uses ECG for apnea detection. while ECG is not the most relevant signal in clinical settings, the popularity of the methods for ECG is partly due to the availability of public ECG datasets (Penzel et al., 2000). These methods generally use 'band-pass' filters to reduce the noise sourced from the baseline wander, muscle artifacts, power line interference, and other sources (Urtnasan et al., 2018; Bahrami and Forouzanfar, 2022). Since ECG signals contain complex patterns, many studies have used extensive preprocessing steps, including automatic feature extraction approaches through deep neural networks, in their pipeline for extracting features (Shen et al., 2021; Chang et al., 2020; Chen et al., 2022; Zarei et al., 2022). These methods are extensively reviewed by Salari et al. (2022). (2) Many studies have used $SpO_2$, which is the most clinically relevant signal for apnea detection (in adults). Similar to the previous category, various types of manual or automatic feature extraction methods have been used in this category (Álvarez et al., 2012; Morillo and Gross, 2013; Uçar et al., 2017; John et al., 2021). (3) The combination of these two signals (ECG and $SpO_2$) have been used, especially to handle 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). (4) Finally, the fourth major category is related to a group of work that uses EEG signals to detect apneic events (Vimala et al., 2019; Zhao et al., 2021; Almuhammadi et al., 2015). As discussed earlier, a common limitation of existing studies is unclear generalizability to other PSG signals. Oftentimes, these models cannot utilize an additional modality not seen in the training phase and handle partially available or noisy modalities.

Multimodal learning approaches aim to use information from different sources to better understand an underlying phenomenon and improve the performance of downstream tasks. Data fusion is an integral part of multimodal learning, which is the process of integrating multiple data sources to build a representation that is more pertinent than any individual source (Stahlschmidt et al., 2022). Depending on the fusion stage, existing techniques can be categorized into i) early, ii) intermediate, and iii) late fusion, where the data from modalities are respectively combined at i) the input (Lim et al., 2019; Neves et al., 2023), ii) through a middle representation (like a hidden layer) between the individual modalities and the final model (Hassan et al., 2022), and iii) using separate models on each modality which are combined later (Wang et al., 2019; Zhang et al., 2019). When some modalities are missing or noisy, data fusion enables combining available information from different sources

to create a more comprehensive and reliable representation, enhancing the robustness of the models (Gaw et al., 2022).

Multimodal information can be especially effective for apnea detection, as exploiting multimodal data not only can outperform using only uni-modal data but can also improve the robustness of the system when one or more modalities are missing. Despite the importance of multimodal learning, there is limited work (Ye et al., 2023; Fayyaz et al., 2023; Van Steenkiste et al., 2020) on multimodal data fusion for apnea detection. We are not aware of any previous study that directly aims at handling missing and noisy data in the context of sleep apnea detection.

When it comes to working with multi-modal data, the choice of a data fusion mechanism presents an additional dimension of the problem. In machine learning pipelines, fusing multimodal data in any of the discussed stages has advantages and disadvantages. Among the existing fusion techniques, gated fusion (Arevalo et al., 2017) is a popular method, with applications in different domains such as computer vision (Hosseinpour et al., 2022; Zhang et al., 2018; Ren et al., 2018; Feng et al., 2020), speech recognition (Fan et al., 2020), and natural language processing (Du et al., 2022). By employing gating mechanisms, gated fusion models selectively weigh and fuse information and capture complementary aspects. In our study, we propose a modified version of the gated fusion mechanism that considers the abnormality of the present modalities (i.e., the extent to which current signal patterns deviate from the others) to identify the optimized way of "fusing" the present modalities.

## 3. Problem setup

Consider a sleep dataset consisting of $I$ sleep studies (e.g., $I$ individual nights in the clinic) shown by $\{S_i\}_{i=1}^{I}$. Each study can be divided into equal-length non-overlapping epochs (windows), considered as a sample $X$, and shown by

$$S_i = \{X^{i,j}\}_{j=1}^{J_i}, \tag{1}$$

where, $J_i$ is the number of epochs in $i$-th study.

Each sample can consist of different PSG signals, considered as distinct modalities:

$$X^{i,j} = (X_1^{i,j}, X_2^{i,j}, ..., X_M^{i,j}), \tag{2}$$

where $M$, is the total number of modalities.

Furthermore, each sample has a label $y$, showing if an apnea event occurs during the sample time window. Thus, $X_m^{i,j}$ is a signal from the modality $m$ that belongs to the epoch $j$ of the study $i$, and is a multivariate time series consisting of $T$ time points:

$$X_m^{i,j} = (X_{1,m}^{i,j}, X_{2,m}^{i,j}, ..., X_{T,m}^{i,j}). \tag{3}$$

In this work, we design and train a model $f$ (parameterized with $\theta$) that can handle missing and noisy modalities and estimates the occurrence of apnea $\hat{y}$ in a given epoch:

$$\hat{y}^{i,j} = f_\theta(X^{i,j}) \tag{4}$$

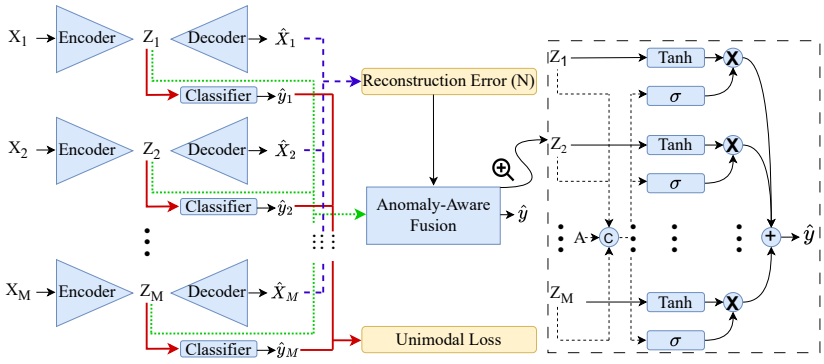


Figure 1: Our proposed pipeline. It consists of four components: (1) Encoder, (2) Decoder, (3) Classifier, and (4) Anomaly-aware Fusion. In the first (unimodal) step of training, an encoder for each modality is used to transform the input into a latent space $Z_m$. Besides, a unimodal classifier and decoder are utilized to detect apnea and reconstruct the input, respectively. In the second (multimodal) step of training, the Anomaly-Aware Fusion module classifies the epochs using the unimodal latent representation ($Z$) and the reconstruction error for each input signal. Notations `C` and `X` denote the concatenation and multiplication operations respectively.

## 4. Method

For detecting apnea with missing or noisy modalities, we propose a two-step method in companion with a multimodal autoencoder network using a transformer backbone (Figure 1). The two steps of our method include unimodal pre-training and multimodal training. The rationale for introducing a two-step training process in our design is that the pipeline can use samples that do not have complete modalities to train the model when dealing with incomplete modalities (akin to an imputing mechanism paradigm).

**Unimodal Pre-training** We use unimodal transformer-based autoencoders that reconstruct a unimodal input $X_m$, by mapping the input space to a lower dimensional latent space $Z_m$. More specifically, an encoder $E_m$ maps the signal to a latent representation ($Z_m = E_m(X_m)$) and a decoder $D_m$ reconstructs the original signal from the latent representation ($\hat{X}_m = D_m(Z_m)$).

The encoder and decoder consist of a series of transformer blocks. Besides being used for signal reconstruction, the decoder is also used for anomaly detection by using the reconstruction error as a proxy. Our method also includes a unimodal classifier $C_m$ to detect the epochs with apneic events using $Z_m$ ($\hat{y}_m = C_m(Z_m)$).

This way, the encoder learns a mapping to a latent space that preserves the fundamental characteristics of the signals, while extracting the key features for apnea detection. Additional details about the design of the transformers are presented in Appendix A.

For modality $m$, the reconstruction loss $\mathcal{L}_R$ can be formulated as the mean squared error (`MSE`) of the original signal and its reconstructed version:

$$\mathcal{L}_R^m = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{t=1}^{T}(X_{t,m}^{i,j} - \hat{X}_{t,m}^{i,j})^2}{N \times T}, \tag{5}$$

where,

$$N = \sum_{i=1}^{I} J_i, \text{ and} \tag{6}$$

$$\hat{X}_{t,m}^{i,j} = D_m(E_m(X_{t,m}^{i,j})). \tag{7}$$

In addition, the classification loss $\mathcal{L}_C$ (related to the apnea detection task) is defined using binary cross entropy (`BCE`) for modality $m$ as:

$$\mathcal{L}_C^m = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{J_i} BCE(y_m^{i,j}, \hat{y}_m^{i,j}), \tag{8}$$

where,

$$\hat{y}^{i,j} = C_m(E_m(X_m^{i,j})). \tag{9}$$

The model (in this first step) is trained by jointly optimizing the two aforementioned loss functions over all modalities:

$$\mathcal{L}_{Unimodal} = \sum_{m=1}^{M} \alpha_m \mathcal{L}_R^m + \beta_m \mathcal{L}_C^m \tag{10}$$

where, the hyperparameters ($\alpha_m$ and $\beta_m$) are used to tune the contribution of each loss.

**Multimodal Training** In the second step of our method, we use a modified version of gated fusion (Arevalo et al., 2017) to detect apneic events. We refer to this module as Anomaly Aware Fusion (`AAF`). It receives the latent representation $Z$ and a distance measure $A$ for each signal. The distance measure $A$ aims to capture the abnormality of the signals (the degree to which the signal deviates from the original), similar to Lee and Kang (2022). Here, $A$ is defined as:

$$A = [a_1, a_2, ..., a_M], \tag{11}$$

where,

$$a_m = [|X_{1,m} - \hat{X}_{1,m}|, ..., |X_{T,m} - \hat{X}_{T,m}|], \tag{12}$$

while the anomaly score $a_m$ captures the absolute distance between the two time-series $X_m$ and $\hat{X}_m$. While closely related, the reconstruction loss in Eq. 12 yields a single value; however, the distance measure in Eq. 5 is a vector that shows the degree of abnormality in each time step.

Specifically, `AAF` can be shown as a function receiving $Z$ and $A$ to predict the final label:

$$\hat{y} = AAF(Z, A). \tag{13}$$

As the encoder-decoder model learns the distribution of (mostly normal) signals, the distance measure (captured by the absolute differences) increases when this model receives an abnormal signal, as the signals containing the abnormal patterns are not reconstructed well.

Furthermore, $Z$ in `AAF` (Eq. 13) refers to:

$$Z = [Z_1, Z_2, ..., Z_M], \tag{14}$$

where, $Z_m$ corresponds to a feature vector in the latent representation space associated with the modality $m$.

The `AAF` module (the right box in Figure 1) takes these steps. First, each feature vector $Z_m$ is fed to a fully connected layer with a $Tanh$ activation function, which is intended to encode an internal representation feature $h_m$ based on the particular modality:

$$h_m = Tanh(W_m Z_m). \tag{15}$$

where, $W_m$s are learnable weights related to the modality $m$. Another fully connected layer is used for each $Z_m$, which controls the contribution of the feature calculated from $Z_m$ to the overall output of `AAF`. The output of this layer is another latent representation named $Z'$:

$$Z_m^{'} = \sigma(W_{Z_m}[Z_1, Z_2, ..., Z_M, A]) \tag{16}$$

where, $W_{Z_m}$s are learnable weights related to modality $m$.

When a sample is fed to the gated fusion network, a gate layer associated with the modality $m$ receives the feature vectors of all modalities and uses them to decide whether the modality $m$ may contribute or not to the internal encoding of the particular input sample:

$$h = Z_1^{'} h_1 + Z_2^{'} h_2 + ... + Z_M^{'} h_M. \tag{17}$$

Finally, a fully connected layer and a sigmoid function $\sigma$ transform $h$ into a label:

$$\hat{y} = \sigma((W_c h) + b_c), \tag{18}$$

where, $W_c$ and $b_c$ are learnable weights and biases. We use the `BCE` loss (as in the previous step) for the classification loss. The encoder and decoder weights are frozen during this step.

## 5. Experiments

**Datasets** A large number of public sleep studies are available, including those available through the National Sleep Research Resource platform (Mueller, 2024). Most of those, however, are not a good fit for evaluating our method. For our study, we looked for sleep recordings that (i) had multimodal data (i.e., not unimodal or only a few modalities), (ii) were supervised (i.e., have expert annotated apnea event labels), and (iii) had a large number of samples (the $n$ size for the sleep studies and patients), to allow training large deep learning models. We use two of the largest public sleep datasets that match the above criteria. These two are 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).

**NCH** - This dataset offers a large and free source that includes both PSG signals and linked electronic health records (which includes demographics and longitudinal clinical data such as encounters, medication, measurements, diagnoses, and procedures). The dataset was collected between 2017 and 2019 at Nationwide Children's Hospital (NCH), Columbus, Ohio, USA. Sleep studies were annotated in real-time by technicians at the time of the study, and then staged and scored by a second technician after the study was completed. We used all studies that have all of the available six modalities.

**CHAT** - We also use recordings from the CHAT study, which is a randomized, single-blind, multicenter trial designed to analyze the efficacy of early removal of adenoids and tonsils (adenotonsillectomy) in 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, Boston, MA. We use the 453 sleep studies collected in the baseline in our work.

**Implementation details** Our model utilizes ECG, EEG, EOG, $SpO_2$, $CO_2$, and respiratory signals to detect apnea. As the signals have different sampling rates, we resampled all signals with a frequency of 128Hz. For model input, we divided each sleep study into 30-second non-overlapping epochs. ECG signals were initially denoised using a band-pass filter, which allows only a certain range of frequencies to pass through while blocking others, with lower and upper cutoff frequencies of 3Hz and 45Hz. Hamilton R-peak detection method (Hamilton and Tompkins, 1986) was utilized to extract R-R intervals and the amplitude of R-peaks from the ECG signal.

We run experiments in a 5-fold cross-validation manner and report the mean and standard deviation of performance of the trained models using the F1-score (harmonic mean of precision and recall) and AUROC (area under the receiver operating characteristic curve) in the test set.

**Baselines** We chose three related and state of the art (SOTA) studies with different architectures to compare with the proposed model. These three studies include architectures based on CNNs (Chang et al., 2020), CNN+LSTM (Zarei et al., 2022), and transformers (Fayyaz et al., 2023) as described in the following.

**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. They only utilized ECG for apnea detection. However, to have a fair comparison, we trained this model with the same modalities that we used for training our model.

**CNN+LSTM** - The second model presented by Zarei et al. (2022) uses an automatic feature extraction method developed by combining CNN and long short-term memory (LSTM) modules using ECG. A stack of fully connected layers is used at the end to classify the events. Similar to the previous baseline, we trained this model with the same modalities that we used for training our model.

**Transformers** - The last model proposed by Fayyaz et al. (2023) uses transformers followed by two layers of a fully connected network as a classifier for detecting apnea using PSG signals.

In addition to the above baselines, which have been designed for PSG studies, we include alternative variations of our method to demonstrate the performance of more general multimodal methods from outside the apnea studies.

### 5.1. Research Questions

We study five major research questions in our experiments to investigate the performance of our model.

**Q1: How do the methods perform with complete modalities?** We carried out an experiment to find the overall performance of our method and the baselines when provided with complete data (i.e., without any added noise or missingness).

Our proposed method follows a mid-fusion approach. In addition to the listed baselines, we investigate the performance of two alternative approaches that follow early and late fusion. For early-fusion, we concatenate all modalities and pass them to an encoder followed by a fully connected classifier. In late-fusion, we pass each modality through an encoder appended by a fully connected classifier. Finally, in an ensemble-wise manner, we consider the network related to each modality as a weak learner. The final label is calculated by averaging the output related to all modalities.

To study the ablation version of our method, we studied the effect of various components and the design choices of our proposed method on its performance in apnea detection. In the "pretraining" scenario, we skipped the first step of the training algorithm and only ran the second step of training (multimodal fine-tuning) to find the effect of fine-tuning on the overall performance of our method in apnea detection. In the "reconstruction loss" scenario, we evaluate the performance of our model in the absence of reconstruction loss (in unimodal pretraining) and the distance measures $A$ (in multimodal training). Finally, in the "gated fusion" scenario, we replaced the customized gated fusion we designed for this study with a fully connected neural network to show its effectiveness in comparison with a general machine learning model.

Table 1 shows the results. Our model has achieved superior performance, measured in terms of F1 score and AUROC, compared to the baselines.

**Q2: How do the methods perform in the presence of missing modalities?** To simulate the effect of missing modalities, we omitted signals within epochs in a random manner. More specifically, we randomly substituted a predefined percent of signals in epochs with zero. The algorithm for this purpose is presented in Appendix D. The results are shown in Figures 2.a (NCH) and 2.b (CHAT).

**Q3: How do the methods perform with noisy modalities?** We investigated how well our model and other top-performing models handle noisy data. In the absence of a dataset with controlled non-synthetic or measured noise, we add synthetic noise to our target datasets. We added white Gaussian noise to the signals to simulate unremovable types of noises (we discuss this further in Section 6), using a process explained further in Appendix D. Additive Gaussian noise can mimic the effect of many random processes that can occur in real-life settings. The results are shown in Figures 2.c (NCH) and 2.d (CHAT). The signal-to-noise ratio (SNR) is a measure of the strength of a desired signal compared

Table 1: The overall performance of our model, the baselines, and ablation versions when using complete data. Mean (± STD).

| Method | CHAT | | NCH | |
|---|---|---|---|---|
| | 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) |
| CNN+LSTM (Zarei et al., 2022) | 81.7 (0.6) | 89.7 (0.7) | 81.7 (0.8) | 89.4 (0.6) |
| Transformer (Fayyaz et al., 2023) | 83.1 (1.0) | 90.0 (0.8) | 82.6 (0.5) | 90.4 (0.4) |
| Early Fusion | 81.9 (2.0) | 89.0 (0.8) | 83.2 (0.8) | 91.1 (0.6) |
| Late Fusion | 79.9 (1.5) | 88.6 (0.3) | 80.8 (1.1) | 89.0 (1.1) |
| W/O pretraining | 78.6 (1.9) | 89.9 (0.7) | 84.1 (0.7) | 92.3 (0.5) |
| W/O reconstruction loss | 85.1 (0.7) | 92.3 (0.6) | 83.9 (0.6) | 92.3 (0.4) |
| Fully-connected Fusion | 82.2 (1.8) | 92.8 (1.1) | 83.6 (1.6) | 93.3 (0.6) |
| Ours (the proposed approach) | **86.6 (0.7)** | **93.3 (0.6)** | **85.2 (0.7)** | **93.4 (0.4)** |

Table 2: Our method's performance (mean AUROC ± STD) with the concurrent occurrence of noise and missingness on the NCH dataset.

| Missing Ratio | NCH | | | | | CHAT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Signal to noise ratio (dB) | | | | | | | | | |
| | 10 | 20 | 30 | 40 | 50 | 10 | 20 | 30 | 40 | 50 |
| 10% | 84.9(1.9) | 90.4(0.9) | 92.6(0.5) | 92.8(0.5) | 92.8(0.5) | 85.5 (0.7) | 85.5 (0.6) | 90.0 (0.1) | 91.7 (0.4) | 91.8 (0.5) |
| 20% | 82.9(1.9) | 89.0(1.1) | 91.8(0.6) | 92.0(0.5) | 92.1(0.5) | 83.1 (0.7) | 82.6 (1.5) | 87.7 (0.6) | 89.3 (0.1) | 89.9 (0.3) |
| 30% | 80.9(2.0) | 87.3(1.0) | 90.6(0.6) | 91.1(0.5) | 91.1(0.5) | 79.8 (0.8) | 79.7 (0.9) | 85.0 (1.1) | 87.5 (0.1) | 87.3 (0.1) |
| 40% | 78.6(1.9) | 85.4(1.2) | 89.3(0.6) | 89.9(0.5) | 89.8(0.5) | 76.6 (0.5) | 76.8 (1.5) | 81.8 (1.5) | 84.4 (0.8) | 84.6 (0.6) |
| 50% | 76.9(2.0) | 83.4(1.0) | 87.6(0.6) | 88.0(0.6) | 88.2(0.4) | 73.5 (0.1) | 73.6 (1.7) | 78.6 (0.9) | 81.0 (1.1) | 81.1 (0.9) |

to the background noise. It is often expressed in decibels (dB), and a higher SNR value indicates a cleaner signal.

**Q4: How do the methods perform when the input has both noisy and missing signals?** We also fed our model with both noisy and missing modalities at once, which combines the two studied scenarios in Q3 and Q4. The results on the NCH and CHAT datasets are shown in Table 2. We report the results of the baseline models for the Q4 scenario, in the Appendix (Table 5).

**Q5: How do the methods perform when specific modalities are completely missing?** We also study the methods with one missing modality at a time, to find the stability of our model when modalities are entirely (versus partially in Q2) unavailable. Results are shown in Table 3.

## 6. Discussion

In this study, we presented a machine learning pipeline for apnea detection with any combination of available PSG modalities. By leveraging the complementary nature of multiple modalities, data fusion techniques can mitigate the impact of missing data, reduce uncer-

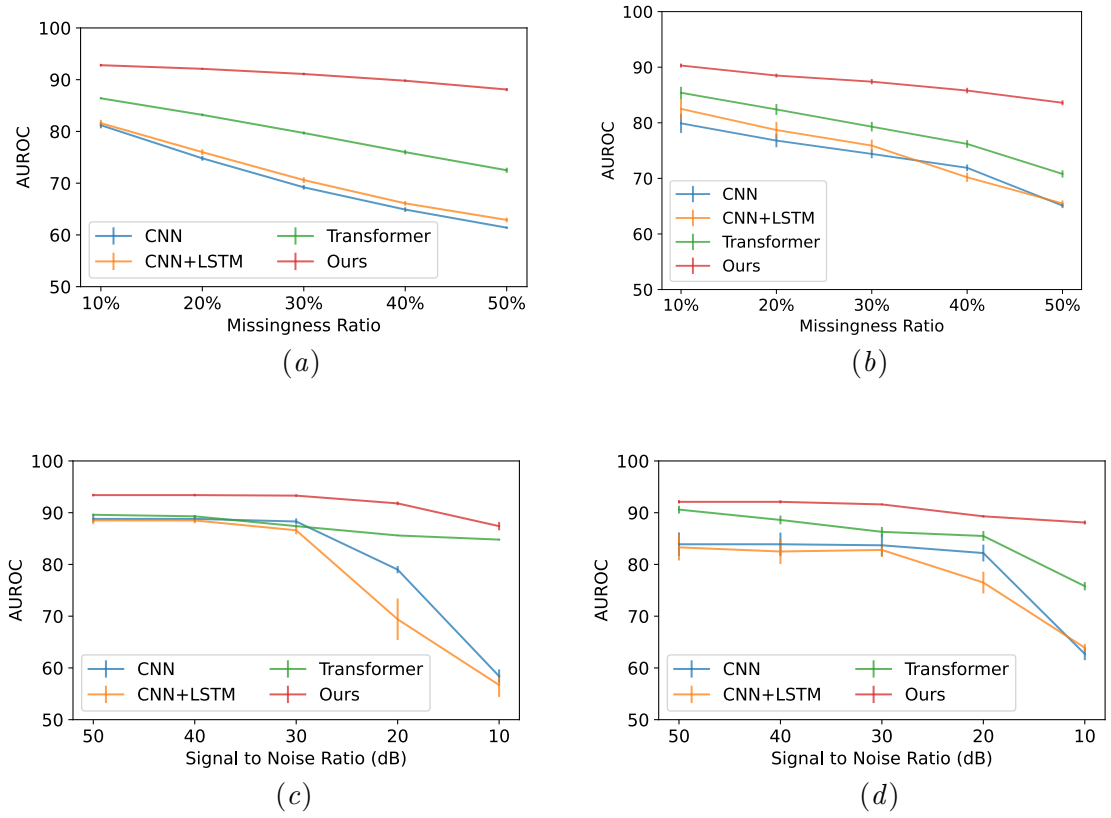


Figure 2: (a) Model performance with missing modalities on the NCH dataset, (b) Models Performance with missing modalities on CHAT dataset, (c) Model performance with noisy modalities on the NCH dataset, and (d) Models performance with noisy modalities on CHAT dataset. Error bars show the standard deviation

Table 3: Performance of our model (mean AUROC ± STD) when specific modalities are missing on the NCH dataset.

| Missing Modality | NCH | | | | CHAT | | | |
|---|---|---|---|---|---|---|---|---|
| | CNN | CNN-LSTM | Transformer | Ours | CNN | CNN-LSTM | Transformer | Ours |
| EOG | 83.3 (2.2) | 83.0 (9.9) | 89.2 (0.0) | **93.0 (0.7)** | 68.1 (4.5) | 57.5 (4.1) | 83.9 (3.6) | **91.9 (1.5)** |
| EEG | 84.4 (2.1) | 75.6 (9.4) | 89.2 (0.5) | **92.6 (0.6)** | 68.9 (3.5) | 51.5 (7.3) | 84.6 (3.2) | **91.3 (1.7)** |
| Resp | 82.9 (1.9) | 87.3 (2.3) | 86.3 (0.8) | **91.4 (0.5)** | 73.1 (11.1) | 56.6 (10.4) | 86.3 (2.7) | **88.3 (1.7)** |
| $SpO_2$ | 64.8 (1.4) | 63.2 (3.0) | 83.5 (1.0) | **91.5 (0.3)** | 51.0 (1.6) | 75.1 (5.8) | 83.5 (1.0) | **89.8 (0.7)** |
| $CO_2$ | 81.1 (3.4) | 80.4 (5.9) | 89.4 (0.5) | **92.9 (0.5)** | 73.3 (18.6) | 88.6 (0.7) | 89.4 (0.5) | **92.5 (0.9)** |
| ECG | 80.8 (2.3) | 85.6 (2.3) | 88.5 (0.5) | **93.1 (0.6)** | 61.8 (6.9) | 54.0 (8.8) | 81.0 (3.6) | **91.8 (0.8)** |
| EOG, EEG | 79.5 (2.9) | 74.7 (12.2) | 88.7 (0.5) | **91.5 (0.8)** | 62.3 (2.2) | 47.8 (5.1) | 82.7 (4.0) | **90.3 (2.1)** |
| None | 86.4(1.2) | 89.4(0.6) | 90.4 (0.4) | **93.4 (0.4)** | 86.8(1.0) | 89.7(0.7) | 90.0 (0.8) | **93.3 (0.6)** |

tainty, improve the overall robustness of the fused information, and compensate for the absence of certain modalities. All of these can allow for higher performance in machine learning tasks (Woo et al., 2023).

We showed that our hybrid pipeline —which combines a transformer-based architecture and a gated fusion— is robust to noisy and missing modalities and outperforms several recent baselines. Specifically, we investigated the effect of imperfect (noisy and missing) signals on the performance of apnea detection models. The performance of the models that are not designed to handle such conditions can degrade drastically as the imperfection of input signals increases. As shown in Figure 2, our model is more resilient to various degrees of missingness and noise in comparison with the baselines.

In this study, we used six different modalities, each of which with various noise types. Some of these types of noise have been comprehensively studied, and efficient removal approaches are available for such noises (Limaye and Deshmukh, 2016; Lai et al., 2018). As a result, some standard noises are removable using well-known denoising approaches without losing too much information and performance degradation. For instance, the "baseline wander" and "power-line interference" noises in ECG signals are removable using a notch filter (a filter that eliminates a single frequency from a spectrum of frequencies) and a high-pass filter (a filter that passes signals with a frequency higher than a certain cutoff frequency). To demonstrate the robustness of our proposed pipeline to this type of noise, we report the performance of our method in the presence of such noise in Appendix C.

In our study, however, we focus on noises that can deeply disturb the signals so that the original signals are not distinguishable. This type of noise is not easily reproducible, like the easy-to-model noises we discussed above. Modeling this type of noise is very hard. Such noise types include patient movements or electrode detachment, which are events that can introduce severe noise.

The clinical implications of our study are also worth noting. Our method can be especially relevant to the ongoing efforts to adopt at-home sleep apnea testing (HSAT) solutions. While there exist several FDA-approved commercial HSAT products (Van Pee et al., 2022; Manoni et al., 2020), inside-clinic PSG is still the gold standard for diagnosing sleep apnea. This is partly due to the challenges that sleep recording can present.

Research on HSATs is especially relevant to pediatric populations, as such tools are almost widely available for adults, but not children (Kirk et al., 2017). In fact, a key barrier to running sleep apnea testing outside the clinic for children is the presence of noise and missingness, as children can move more frequently, be less cooperative, and pull the probes, among other issues.

Additionally, a key difference between apnea detection in adults versus children relates to using $SpO_2$ (for adults) versus EEG (for children) as the main signal for apnea detection (Bandla and Gozal, 2000). In children, the brain activates early on to regulate breathing and sleep disorders, therefore it functions as a better signal for apnea detection. As we show in Table 3, even without EEG and EOG (i.e., the two critical but challenging to collect signals in children), our method performs very well.

The two datasets that we use in this study are related to pediatric sleep studies, further showcasing the performance of our pipeline on pediatric cases. We also report the performance of our model across different ages in Appendix E to further highlight the potential of our pipeline in informing the efforts targeting younger age patients.

Some limitations of our study are worth noting. A common concern in apnea detection models relates to the lack of sub-typing. Although apnea and hypopnea have three different types (i.e., obstructive, central, and mixed) (Javaheri et al., 2017), our work, similar to

most prior studies, only tries to train a model that detects apnea, not its type and severity. Another limitation relates to model interpretability. While the fusion mechanism in our model compensates for noisy or missed signals by dynamically adjusting its decision-making process and leveraging signals from other modalities, the precise mechanism of compensation remains unclear due to the model's "opaque box" nature. Moreover, in this study, we manually added noise to the lab-recorded PSG data. While we are not aware of any large study collecting at-home sleep data, evaluating our model using data recorded using HSAT devices with actual noise could have demonstrated our model's performance more comprehensively. Additionally, our model may underperform on unbalanced datasets (i.e., those consisting mainly of data from healthy individuals without apnea).

In the future, we plan to investigate whether applying our method can also help to improve sleep staging performance in comparison with existing methods. Sleep staging is mostly done using EEG and EOG.

## Acknowledgments

We would like to express our gratitude to Dr. Abigail Strang for her invaluable contribution to this work. We also acknowledge the support provided by the Google Cloud Research Credits Program with the award GCP279907463. Our study was supported by the NIH award U54-GM104941

## References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI )*, pages 265–283, Savannah, GA, USA, 2016. ACM.

Wafaa S Almuhammadi, Khald AI Aboalayon, and Miad Faezipour. Efficient obstructive sleep apnea classification based on EEG signals. In *2015 Long Island Systems, Applications and Technology*, pages 1–6. IEEE, 2015.

Daniel Álvarez, Roberto Hornero, J Víctor Marcos, and Félix Del Campo. Feature selection from nocturnal oximetry using genetic algorithms to assist in obstructive sleep apnoea diagnosis. *Medical engineering & physics*, 34(8):1049–1057, 2012.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

Mahsa Bahrami and Mohamad Forouzanfar. 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.

Hari PR Bandla and David Gozal. Dynamic changes in EEG spectra during obstructive apnea in children. *Pediatric pulmonology*, 29(5):359–365, 2000.

Edward O Bixler, Alexandros N Vgontzas, Hung-Mo Lin, Duanping Liao, Susan Calhoun, Antonio Vela-Bueno, Fred Fedok, Vukmir Vlasic, and Gavin Graff. Sleep disordered breathing in children in a general population sample: prevalence and risk factors. *Sleep*, 32(6):731–736, 2009.

Hung-Yu Chang, Cheng-Yu Yeh, Chung-Te Lee, and Chun-Cheng Lin. A sleep apnea detection system based on a one-dimensional deep convolution neural network model using single-lead electrocardiogram. *Sensors*, 20(15):4157, 2020.

Xianhui Chen, Ying Chen, Wenjun Ma, Xiaomao Fan, and Ye Li. Toward sleep apnea detection with lightweight multi-scaled fusion network. *Knowledge-Based Systems*, 247: 108783, 2022.

François Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

Yongping Du, Yang Liu, Zhi Peng, and Xingnan Jin. Gated attention fusion network for multimodal sentiment classification. *Knowledge-Based Systems*, 240:108107, 2022.

Cunhang Fan, Jiangyan Yi, Jianhua Tao, Zhengkun Tian, Bin Liu, and Zhengqi Wen. Gated recurrent fusion with joint training framework for robust end-to-end speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:198–209, 2020.

Hamed Fayyaz, Abigail Strang, and Rahmatollah Beheshti. Bringing at-home pediatric sleep apnea testing closer to reality: A multi-modal transformer approach. In *Machine Learning for Healthcare Conference*. PMLR, 2023.

Shuanglang Feng, Heming Zhao, Fei Shi, Xuena Cheng, Meng Wang, Yuhui Ma, Dehui Xiang, Weifang Zhu, and Xinjian Chen. Cpfnet: Context pyramid fusion network for medical image segmentation. *IEEE transactions on medical imaging*, 39(10):3008–3018, 2020.

Nathan Gaw, Safoora Yousefi, and Mostafa Reisi Gahrooei. Multimodal data fusion for systems improvement: A review. *IISE Transactions*, 54(11):1098–1116, 2022.

Patrick S Hamilton and Willis J Tompkins. Quantitative investigation of qrs detection rules using the mit/bih arrhythmia database. *IEEE transactions on biomedical engineering*, BME-33(12):1157–1165, 1986.

Md Rafiul Hassan, Shamsul Huda, Mohammad Mehedi Hassan, Jemal Abawajy, Ahmed Alsanad, and Giancarlo Fortino. Early detection of cardiovascular autonomic neuropathy: A multi-class classification model based on feature selection and deep learning feature fusion. *Information Fusion*, 77:70–80, 2022.

Hamidreza Hosseinpour, Farhad Samadzadegan, and Farzaneh Dadrass Javan. Cmgfnet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, 184:96–115, 2022.

Shahrokh Javaheri, Ferran Barbe, Francisco Campos-Rodriguez, Jerome A Dempsey, Rami Khayat, Sogol Javaheri, Atul Malhotra, Miguel A Martinez-Garcia, Reena Mehra, Allan I Pack, et al. Sleep apnea: types, mechanisms, and clinical cardiovascular consequences. *Journal of the American College of Cardiology*, 69(7):841–858, 2017.

Arlene John, Koushik Kumar Nundy, Barry Cardiff, and Deepu John. 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.

E Kasasbeh, David S Chi, and G Krishnaswamy. Inflammatory aspects of sleep apnea and their cardiovascular consequences. *Southern medical journal*, 99(1):58–68, 2006.

Leila Kheirandish-Gozal and David Gozal. *Sleep disordered breathing in children: a comprehensive clinical guide to evaluation and treatment*. Springer Science & Business Media, 2012.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Valerie Kirk, Julie Baughn, Lynn D'Andrea, Norman Friedman, Anjalee Galion, Susan Garetz, Fauziya Hassan, Joanna Wrede, Christopher G Harrod, and Raman K Malhotra. 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.

Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171, 2022.

Kristen L Knutson, Armand M Ryden, Bryce A Mander, and Eve Van Cauter. Role of sleep duration and quality in the risk and severity of type 2 diabetes mellitus. *Archives of internal medicine*, 166(16):1768–1774, 2006.

Chi Qin Lai, Haidi Ibrahim, Mohd Zaid Abdullah, Jafri Malin Abdullah, Shahrel Azmin Suandi, and Azlinda Azman. Artifacts and noise removal for electroencephalogram (EEG): A literature review. In *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pages 326–332. IEEE, 2018.

Harlin Lee, Boyue Li, Shelly DeForte, Mark L Splaingard, Yungui Huang, Yuejie Chi, and Simon L Linwood. A large collection of real-world pediatric sleep studies. *Scientific Data*, 9(1):1–12, 2022.

Yunseung Lee and Pilsung Kang. Anovit: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder. *IEEE Access*, 10:46717–46724, 2022. doi: 10.1109/ACCESS.2022.3171559.

Teck-Yian Lim, Amin Ansari, Bence Major, Daniel Fontijne, Michael Hamilton, Radhika Gowaikar, and Sundar Subramanian. Radar and camera early fusion for vehicle detection in advanced driver assistance systems. In *Machine learning for autonomous driving*

*workshop at the 33rd conference on neural information processing systems*, volume 2, 2019.

Hrishikesh Limaye and VV Deshmukh. Ecg noise sources and various noise removal techniques: A survey. *International Journal of Application or Innovation in Engineering & Management*, 5(2):86–92, 2016.

Ziyi Liu, Jiaqi Zhang, Yongshuai Hou, Xinran Zhang, Ge Li, and Yang Xiang. Machine learning for multimodal electronic health records-based research: Challenges and perspectives. In *Health Information Processing*, pages 135–155. Springer Nature Singapore, 2023. ISBN 978-981-19-9865-2.

Faith S Luyster, Patrick J Strollo Jr, Phyllis C Zee, and James K Walsh. Sleep: a health imperative. *Sleep*, 35(6):727–734, 2012.

Alessandro Manoni, Federico Loreti, Valeria Radicioni, Daniela Pellegrino, Luigi Della Torre, Alessandro Gumiero, Damian Halicki, Paolo Palange, and Fernanda Irrera. A new wearable system for home sleep apnea testing, screening, and classification. *Sensors*, 20(24):7014, 2020.

Carole L Marcus, Lee J Brooks, Sally Davidson Ward, Kari A Draper, David Gozal, Ann C Halbower, Jacqueline Jones, Christopher Lehmann, Michael S Schechter, Stephen Sheldon, et al. Diagnosis and management of childhood obstructive sleep apnea syndrome. *Pediatrics*, 130(3):e714–e755, 2012.

Carole L Marcus, Reneé H Moore, Carol L Rosen, Bruno Giordani, Susan L Garetz, H Gerry Taylor, Ron B Mitchell, Raouf Amin, Eliot S Katz, Raanan Arens, et al. A randomized trial of adenotonsillectomy for childhood sleep apnea. *N Engl J Med*, 368:2366–2376, 2013.

Daniel Sánchez Morillo and Nicole Gross. Probabilistic neural network approach for the detection of sahs from overnight pulse oximetry. *Medical & biological engineering & computing*, 51(3):305–315, 2013.

Remo Mueller. Sleep data - national sleep research resource - nsrr, 2024. URL http://www.sleepdata.org/.

Ghulam Muhammad, Fatima Alshehri, Fakhri Karray, Abdulmotaleb El Saddik, Mansour Alsulaiman, and Tiago H Falk. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Information Fusion*, 76:355–375, 2021.

Francisco Soares Neves, Rafael Marques Claro, and Andry Maykol Pinto. End-to-end detection of a landing platform for offshore uavs based on a multimodal early fusion approach. *Sensors*, 23(5):2434, 2023.

Rahul Krishnan Pathinarupothi, Ekanath Srihari Rangan, EA Gopalakrishnan, R Vinaykumar, KP Soman, 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.

Milena K Pavlova and Véronique Latreille. Sleep disorders. *The American journal of medicine*, 132(3):292–299, 2019.

Thomas Penzel, George B Moody, Roger G Mark, Ary L Goldberger, and J Hermann Peter. The apnea-ECG database. In *Computers in Cardiology 2000. Vol. 27 (Cat. 00CH37163)*, pages 255–258. IEEE, 2000.

Antonio G Ravelo-García, Jan F Kraemer, Juan L Navarro-Mesa, Eduardo Hernández-Pérez, Javier Navarro-Esteva, Gabriel Juliá-Serdá, Thomas Penzel, and Niels Wessel. Oxygen saturation and rr intervals feature selection for sleep apnea detection. *Entropy*, 17(5):2932–2957, 2015.

Susan Redline, Raouf Amin, Dean Beebe, Ronald D Chervin, Susan L Garetz, Bruno Giordani, Carole L Marcus, Renee H Moore, Carol L Rosen, Raanan Arens, 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.

Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3253–3261, 2018.

Jessica Vensel Rundo and Ralph Downey III. Polysomnography. *Handbook of clinical neurology*, 160:381–392, 2019.

Nader Salari, Amin Hosseinian-Far, Masoud Mohammadi, Hooman Ghasemi, Habibolah Khazaie, Alireza Daneshkhah, and Arash Ahmadi. Detection of sleep apnea using machine learning algorithms based on ECG signals: A comprehensive systematic review. *Expert Systems with Applications*, 187:115950, 2022.

Daniel J Schwartz, William C Kohler, and Gillian Karatinos. Symptoms of depression in individuals with obstructive sleep apnea may be amenable to treatment with continuous positive airway pressure. *Chest*, 128(3):1304–1309, 2005.

Martha Schwartz, Luis Acosta, Yuan-Lung Hung, Mariela Padilla, and Reyes Enciso. Effects of cpap and mandibular advancement device treatment in obstructive sleep apnea patients: a systematic review and meta-analysis. *Sleep and Breathing*, 22:555–568, 2018.

Qi Shen, Hengji Qin, Keming Wei, and Guanzheng Liu. 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.

Marc Spielmanns, David Bost, Wolfram Windisch, Peter Alter, Tim Greulich, Christoph Nell, Jan Henrik Storre, Andreas Rembert Koczulla, and Tobias Boeselt. Measuring sleep quality and efficiency with an activity monitoring device in comparison to polysomnography. *Journal of clinical medicine research*, 11(12):825, 2019.

Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23(2):bbab569, 2022.

Shahrad Taheri. The link between short sleep duration and obesity: we should recommend more sleep to prevent obesity. *Archives of disease in childhood*, 91(11):881–884, 2006.

Seda Arslan Tuncer, Beyza Akılotu, and Suat Toraman. A deep learning-based decision support system for diagnosis of osas using ptt signals. *Medical hypotheses*, 127:15–22, 2019.

Muhammed Kürşad Uçar, Mehmet Recep Bozkurt, Cahit Bilgin, and Kemal Polat. Automatic detection of respiratory arrests in osa patients using ppg and machine learning techniques. *Neural Computing and Applications*, 28(10):2931–2945, 2017.

Erdenebayar Urtnasan, Jong-Uk Park, Eun-Yeon Joo, and Kyoung-Joung Lee. 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.

Bart Van Pee, Frederik Massie, Steven Vits, Pauline Dreesen, Susie Klerkx, Jagdeep Bijwadia, Johan Verbraecken, and Jeroen Bergmann. A multicentric validation study of a novel home sleep apnea test based on peripheral arterial tonometry. *Sleep*, 45(5):zsac028, 2022.

Tom Van Steenkiste, Dirk Deschrijver, and Tom Dhaene. Sensor Fusion using Backward Shortcut Connections for Sleep Apnea Detection in Multi-Modal Data. In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 116 of *Proceedings of Machine Learning Research*, pages 112–125. PMLR, 13 Dec 2020.

Fernando Vaquerizo-Villar, Daniel Álvarez, Leila Kheirandish-Gozal, Gonzalo C Gutiérrez-Tobal, Javier Gómez-Pilar, Andrea Crespo, Felix Del Campo, David Gozal, and Roberto Hornero. 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.

V Vimala, K Ramar, and M Ettappan. An intelligent sleep apnea classification system based on EEG signals. *Journal of medical systems*, 43(2):36, 2019.

Siwei Wang, Xinwang Liu, En Zhu, Chang Tang, Jiyuan Liu, Jingtao Hu, Jingyuan Xia, and Jianping Yin. Multi-view clustering via late fusion alignment maximization. In *IJCAI*, pages 3778–3784, 2019.

Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2776–2784, 2023.

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

Pengfei Ye, Han Qin, Xiaojun Zhan, Zhan Wang, Chang Liu, Beibei Song, Yaru Kong, Xinbei Jia, Yuwei Qi, Jie Ji, et al. Diagnosis of obstructive sleep apnea in children based on the xgboost algorithm using nocturnal heart rate and blood oxygen feature. *American Journal of Otolaryngology*, 44(2):103714, 2023.

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

Xinyi Zhang, Hang Dong, Zhe Hu, Wei-Sheng Lai, Fei Wang, and Ming-Hsuan Yang. Gated fusion network for joint image deblurring and super-resolution. *arXiv preprint arXiv:1807.10806*, 2018.

Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7469–7478, 2019.

Xiaoyun Zhao, Xiaohong Wang, Tianshun Yang, Siyu Ji, Huiquan Wang, Jinhai Wang, Yao Wang, and Qi Wu. Classification of sleep apnea based on EEG sub-band signal characteristics. *Scientific Reports*, 11(1):5824, 2021.

Mark Zimmerman, Joseph B McGlinchey, Diane Young, and Iwona Chelminski. Diagnosing major depressive disorder i: A psychometric evaluation of the dsm-iv symptom criteria. *The Journal of nervous and mental disease*, 194(3):158–163, 2006.

## Appendix A. Transformer Backbone

The encoder and decoder consist of multiple layers of transformer, each consisting of two components. Both components have a residual connection and a normalization layer. The first is the multi-head self-attention (MHSA), which allows the different parts of the input sequence to interact:

$$X^{'} = LayerNorm(X) \tag{19}$$

$$MHSA(X^{'}) = Concat(h_1, ..., h_n)W^C \tag{20}$$

where:

$$h_i = Softmax(\frac{W_i^K X^{'}(W_i^N X^{'})^T}{\sqrt{d_k}})W_i^V X^{'}. \tag{21}$$

The second component is a fully connected network (FCN) which applies a nonlinear transformation to each position in the sequence:

$$X^{''} = LayerNorm(X^{'} + MHSA(X^{'})) \tag{22}$$

$$FCN(X^{''}) = ReLU(X^{''}W_1 + b_1)W_2 + b_2, \tag{23}$$

$$output = X^{''} + FCN(X^{''}) \tag{24}$$

$W_i^N, W_i^K, W_i^V$ are learnable weights related to the head $i$ of the self-attention module. $d_k$ is the dimension of queries and keys in self-attention. ($W_1$, $b_1$) and ($W_2$, $b_2$) are the learnable weights and biases for the first and second layers, respectively.

## Appendix B. Additional training 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. We use L2 weight regularization with $\lambda = 10^{-3}$ and dropout (rate=0.25) to avoid overfitting. We trained the model for 100 epochs in a 5-fold cross-validation manner. our model was implemented using Keras (Chollet et al., 2015) inside the TensorFlow (Abadi et al., 2016) framework.

**Hyper-parameters tuning** We performed a grid search to empirically find the best values. We ran the experiments with models with different numbers of transformer layers and heads in each layer. For our task, the best performance achieved with the model has 5 layers and 4 heads in each layer. in our architecture, each layer consists of two layers of MLP with 16 and 32 units. Fully connected layers in a gated fusion network have 8 units, which have been selected among 4, 8, 16, and 32 units.

## Appendix C. Extended Results

The performance of SOTA models in terms of AUROC with the concurrent occurrence of noise and missingness on the NCH dataset is shown in Table 5. The performance of the proposed model in terms of AUROC when just utilizing ECG signals with Baseline wandering and Powerline interference noises are shown in Table 4.

| Noise type | AUROC | |
|---|---|---|
| | CHAT | NCH |
| Baseline wandering | 82.2 (0.7) | 82.5 (0.7) |
| Powerline interference | 82.4 (0.8) | 82.8 (0.7) |
| Without Noise | 83.1 (0.6) | 83.4 (0.6) |

Table 4: Performance of our model with noisy ECG signals. Mean(± STD)

## Appendix D. Noise and missingness algorithms

Additive white Gaussian noise generation and Random epoch-channel omission are shown in algorithms 2 and 1, respectively.

**Algorithm 1:** Random epoch-channel omission

**Input:** $\{S_i\}_{i=1}^{I}, omission\_ratio$
**Output:** $\{S_i\}_{i=1}^{I}$
**for** $i \leftarrow 1$ ***to*** $I$ **do**
  **for** $j \leftarrow 1$ ***to*** $J_i$ **do**
    **for** $m \leftarrow 1$ ***to*** $M$ **do**
      $rnd \leftarrow$ Generate a random number in (0,1)
      **if** $rnd \leq omission\_ratio$ **then**
        $X_m^{i,j} = 0$
    **end**
  **end**
**end**

## Appendix E. Model performance across different ages

We separated out patients according to their age range to study the performance of the methods across different ages. The results for the NCH dataset are shown in Figure 3. One can observe that the model's discriminative performance remains consistently over 90%, while the performance is slightly lower in younger ages. The patients' age distribution is shown in Figure 3.

Table 5: Performance of the baselines and our model in terms of AUROC with concurrent occurrence of noise and missingness on NCH dataset. The mean (standard deviation) values are shown.

| Missing ratio | SNR | CNN-LSTM | CNN | Transformer | Ours |
|---|---|---|---|---|---|
| 10% | 10 | 57.0(3.3) | 58.8(2.0) | 82.1(0.2) | 84.9(1.9) |
| | 20 | 66.6(5.5) | 73.7(1.1) | 83.0(0.3) | 90.4(0.9) |
| | 30 | 80.4(0.9) | 80.5(0.7) | 84.6(0.5) | 92.6(0.5) |
| | 40 | 81.8(0.9) | 81.4(1.0) | 86.3(0.5) | 92.8(0.5) |
| | 50 | 81.8(1.0) | 81.2(0.9) | 86.5(0.3) | 92.8(0.5) |
| 20% | 10 | 57.4(2.4) | 58.8(1.7) | 79.4(0.2) | 82.9(1.9) |
| | 20 | 64.6(3.7) | 69.0(0.7) | 80.0(0.5) | 89.0(1.1) |
| | 30 | 74.9(0.5) | 74.1(0.5) | 81.8(0.4) | 91.8(0.6) |
| | 40 | 75.9(0.7) | 74.7(0.9) | 83.1(0.4) | 92.0(0.5) |
| | 50 | 75.9(0.9) | 74.5(1.0) | 83.3(0.4) | 92.1(0.5) |
| 30% | 10 | 56.8(2.0) | 58.6(1.2) | 76.2(0.5) | 80.9(2.0) |
| | 20 | 62.3(2.2) | 65.2(0.6) | 77.1(0.3) | 87.3(1.0) |
| | 30 | 70.2(0.6) | 68.7(0.5) | 78.3(0.5) | 90.6(0.6) |
| | 40 | 70.5(1.2) | 69.3(0.8) | 79.5(0.5) | 91.1(0.5) |
| | 50 | 70.4(1.2) | 69.3(0.8) | 80.0(0.5) | 91.1(0.5) |
| 40% | 10 | 56.6(0.9) | 57.6(0.9) | 73.1(0.7) | 78.6(1.9) |
| | 20 | 61.0(1.6) | 62.4(0.5) | 73.8(0.9) | 85.4(1.2) |
| | 30 | 66.0(0.4) | 64.7(0.3) | 74.8(0.8) | 89.3(0.6) |
| | 40 | 66.3(1.1) | 65.0(0.7) | 76.1(1.0) | 89.9(0.5) |
| | 50 | 66.4(1.1) | 64.9(0.9) | 76.3(1.1) | 89.8(0.5) |
| 50% | 10 | 56.0(1.0) | 56.6(0.9) | 69.7(0.9) | 76.9(2.0) |
| | 20 | 58.6(1.4) | 59.7(0.5) | 70.4(0.9) | 83.4(1.0) |
| | 30 | 62.3(0.9) | 60.9(0.4) | 71.3(0.8) | 87.6(0.6) |
| | 40 | 62.6(0.8) | 61.1(0.7) | 72.3(1.2) | 88.0(0.6) |
| | 50 | 62.5(0.4) | 61.2(0.7) | 72.7(1.2) | 88.2(0.4) |

## Appendix F. Computationl complexity

Addressing concerns about the potentially high number of trainable parameters and computational costs associated with unimodal modules, we compared the trainable parameters of our proposed model with the aforementioned baseline models, shown in 6.

**Algorithm 2:** Additive white Gaussian noise generation

**Input:** $\{S_i\}_{i=1}^{I}, target\_snr, noise\_occurrence\_chance$
**Output:** $\{S_i\}_{i=1}^{I}$
**for** $i \leftarrow 1$ ***to*** $I$ **do**
  **for** $j \leftarrow 1$ ***to*** $J_i$ **do**
    **for** $m \leftarrow 1$ ***to*** $M$ **do**
      $signal\_average\_power = \frac{\sum_{t=1}^{T}(X_{t,m}^{i,j})^2}{T}$
      $signal\_average\_power\_db = 10 * log_{10}(signal\_average\_power)$
      $noise\_average\_power\_db = signal\_average\_power\_db - target\_snr$
      $noise\_average\_power = 10^{noise\_average\_power\_db}$
      $noise \sim Normal(0, \sqrt{noise\_average\_power})$
      $rnd \leftarrow$ Generate a random number in (0,1)
      **if** $rnd \leq noise\_occurrence\_chance$ **then**
        $X_m^{i,j} \leftarrow X_m^{i,j} + noise$
    **end**
  **end**
**end**

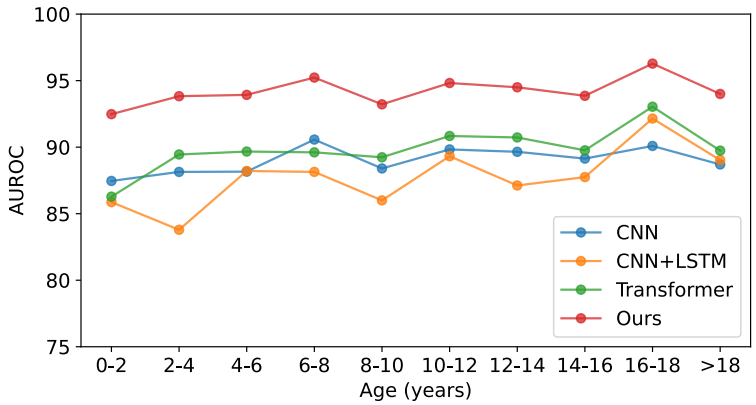


Figure 3: Performance of the proposed and state-of-the-art models across different age groups on the NCH dataset.

| Model | Number of trainable parameters |
|---|---|
| CNN | 177K |
| CNN+LSTM | 245K |
| Transformer | 138K |
| Ours | 446K |

Table 6: Comparison of trainable parameters across different architectures.