Contactless Oxygen Monitoring with Radio Waves and Gated Transformer

Hao He[*†] HAOHE@MIT.EDU
Massachusetts Institute of Technology

Yuan Yuan[*†] MIAYUAN@MIT.EDU
Massachusetts Institute of Technology

Ying-Cong Chen[*] YINGCONG.IAN.CHEN@GMAIL.COM
Hong Kong University of Science and Technology (Guangzhou)

Peng Cao PENGCAO@MIT.EDU
Massachusetts Institute of Technology

Dina Katabi DK@MIT.EDU
Massachusetts Institute of Technology

Abstract

With the increasing popularity of telehealth, it is crucial to ensure accurate monitoring of basic physiological signals at home with minimal patient overhead. In this paper, we propose a contactless approach for monitoring blood oxygen levels simply by analyzing radio signals in a patient's room, without the need for wearable devices. Our method extracts a patient's respiration from radio signals that bounce off their body, and we use a novel neural network, called *Gated BERT-UNet*, to estimate blood oxygen saturation from the breathing signal. We designed our model to adapt to a patient's medical indices, such as gender and sleep stages, to provide personalized inference. Specifically, it uses multiple predictive heads, controlled by a gate, to make predictions for different sub-populations. Our extensive empirical results demonstrate that our model achieves high accuracy on both medical and radio-frequency datasets. It outperforms past work on contactless oxygen monitoring, reducing the mean absolute error in oxygen saturation from 2.0% to 1.3%.

1. Introduction

Remote health monitoring and telehealth have gained popularity in recent years due to their ability to reduce healthcare costs and improve access to healthcare, particularly for those in remote areas (Al-Khafajiy et al., 2019). Additionally, remote health monitoring can help track a patient's long-term physiological state, including older individuals who live alone at home, and enable caregivers to provide timely assistance (Celler et al., 1995; Tian et al., 2018; Fan et al., 2020b). However, the success of remote health monitoring services depends on the availability of solutions that can continuously monitor a person's physiological signals at home with minimal patient overhead.

[*] Equal contribution. Yingcong participated in this work when he was a postdoc at MIT.
[†] Corresponding authors. The theoretical results were mainly derived by Hao.

Oxygen saturation is an important physiological signal whose at-home monitoring would benefit very old adults and individuals at high risk for low blood oxygen (Moss et al., 2005). Oxygen saturation refers to the amount of oxygen in the blood, measured as the fraction of oxygen-saturated hemoglobin relative to the total blood hemoglobin. Normal oxygen levels range from 94% to 100%. Levels below this range can be dangerous, if severe can lead to brain and lung failure (Díaz-Regañón et al., 2002; Lapinsky et al., 1999).

Currently, measuring oxygen saturation requires wearing a pulse oximeter on the finger and actively measuring oneself. While pulse oximeters are helpful, they can be impractical in some at-home monitoring scenarios. Very old adults in their late 80s and 90s, who are at high risk for low blood oxygen (NLM, 2020), should regularly monitor their oxygen levels, but many suffer from dementia or cognitive impairment, which prevents them from measuring themselves. Patients recovering from COVID-19 or pneumonia may experience delirium (Han et al., 2020), which can affect their reasoning and ability to measure their oxygen levels. Additionally, blood oxygen levels tend to drop during sleep, making it particularly important to track oxygen levels overnight (Palma et al., 2008; Gries and Brooks, 1996). However, people cannot actively measure themselves while asleep.

The above use cases motivate us to try to complement pulse oximetry with a new approach that can work passively and continuously, assessing blood oxygen throughout the night without requiring the person to wear a sensor or actively measure themselves. Prior attempts at estimating blood oxygen passively, without a wearable sensor, rely on cameras (Mathew et al., 2021; Van Gastel et al., 2016; van Gastel et al., 2019; Shao et al., 2015; Bal, 2015; Guazzi et al., 2015). For example, (Mathew et al., 2021) proposes a convolution neural network that analyzes a video of the person's palm to estimate his/her blood oxygen. While this method does not require wearable sensors, it cannot work continuously since the user cannot keep their hand in front of the camera for a long time. It also cannot operate in dark settings and thus cannot monitor one's oxygen during sleep. To bypass the limitations imposed by cameras, we propose a different sensing modality, radio-frequency (RF) signals.

We propose to monitor oxygen saturation by analyzing the radio signals that bounce off a person's body. Recent research has demonstrated the feasibility of monitoring breathing, heart rate, and even sleep stages by transmitting a very low power RF signal and analyzing its reflections off a person's body (Adib et al., 2015; Yue et al., 2018). Building on these advances, we use RF signals to track a person's breathing signal and train a neural network model to infer oxygen from respiration. Such a design can measure a person's oxygen without any physical contact or wearable devices. Thus, it does not burden the patient or interfere with their sleep. Further, since RF signals are independent of lighting conditions, our approach can monitor oxygen saturation (SpO2) during sleep and in dark settings. In addition, using breathing as an intermediate representation allows us to leverage existing medical datasets that contain continuous breathing signals paired with SpO2 measurements. This enables us to train our model on a large dataset and directly test it on breathing signals extracted from RF signals.

We propose a deep learning model for inferring oxygen saturation from breathing signals, with a focus on personalized prediction by adapting to the patient's medical indices. In many cases, medical indices such as gender and disease diagnoses are binary or categorical. However, shown by our theoretical analysis, simply adding them as additional inputs to the model can lead to sub-optimal performance. To address this, we propose Gated BERT-

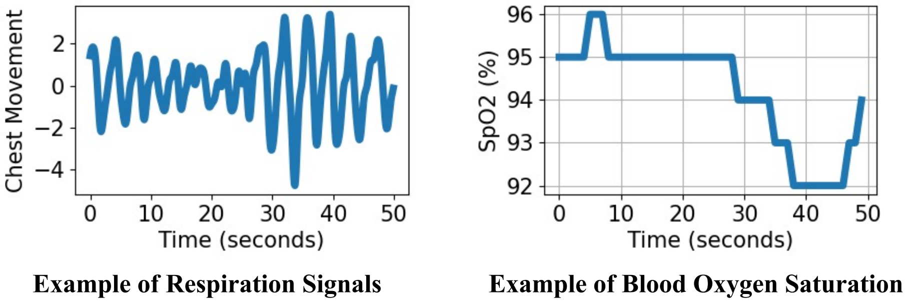
UNet, a transformer model with multiple predictive heads, each suited to a specific subset of patients based on their categorical indices. Our approach allows for both simple index like gender/race and dense categorical variables such as a time series of sleep stages. To evaluate our model, we conduct experiments on medical and RF datasets. Our results demonstrate an average absolute error in predicting oxygen saturation of 1.3%, which outperforms state-of-the-art camera-based models (Mathew et al., 2021). Our contributions are twofold:

- The paper introduces the task of inferring oxygen saturation from RF signals, and the task of inferring oxygen from breathing signals. Our approach allows for passive and continuous tracking of oxygen saturation at home, a task that could benefit older adults and facilitate overnight oxygen monitoring.
- The paper introduces *Gated BERT-UNet*, a novel transformer model that can adapt to auxiliary categorical variables. Experimental results show good performance on both medical and radio-frequency datasets.

Generalizable Insights about Machine Learning in the Context of Healthcare Healthcare applications usually requires machine learning model to adapt to different sub-populations, such as race/gender groups. The most common approach is to provide the model with those group indices as input, but our analysis shows that this might not be the best solution when the target function changes drastically across the group, as models can have intrinsic biases that smooth to the input. To address this issue, We propose a novel gating mechanism. It offers several advantages: (1) allows both scalar indices like race/gender and time series indices like sleep stages overnight; (2) allows indices that are not accessible during inference by an internal estimation; (3) automatically identifies groups that have similar target functions to share a gate which improves the learning efficiency.

2. Related Work

Monitoring Oxygen Saturation. The most exact measurements of oxygen saturation are invasive and require arterial blood samples. The non-invasive and widely-common method for measuring oxygen saturation (SpO2) uses a pulse oximeter, a small device worn on the finger. To enable remote SpO2 measurement, extensive research investigate use camera to perform photoplethysmography (Van Gastel et al., 2016; van Gastel et al., 2019; Shao et al., 2015; Bal, 2015; Guazzi et al., 2015). However those methods have limitations like susceptible to noise, sensitive to motion, requiring ambient light. Recently, deep learning has been considered to aid SpO2 monitoring. Ding et al. tried to monitor SpO2 using smartphones. But their solution requires the fingertip to be pressed against the camera, and hence cannot provide continuous overnight measurements. A more recent work (Mathew et al., 2021) has estimated SpO2 in a contactless way with regular RGB cameras. Their method first extracts the region of interest from the video of the person's palm, then uses a CNN model to estimate SpO2. While this approach is contactless, it still requires the user to keep their hand in front of the video camera for the duration of the monitoring, which is not practical for continuous or overnight monitoring. Our work differs from all prior works in that we predict oxygen from other modalities, namely breathing or radio waves, and enable oxygen sensing in a completely contactless and passive manner.



Example of Respiration Signals **Example of Blood Oxygen Saturation**

Figure 1: Breathing signal and corresponding oxygen (SpO2) signal.

Contactless Health Sensing with Radio Signals. The past decade has seen a rapid growth in research on passive sensing using RF signals. Early work has demonstrated the possibility of sensing one's breathing and heart rate using radio signals (Adib et al., 2015). Later, researchers have shown that by analyzing the RF signals that bounce off the human body, they can monitor a variety of health metrics including sleep, gait, falls, and even human emotions (Zhao et al., 2017; Hsu et al., 2017; Tian et al., 2018; Zhao et al., 2016; Fan et al., 2020a). Our work is the first to enable SpO2 monitoring from radio waves.

Adaptation of ML Models to Medical Indices. Previous deep learning models that operate on physiological signals typically do not adapt to a person's medical indices. For instance, models that infer sleep stages from respiration (Zhao et al., 2017), detect arrhythmia from ECG (Kiyasseh et al., 2020), and classify emotion from EEG signals (Murugappan et al., 2010) do not account for individual differences in medical indices. A recent survey (Rim et al., 2020) collected 147 papers on learning with physiological signals, and showed that none of them adapts to a person's medical indices. While a few approaches exist for leveraging side variables in deep learning models, they typically treats such variables as additional input features when they are available at inference time (Narayan et al., 2017; Shen et al., 2016). For variables that are only accessible during training, they are often used as extra supervisors to regularize the model via multi-task learning (Liu et al., 2019; Mordan et al., 2018). In this paper, we propose a novel gating mechanism that better handles categorical variables, and we show that it outperforms these previous approaches.

3. Our System: RF-Oximeter

3.1. Overview

As introduced above, we aim to develop a system, which we call RF-Oximeter, that estimates oxygen saturation from radio frequency signals in a contactless manner. Figure 1 shows an example of the breathing signals (our model's input) and the corresponding ground truth oxygen (SpO2) time series (our model's target). Our proposed RF-Oximeter contains two steps, elaborated as follows.

Breathing signal extraction from RF signals. We leverage past work on extracting breathing signals from the RF signal that bounces off people's bodies. Specifically, our system is equipped with a multi-antenna Frequency-Modulated Carrier Waves (FMCW)

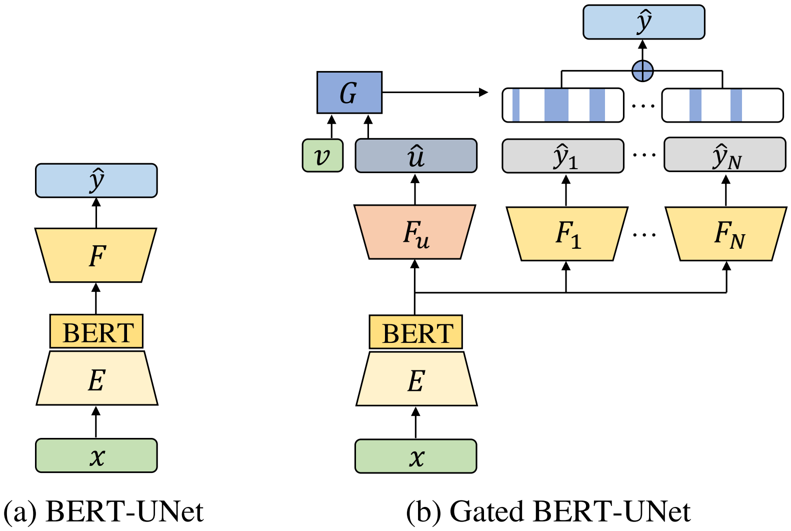


Figure 2: Illustrations of the proposed models: (a) The backbone BERT-UNet model, which contains an encoder E and a predictor F as the UNet structure, and a BERT module at the bottleneck of the UNet. (b) The Gated BERT-UNet, where a gate G, controlled by the accessible variable v and the predicted inaccessible variable $\hat{u}$, is used to select among multiple predictive heads.

radio, which is commonly used in passive health monitoring (Rahman et al., 2015; Fan et al., 2020b). The radio transmits a very low power RF signal and captures its reflections from the environment. We process these reflections using the algorithm in Yue et al. (2018) to infer the subject's breathing signal. Past work shows that breathing signals extracted in this manner are highly accurate; Specifically, their correlation with an FDA-approved breathing belt on the person ranges from 91% to 99%, depending on the distance from the radio and the distance between people (Yue et al., 2018).

Oxygen saturation estimation from breathing signals. In the following paragraphs, we describe our deep learning model, which takes as input a breathing signal and predicts the corresponding oxygen time series for the same interval. We propose a novel architecture that leverages the subject's physiological indices to improve the accuracy of oxygen prediction.

3.2. Machine Learning Task Formulation

we formulate the problem of oxygen saturation prediction as a sequence-to-sequence dense regression task, where the respiration signals are denoted by $x \in \mathbb{R}^{1\times f_b T}$ and the corresponding oxygen signals are denoted by $y \in \mathbb{R}^{1\times f_o T}$. The model h takes the respiration signal over a T-second interval and predicts the oxygen level over the same period, i.e., $y^1, \cdots, y^{f_o T} = h(x^1, \cdots, x^{f_b T})$, where superscripts denote time steps, and f_b and f_o are the sampling frequencies of respiration and oxygen, respectively. The default sampling rates for respiration and oxygen are $f_b = 10Hz$ and $f_o = 1Hz$.

3.3. Backbone Model: BERT-UNet

In this section we propose our backbone model. In the next section, we augment this backbone to adapt to the person's medical indices. Our backbone model, BERT-UNet, is a combination of a BERT module (Devlin et al., 2018) and UNet (Ronneberger et al., 2015). As shown in Figure 2(a), our BERT-UNet model consists of an encoder $E(\cdot;\theta_e)$ and an oxygen predictor $F(\cdot;\theta_f)$. The encoder is composed of a fully convolutional network (FCN) followed by a bidirectional transformers (BERT) module (Devlin et al., 2018). The FCN extracts local features from the raw respiration signals, then the BERT module captures long-term temporal dependencies based on those features. The predictor F is composed of several deconvolutional layers, which up-samples the extracted features to the same time resolution of oxygen saturation. Formally, we have $E : \mathbb{R}^{1\times f_b T} \to \mathbb{R}^{n\times \alpha f_b T}$ and $F : \mathbb{R}^{n\times \alpha f_b T} \to \mathbb{R}^{1\times f_o T}$ where n is the dimension of the respiration feature and α is the down-sampling factor ($\alpha = 1/240$ in our experiments). The model is trained with a combination of the L_1 loss and the correlation loss given below:

$$\mathcal{L}(\hat{y}, y) = \frac{\|\hat{y} - y\|_1}{f_o T} - \lambda \frac{\sum_i (\hat{y}^i - \mu_{\hat{y}})(y^i - \mu_y)}{\sqrt{\sum_i (\hat{y}^i - \mu_{\hat{y}})^2 \sum_i (y^i - \mu_y)^2}}. \quad (1)$$

Here $\hat{y} = F(E(x;\theta_e);\theta_f)$ is the model prediction, y is the ground truth oxygen, μ_y and $\mu_{\hat{y}}$ are the mean values of y and $\hat{y}$, and λ is a hyper-parameter to balance the two loss terms. We choose the L_1 loss over other regression loss functions, since it is more robust to outliers and empirically has better performance. We also find that the correlation loss helps in matching the fluctuations of the predicted oxygen to the ground truth.

3.4. Leveraging Categorical Bio-indices

In most medical applications, there are useful side variables. Adapting to such variables will likely improve performance and make the results more personalized. The standard way for incorporating such variables in a deep learning model is to provide them as extra input features. We argue however that such a design is not suitable for categorical variables. In many tasks, the relevant medical variables are binary or categorical. For example, the relevant variables for oxygen saturation include gender, whether the person is a smoker, whether they have asthma, etc. The binary nature of these variables induces discontinuity in the learned function over the physiological indices. Take our task as an example. Consider oxygen saturation $y = g_s(x)$ as a function of respiration x and gender s (0 for male and 1 for female). $g_0(x)$ and $g_1(x)$ can be two different functions since men and women have differences in their oxygen transport systems (Reybrouck and Fagard, 1999). Below, we show that, if these functions are significantly different, simply introducing the variable as an extra input feature is not an appropriate model.

Formally, consider a learning target $y = g_s(x)$, which is a mapping from data $x \in \mathbb{R}^n$ to label $y \in \mathbb{R}$. Let the mapping depend on a binary variable $s \in \{0, 1\}$. Say we learn the mapping via a neural network $f(W[x, s])$ whose input $[x, s]$ is the concatenation of data x and variable s. W denotes the weight matrix of the first layer. Assuming the data x follows the distribution $p(x)$, we define the error of any neural network with respect to the ground truth as:

$$\mathcal{E}(f, W) = \mathbb{E}_{x \sim p(x)}\left[\sum_{s=0}^{1}(f(W[x, s]) - g_s(x))^2\right]. \tag{2}$$

In the Appendix A, we prove the following lower bound.

Theorem 1 (Informal) *If the neural network (f, W)'s first layer weight matrix has a column rank deficiency, then its error is lower-bounded as follows:*

$$\mathcal{E}(f, W) \geq \mathcal{D}(g_0, g_1) \triangleq \min_{c \in R^m} \mathcal{D}_c(g_0, g_1)$$

$$\mathcal{D}_c(g_0, g_1) \triangleq \mathbb{E}_x\left[\frac{p(x+c)(g_0(x) - g_1(x+c))^2}{p(x) + p(x+c)}\right], \forall c \in \mathbb{R}^n$$

Remark. Theorem 1 says any model that takes the side variables as input (with a column rank deficient weight W) cannot perfectly fit the ground truth function, and its error is at least as large as the "distance" between g_0 and g_1. For any vector c, we interpret $\mathcal{D}_c(g_0, g_1)$ as the distance between $g_0(x)$ and $g_1(x+c)$ since $\mathcal{D}_c(g_0, g_1) = 0$ if $g_0(x) \equiv g_1(x+c)$. Similarly, $\mathcal{D}(g_0, g_1)$ can be interpreted as the minimum expected distance under any shift alignment in the input space. Further, note that the condition that the weight matrix being column rank deficient is practical. Neural networks may learn a low rank weight matrix since sometimes they compress the information during training. For example, convolutional neural network (CNN) learn highly structured filters in the low-level layers, which are intrinsically low rank (Nakkiran et al., 2015). Note that our theorem does not violate the universal approximation theorem of neural network (Cybenko, 1989; Hornik, 1991) since it puts rank constrains on the weights.

3.5. Final Model: Gated BERT-UNet

Motivated by our theoretical insight, we propose *Gated BERT-UNet*, a new model that augments BERT-UNet with multiple predictive heads. It selects the most suitable head for a person via a gate controlled by the subject's categorical indices. The model supports both variables available at the time of inference (e.g., gender), as well as dense categorical variables concurrently learned from the input signals (e.g., sleep stages).

Figure 2(b) illustrates the design of our model. It has a gate function $G(v, u) : \mathcal{V} \times \mathcal{U} \to \{1, 2, \cdots, N\}$ where $v \in \mathcal{V}$ and $u \in \mathcal{U}$ are accessible/inaccessible variables, N is the number of gate statuses. We use the term *accessible variable* to refer to variables easily available during inference time, e.g., gender and the term *inaccessible variable* to refer to information that is not available during inference but contained in the train dataset, like a person's sleep stages. Inaccessible variables are typically dense time series (e.g., sleep stages). Their prediction are learned concurrently with the main task under a full supervision. The construction of the gate function $G(v, u)$ is described in the next section.

The model has N heads $\{F_i\}_{i=1}^N$ that adapt the prediction $\hat{y}_i = F_i(E(x))$ to the gate status. It also has an extra predictor F_u to infer inaccessible variables u. During testing time, based on the accessible variables v and estimated inaccessible variables $\hat{u}$, we evaluate the gate status $s = G(v, \hat{u})$.

In the case of oxygen prediction, $\hat{y}_i$ and the gate status s are time series. As shown in Figure 2(b), the final prediction at each time step, is the gated combination of every head's output, i.e. $\forall t = 1, \ldots, f_o T, \hat{y}^t = \sum_{i=1}^{N} \mathbf{1}[s^t = i]\hat{y}_i^t$. We train Gated BERT-UNet (GBU) with the following loss,

$$\mathcal{L}_{\texttt{GBU}}(\hat{y}, \hat{u}, y, u) = \mathcal{L}(\hat{y}, y) + \frac{\lambda_u}{f_o T} \sum_{t=1}^{f_o T} \mathcal{L}_{\texttt{CE}}(\hat{u}^t, u^t), \tag{3}$$

where $\mathcal{L}$ is the main loss defined in Equation 1, $\mathcal{L}_{\texttt{CE}}$ is the cross entropy loss to train the branch for predicting inaccessible variables and λ_u is a balancing factor.

3.6. From Medical Indices to Gate Status

The number of heads in a Gated BERT-UNet model puts an upper bound on the number of possible gate statuses. For example, if a Gated BERT-UNet model has 6 heads, then the gate can take only 1 of 6 statuses. Typically, we have much more variable statuses than gate statuses. To find a proper mapping from variable status to gate status, we rely on gradient similarity. For example, say we want to check whether male smokers should be in the same group as female smokers, we take a pretrained backbone BERT-UNet and compute its averaged gradient (w.r.t the loss function) over all male smokers and all female smokers in the dataset. Then we check the cosine similarity between the two gradients. If the gradients are similar, which means the two categories move the loss function in the same direction, we can use the same predictor for them. On the other hand, if the gradients are vastly different, it is preferable to separate such categories and assign them to different gate statuses. In Appendix B, we illustrate this process in the context of our oxygen task.

4. Datasets & Metrics

Medical Datasets. We leverage three publicly available medical datasets: Sleep Heart Health Study (*SHHS*), Osteoporotic Fractures in Men Study (*MrOS*), and Multi-Ethnic Study of Atherosclerosis (*MESA*) (Zhang et al., 2018; Blackwell et al., 2011; Chen et al., 2015). The first two datasets have males and females, whereas the last one has male subjects. These datasets were collected during sleep studies. For each subject, they include the respiration signals throughout the night along with the corresponding blood oxygen saturation time series, where the respiration signals are collected using a breathing belt around the chest or abdomen to measure the inhale-exhale motion, and the oxygen saturation is measured using a pulse oximeter. They also contain a variety of side variables including sleep stages, which for every time instance assign to the subject one of the following: Awake, Rapid Eye Movement (REM) or one of three non-REM stages N1, N2, and N3.

SHHS, MESA, and MrOS contain 2651, 2056, and 1026 subjects, respectively. We randomly split subjects, 70% to the training and validation sets, 30% to the testing sets, and keep the same splits in all experiments. We note that the subjects in these studies have an age range between 40 and 95, and some of them suffer from a variety of diseases such as chronic bronchitis, cardiovascular diseases, and diabetes. This allows for a wider range of oxygen variability beyond the typical range of healthy individuals.

RF Datasets. We collected an RF dataset that contains radio signals paired with SpO2 measurements. The data is collected from two hospital sleep labs. In total, there are 400

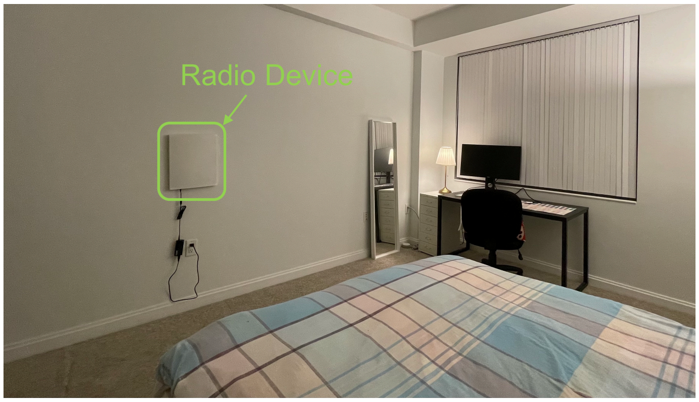


Figure 3: The radio device used to collect RF signals.

Table 1: Performances on medical datasets. Asterisks * indicates model using the physiological variables (*Gender* and *Sleep Stages* in our task). Best results are **boldfaced**.

Model	SHHS			MESA			MrOS			Overall		
	$\text{Corr}^{\uparrow}$	$\text{MAE}^{\downarrow}$	$\text{RMSE}^{\downarrow}$	$\text{Corr}^{\uparrow}$	$\text{MAE}^{\downarrow}$	$\text{RMSE}^{\downarrow}$	$\text{Corr}^{\uparrow}$	$\text{MAE}^{\downarrow}$	$\text{RMSE}^{\downarrow}$	$\text{Corr}^{\uparrow}$	$\text{MAE}^{\downarrow}$	$\text{RMSE}^{\downarrow}$
CNN	0.47	1.72	1.81	0.46	1.62	1.72	0.49	1.77	1.82	0.47	1.73	1.78
CNN-RNN	0.48	1.70	1.80	0.50	1.56	1.65	0.52	1.76	1.84	0.49	1.67	1.77
BERT-UNet	0.51	1.67	1.77	0.52	1.55	1.66	0.54	1.75	1.84	0.51	1.64	1.76
BERT-UNet + VarAug*	**0.52**	1.65	1.76	0.52	1.51	1.62	**0.55**	1.68	1.78	0.52	1.61	1.72
Gated BERT-UNet*	**0.52**	**1.61**	**1.72**	**0.53**	**1.50**	**1.61**	**0.55**	**1.65**	**1.75**	**0.53**	**1.58**	**1.70**

hours of data from 49 overnight recordings of 32 subjects. The dataset contains subjects of different genders and races. Some subjects are healthy volunteers while others are patients with sleep problems. Thus, the ground truth oxygen saturation distribution is wider than normal ranges. A radio device is installed in the room to collect RF signals, as shown in Fig. 3. The radio signals are synchronized with the SpO2 measurements and processed with the algorithm in Yue et al. (2018) to extract the person's breathing signals.

Metrics. Let $\hat{y}$, y denote the predicted and ground truth oxygen saturation. Following the previous work (Mathew et al., 2021), we use three standard metrics for evaluation: (1) *Correlation*: $\frac{\sum_t (y^t - \mu_y)(\hat{y}^t - \mu_{\hat{y}})}{\sqrt{\sum_t (y^t - \mu_y)^2 \sum_t (\hat{y}^t - \mu_{\hat{y}})^2}}$ where μ_y and $\mu_{\hat{y}}$ are the averaged oxygen saturation; (2) *Mean Averaged Error (MAE)*: $\frac{1}{T}\sum_{t=1}^{T} |y^t - \hat{y}^t|$; (3)*Rooted Mean Squared Error (RMSE)*: $\sqrt{\frac{1}{T}\sum_{t=1}^{T}(y^t - \hat{y}^t)^2}$. In our experiments, the models are trained to take a subject's whole night respiration signals as input and predict the corresponding oxygen levels. However, during the evaluation, we divide model's prediction and ground truth oxygen into non-overlapping 240-second segments to compute the metrics. It is because that the correlation metrics are sensitive to the segment length. For a fair comparison, we follow (Mathew et al., 2021) and use 240-second intervals. More details can be found in Appendix B.

5. Experiment

We first evaluate and compare different variants of our model in both medical and RF datasets. We further compare our model to the SOTA in contactless monitoring of SpO2.

Baselines. To compare the performance of different 'backbone' models, we consider: (a) *CNN* is a fully convolutional model composed of eight 1-D convolutional layers and seven 1-D deconvolutional layers; (b) *CNN-RNN* augments the CNN model with a recurrent unit in the bottleneck to better captures the long-term temporal relationships of the data. (c) *BERT-UNet* further makes two improvements to the CNN-RNN model. First, it replaces the recurrent unit with an attention module to improve global temporal modeling. Second, it adds skip links between encoding and decoding layers to better capture the signals' locality. To assess the effectiveness of different 'schemes for incorporating bio-indices', we consider: (a) *BERT-UNet + VarAug* uses BERT-UNet as its backbone and takes accessible variables as extra inputs and inaccessible variables as auxiliary tasks. (b) Our *Gated BERT-UNet* model uses multi-heads that are gated by physiological variables.

Train and Evaluation Protocols. Since the RF dataset is not large enough for training, we train all the models on the medical datasets and save the RF data for the test. Specifically, all models are trained on the union of the training sets from the three medical datasets, and tested on each test set, and on the union of the three test sets. We then directly test those models (trained on respiration signals from the medical datasets) on the RF dataset, i.e., do inferences on the respiration signals extracted from the RF signals.

Side Variables. We use gender as the accessible variable and sleep stages as the inaccessible variable. In Gated BERT-UNet, we use gradient similarity (details in Appendix B) to map gate status as described in the method section, which results in the following 6 categories: (male, awake), (male, REM), (male, N1+N2+N3), (female, awake), (female, REM), (female, N1+N2+N3). The sleep stages themselves are learned from the input since they are an inaccessible variable. In the baseline VarAug, the gender variable is provided as an additional input and the sleep stages are used as an auxiliary task in a multitask model.

5.1. Evaluation on Medical Datasets

Quantitative Results. The results are shown in Table 1. Since there is no past work that predicts oxygen from breathing or radio signals, all models in the table refer to variants of our neural network. The table shows that all variants achieve relatively low prediction errors with an average MAE that ranges from 1.58 to 1.73 percent, and an average RMSE that ranges from 1.70 to 1.78 percent. Such a relatively low RMSE shows our model rarely has predictions that largely deviate from the ground truth. All variants also achieve reasonable high correlations ranging from 0.47 to 0.53. Such a correlation level indicates our model's prediction capture the dynamics of the ground truth SpO2 which is also visually shown in Figure 4 and Figure 6. These quantitative results highlight that our system can be useful for continuous monitoring of patients' oxygen at home to alert patients for changes in their oxygen saturation, which they may check with their doctor or health professional.

The upper rows in the table show the results of variants that do not leverage side variables. The table shows that the BERT-UNet model consistently outperforms the CNN model and the CNN-RNN model on all datasets, in all metrics. This indicates that BERT-UNet is a preferable architecture for this task. The bottom rows in the table show the results of models leveraging accessible and inaccessible medical variables. The table shows that BERT-UNet + VarAug and Gated BERT-UNet outperform models that do not leverage

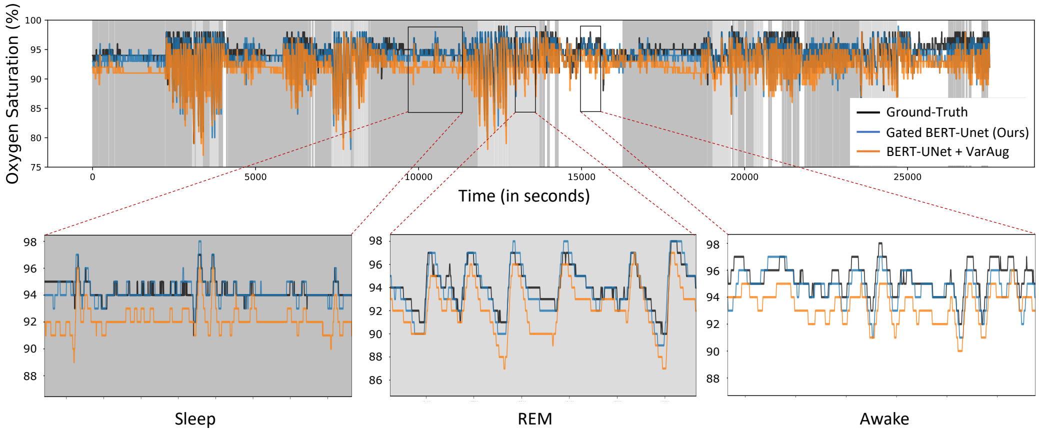
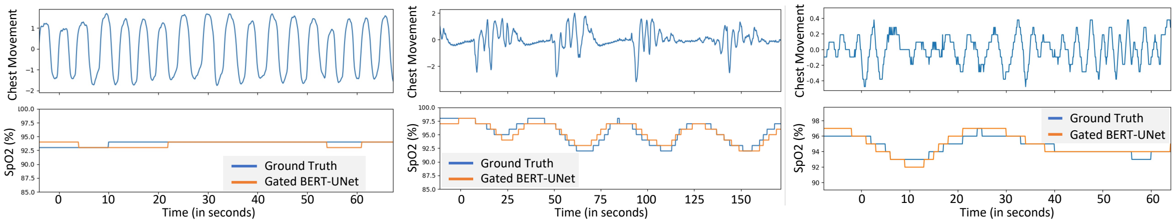


Figure 4: An illustrative example of the oxygen level predicted by the BERT-UNet + VarAug baseline (Orange Curve), and our Gated BERT-UNet model (Blue Curve). The background color indicates sleep stages. The 'dark grey', 'light grey' and 'white' corresponds to 'non-REM Sleep', 'REM' and 'Awake'.

physiological indices, demonstrating the benefit of leveraging such variables. In addition, Gated BERT-UNet outperforms VarAug on all three datasets, demonstrating that a gated multi-head approach works best for such categorical side variables.

Qualitative Results. To better understand how the model works, we visualize a few examples of its prediction. Figure 4 visualizes the predicted oxygen saturation of the Gated BERT-UNet model and the VarAug model on a male subject in the SHHS dataset. As the ground-truth oxygen saturation are integers, we round the predicted oxygen values. The background color indicates different sleep stages. The 'light grey' and 'white' correspond to 'REM' and 'Awake', respectively. For clarity, we use one color 'dark grey' for all three non-REM 'Sleep' stages, N1, N2, and N3. We observe that Gated BERT-UNet consistently outperforms VarAug over the whole night. The small panel focus on different sleep stages. In general, different sleep stages tend to show different behavior and hence the importance of using a gated model. Specifically, oxygen is typically more stable during non-REM 'Sleep' than during REM and Awake stages. The figures show that Gated BERT-UNet can track the ups and downs in oxygen and is significantly more accurate than VarAug. This experiment demonstrates that the way we incorporate the sleep stages into the model improves performance across different sleep stages. Please see the supplement for visualizations on other datasets.

Next, we demonstrate several visual results on different patterns of breathing signals and the corresponding ground truth/predicted oxygen saturation, as shown in Figure 5. Figure 5(a) shows a normal breathing pattern, which leads to constant oxygen saturation. In contrast, Figure 5(b,c) present two different abnormal breathing signals and the resulting fluctuated oxygen predictions. These figures show the diversity of the oxygen and breathing patterns as well as the complexity of their relationship. The model however is able to capture this relationship for highly diverse patterns.



(a) Normal Breathing (b) Sleep Apnea (c) Irregular Rhythm

Figure 5: Visualization of breathing signals and corresponding predicted oxygen saturation. The top panels are breathing signals. The bottom panels are oxygen saturation with the ground truth in blue and our model's predictions in orange.

Table 2: Performances on the RF dataset.

Model	Corr$^{\uparrow}$	MAE$^{\downarrow}$	RMSE$^{\downarrow}$
CNN	0.45	1.65	1.73
CNN-RNN	0.49	1.85	1.93
BERT-UNet	0.48	1.49	1.58
BERT-UNet + VarAug*	0.49	1.49	1.59
Gated BERT-UNet*	**0.52**	**1.32**	**1.54**

5.2. Evaluation on RF Dataset

The results are shown in Table 2. We observe that all model variants work well on the RF dataset with low prediction error and high correlation. The upper 3 rows compare models that do not leverage physiological variables. As shown, CNN-RNN and Bert-UNet perform significantly better than the vanilla CNN, which shows the importance of modeling temporal information. The lower 2 rows present the results of BERT-UNet+VarAug and Gated BERT-UNet. The performance of BERT-UNet+VarAug is similar to BERT-UNet, while Gated BERT-UNet is better than all other variants. This demonstrates that the gating design leverages auxiliary variables better. Overall, the MAEs, RMSEs and correlations on the RF dataset are comparable to those on the medical datasets. This indicates that our model is directly applicable to respiration signals from RF. We have also visualized the prediction results in Figure 6. As shown, our model can accurately track the fluctuation of ground truth SpO2.

5.3. Results for Different Skin Colors

Since pulse oximeters rely on measuring light absorbance through the finger, they are known affected by skin color and tend to overestimate blood oxygen saturation in subjects with dark skin (Feiner et al., 2007; Sjoding et al., 2020). This issue was also the topic of a recent news outbreak after a large study that looked at tens of thousands of white and black COVID patients found that the "reliance on pulse oximetry to triage patients and adjust supplemental oxygen levels may place black patients at increased risk for hypoxaemia" (Sjoding et al., 2020).

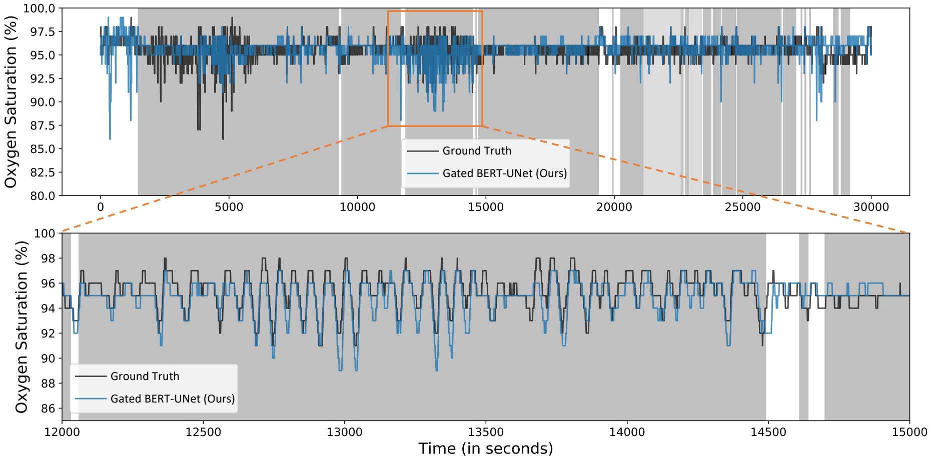


Figure 6: Illustration of the oxygen prediction performance on the RF dataset.

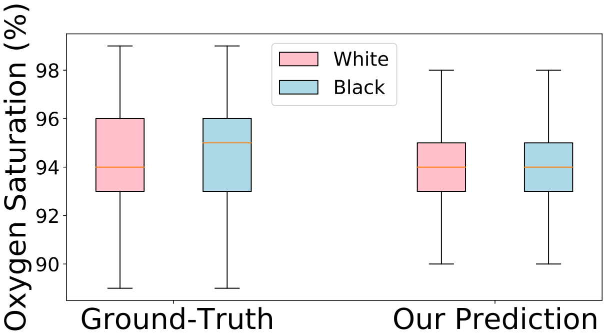


Figure 7: Box plot compares the ground-truth oxygen saturation measured by pulse oximeter and our predicted oxygen saturation for the while and black races.

In contrast, since neither breathing nor RF signals is affected by skin color, a model that estimates oxygen saturation from breathing or RF signals has no intrinsic bias against skin color. To understand model predictions and their relationship to skin color, we plot in Figure 7 the distributions of the ground-truth oxygen and the Gated BERT-UNet predicted oxygen for different races, for the union of all datasets. Interestingly, the ground-truth measurements from oximetry show a clear discrepancy between black and white subjects. In particular, they show that black subjects have higher average blood oxygen. This is compatible with past findings that pulse oximeter overestimates blood oxygen in dark-skinned subjects (Feiner et al., 2007). In contrast, our model appears to correct or reduce the bias. The predictions result in a decrease in the average oxygen saturation for black subjects and demonstrate significantly more similar oxygen distributions for both races.

Remarks. While the primary results presented above are promising, we would like to exercise caution and emphasize the need for a more clinically rigorous study to validate the effectiveness of our approach in mitigating racial bias in blood oxygen measurement.

Table 3: Comparison with previous state-of-the-art oxygen saturation monitoring methods.

Method	Corr↑	MAE↓	RMSE↓	Contactless
Ding et al. 2018	0.26	2.43	2.85	No
Mathew et al. 2021	0.46	1.97	2.16	Yes
Gated BERT-UNet (ours)	**0.52**	**1.32**	**1.54**	Yes

5.4. Comparison with Past Works

We compare our approach with two recent deep-learning camera-based SpO2 monitoring methods. The first method (Ding et al., 2018) makes the user press his/her finger against a smartphone camera, and uses a CNN to estimate SpO2 from the video. The second method (Mathew et al., 2021) uses an RGB camera to capture a video of the person's hand. It then extracts the regions of interest from the video and uses a CNN model to infer SpO2. Although these systems' inputs and datasets are not exactly the same, we make a reasonable comparison by following the setup in Mathew et al. (2021).

Table 3 reports the output of the three models. The results of the baselines are taken from (Mathew et al., 2021), whereas the results of our RF-based model are computed as described. The results show that our model significantly improves the correlation and reduces the MAE and RMSE in comparison to past work.

6. Concluding Remarks

This paper introduces the new task of inferring oxygen saturation from radio signals. It develops a new gated transformer architecture to deliver this application and adapt deep models to auxiliary categorical variables. We believe the proposed design for adapting deep models to auxiliary categorical variables is generally applicable to other tasks to leverage categorical side information.

Our work has a broader impact because it can help in tracking oxygen saturation in older adults during sleep, and providing an initial notification to a caregiver if excessively low oxygen is detected. More extensive measurements with a medical device can then be performed to confirm the results and take medical decisions.

The work also has some limitations. First, the paper focuses on special use cases (e.g., at-home oxygen monitoring in very old adults or during sleep), but is not suitable for other use cases (e.g., measuring oxygen to optimize performance during exercise). Second, the results offer a preliminary demonstration that the morphology and dynamics of breathing signals can provide sufficient information for estimating oxygen saturation. Yet, before integrating this system into clinical practice, it's essential to undertake clinical studies that evaluate its efficacy across varied disease conditions. Notably, factors causing hypoxia and oxygen desaturation can range from mechanical (like hypoventilation) to metabolic or respiratory. The origin of these desaturations may affect our model's precision in determining oxygen levels. Further controlled experiments are necessary to comprehensively gauge the capability of our method. Lastly, it's worth noting that all human subject research presented in this study has been conducted with IRB approval.

Acknowledgments

This work received support from the National Science Foundation and MIT-IBM Watson AI Lab. We express gratitude to Hao Wang, the members of the NETMIT group, and the anonymous reviewers / meta-reviewers for their valuable feedback, insightful comments, and constructive suggestions, which significantly improved the quality of this paper.

References

Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C Miller. Smart homes that monitor breathing and heart rate. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2015.

Mohammed Al-Khafajiy, Thar Baker, Carl Chalmers, Muhammad Asim, Hoshang Kolivand, Muhammad Fahim, and Atif Waraich. Remote health monitoring of elderly through wearable sensors. *Multimedia Tools and Applications*, 78(17):24681–24706, 2019.

Ufuk Bal. Non-contact estimation of heart rate and oxygen saturation using ambient light. *Biomedical optics express*, 6(1):86–97, 2015.

Terri Blackwell, Kristine Yaffe, Sonia Ancoli-Israel, Susan Redline, Kristine E Ensrud, Marcia L Stefanick, Alison Laffan, Katie L Stone, and Osteoporotic Fractures in Men Study Group. Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic fractures in men sleep study. *Journal of the American Geriatrics Society*, 59(12):2217–2225, 2011.

BG Celler, W Earnshaw, ED Ilsar, L Betbeder-Matibet, MF Harris, R Clark, T Hesketh, and NH Lovell. Remote monitoring of health status of the elderly at home. a multidisciplinary project on aging at the university of new south wales. *International journal of bio-medical computing*, 40(2):147–155, 1995.

Xiaoli Chen, Rui Wang, Phyllis Zee, Pamela L Lutsey, Sogol Javaheri, Carmela Alcántara, Chandra L Jackson, Michelle A Williams, and Susan Redline. Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (mesa). *Sleep*, 38(6), 2015.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Genaro Díaz-Regañón, Eduardo Miñambres, Marisol Holanda, Segundo González-Herrera, Francisco López-Espadas, and Carlos Garrido-Díaz. Usefulness of venous oxygen saturation in the jugular bulb for the diagnosis of brain death: report of 118 patients. *Intensive care medicine*, 28(12):1724–1728, 2002.

Xinyi Ding, Damoun Nassehi, and Eric C Larson. Measuring oxygen saturation with smartphone cameras using convolutional neural networks. *IEEE journal of biomedical and health informatics*, 23(6):2603–2610, 2018.

Lijie Fan, Tianhong Li, Rongyao Fang, Rumen Hristov, Yuan Yuan, and Dina Katabi. Learning longterm representations for person re-identification using radio signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020a.

Lijie Fan, Tianhong Li, Yuan Yuan, and Dina Katabi. In-home daily-life captioning using radio signals. In *ECCV*, pages 105–123. Springer, 2020b.

John R Feiner, John W Severinghaus, and Philip E Bickler. Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: the effects of oximeter probe type and gender. *Anesthesia & Analgesia*, 2007.

Robert E Gries and Lee J Brooks. Normal oxyhemoglobin saturation during sleep: how low does it go? *Chest*, 110(6):1489–1492, 1996.

Alessandro R Guazzi, Mauricio Villarroel, Joao Jorge, Jonathan Daly, Matthew C Frise, Peter A Robbins, and Lionel Tarassenko. Non-contact measurement of oxygen saturation with an rgb camera. *Biomedical optics express*, 6(9):3320–3338, 2015.

Dong Han, Chenyang Wang, Xiaojing Feng, and Jing Wu. Delirium during recovery in patients with severe covid-19: Two case reports. *Frontiers in Medicine*, 7, 2020. ISSN 2296-858X. doi: 10.3389/fmed.2020.573791. URL https://www.frontiersin.org/article/10.3389/fmed.2020.573791.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

Chen-Yu Hsu, Yuchen Liu, Zachary Kabelac, Rumen Hristov, Dina Katabi, and Christine Liu. Extracting gait velocity and stride length from surrounding radio signals. In *Proc. of the 2017 CHI Conf. on Human Factors in Computing Systems*, 2017.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clops: Continual learning of physiological signals. *arXiv preprint arXiv:2004.09578*, 2020.

SE Lapinsky, M Aubin, S Mehta, P Boiteau, and AS Slutsky. Safety and efficacy of a sustained inflation for alveolar recruitment in adults with respiratory failure. *Intensive care medicine*, 25(11):1297–1301, 1999.

Albert Lecube, Gabriel Sampol, Patricia Lloberes, Odile Romero, Jordi Mesa, Cristina Hernández, and Rafael Simó. Diabetes is an independent risk factor for severe nocturnal hypoxemia in obese patients. a case-control study. *PloS one*, 4(3):e4692, 2009.

Yifan Liu, Bohan Zhuang, Chunhua Shen, Hao Chen, and Wei Yin. Auxiliary learning for deep multi-task learning. *arXiv preprint arXiv:1909.02214*, 2019.

Joshua Mathew, Xin Tian, Min Wu, and Chau-Wai Wong. Remote blood oxygen estimation from videos using neural networks. *arXiv preprint arXiv:2107.05087*, 2021.

Taylor Mordan, Nicolas Thome, Gilles Henaff, and Matthieu Cord. Revisiting multi-task learning with rock: a deep residual auxiliary block for visual detection. In *NeurIPS*, 2018.

Mark Moss, Mark Franks, Pamela Briggs, David Kennedy, and Andrew Scholey. Compromised arterial oxygen saturation in elderly asthma sufferers results in selective cognitive impairment. *Journal of Clinical and Experimental Neuropsychology*, 27(2):139–150, 2005.

Murugappan Murugappan, Nagarajan Ramachandran, Yaacob Sazali, et al. Classification of human emotion from eeg using discrete wavelet transform. *Journal of biomedical science and engineering*, 3(04):390, 2010.

Preetum Nakkiran, Raziel Alvarez, Rohit Prabhavalkar, and Carolina Parada. Compressing deep neural networks using a rank-constrained topology. 2015.

Shashi Narayan, Nikos Papasarantopoulos, Shay B Cohen, and Mirella Lapata. Neural extractive summarization with side information. *arXiv preprint arXiv:1704.04530*, 2017.

NLM. Aging changes in the lungs. https://medlineplus.gov/ency/article/004011.htm, 2020. Accessed: 2020-10-01.

Mustafa Özdal, Zarife Pancar, Vedat Çinar, and Murat Bilgiç. Effect of smoking on oxygen saturation in healthy sedentary men and women. *EC Pulmonology and Respiratory Medicine*, 4(6):178–182, 2017.

David T Palma, George M Philips, Miguel R Arguedas, Susan M Harding, and Michael B Fallon. Oxygen desaturation during sleep in hepatopulmonary syndrome. *Hepatology*, 47 (4):1257–1263, 2008.

Tauhidur Rahman, Alexander T Adams, Ruth Vinisha Ravichandran, Mi Zhang, Shwetak N Patel, Julie A Kientz, and Tanzeem Choudhury. Dopplesleep: A contactless unobtrusive sleep sensing system using short-range doppler radar. In *UbiComp*, pages 39–50, 2015.

M Ramonatxo. Gender influence on the oxygen consumption of the respiratory muscles in young and older healthy individuals. *Int J Sports Med*, 24:559–564, 2003.

Tony Reybrouck and Robert Fagard. Gender differences in the oxygen transport system during maximal exercise in hypertensive subjects. *Chest*, 115(3):788–792, 1999.

A Ricart, T Pages, G Viscor, C Leal, and JL Ventura. Sex-linked differences in pulse oxymetry. *British journal of sports medicine*, 42(7):620–621, 2008.

Beanbonyka Rim, Nak-Jun Sung, Sedong Min, and Min Hong. Deep learning in physiological signal data: A survey. *Sensors*, 20(4):969, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015.

Dangdang Shao, Chenbin Liu, Francis Tsow, Yuting Yang, Zijian Du, Rafael Iriya, Hui Yu, and Nongjian Tao. Noncontact monitoring of blood oxygen saturation using camera and dual-wavelength imaging system. *IEEE Transactions on Biomedical Engineering*, 2015.

Xiaoyong Shen, Aaron Hertzmann, Jiaya Jia, Sylvain Paris, Brian Price, Eli Shechtman, and Ian Sachs. Automatic portrait segmentation for image stylization. In *Computer Graphics Forum*, volume 35, pages 93–102, 2016.

Michael W Sjoding, Robert P Dickson, Theodore J Iwashyna, Steven E Gay, and Thomas S Valley. Racial bias in pulse oximetry measurement. *New England Journal of Medicine*, 383(25):2477–2478, 2020.

Yonglong Tian, Guang-He Lee, Hao He, Chen-Yu Hsu, and Dina Katabi. Rf-based fall monitoring using convolutional neural networks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–24, 2018.

Mark Van Gastel, Sander Stuijk, and Gerard De Haan. New principle for measuring arterial blood oxygenation, enabling motion-robust remote monitoring. *Scientific reports*, 6(1): 1–16, 2016.

Mark van Gastel, Wim Verkruysse, and Gerard de Haan. Data-driven calibration estimation for robust remote pulse-oximetry. *Applied Sciences*, 9(18):3857, 2019.

Shichao Yue, Hao He, Hao Wang, Hariharan Rahul, and Dina Katabi. Extracting multi-person respiration from entangled rf signals. In *UbiComp*, 2018.

Guo-Qiang Zhang, Licong Cui, Remo Mueller, Shiqiang Tao, Matthew Kim, Michael Rueschman, Sara Mariani, Daniel Mobley, and Susan Redline. The national sleep research resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, 2018.

Mingmin Zhao, Fadel Adib, and Dina Katabi. Emotion recognition using wireless signals. In *MobiCom*, 2016.

Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. Learning sleep stages from radio signals: A conditional adversarial architecture. In *ICML*, 2017.

Appendix A. Theoretical Results

Definition 2 (Column rank deficient) *We call a matrix* $A = [a_1, \cdots, a_n] \in \mathbb{R}^{m\times n}$ *is column rank deficient if any column of it can be linearly represented by the rest of the columns, i.e.,* $\forall i \in [n], a_i \in span(\{a_j\}_{j\neq i})$.

Theorem 1 (Formal) *If the neural network* (f, W)*'s first layer weight matrix is column rank deficient, then its error is lower-bounded as follows:*

$$\mathcal{E}(f, W) \geq \mathcal{D}(g_0, g_1) \triangleq \min_{c\in R^m} \mathcal{D}_c(g_0, g_1) \tag{4}$$

$$\textit{where,}\ \ \mathcal{E}(f, W) \triangleq \mathbb{E}_{x\sim p(x)}\left[\sum_{s=0}^{1}(f(W[x, s]) - g_s(x))^2\right]$$

$$\mathcal{D}_c(g_0, g_1) \triangleq \mathbb{E}_{x\sim p(x)}\left[\frac{p(x + c)(g_0(x) - g_1(x + c))^2}{p(x) + p(x + c)}\right]$$

Proof Let us decompose the first layer weight matrix $W \in \mathbb{R}^{m\times(n+1)}$ as $W = [A, \alpha]$ where $A \in \mathbb{R}^{m\times n}$ and $\alpha \in \mathbb{R}^m$, m is the size of the first layer neurons, $n+1$ is the input dimension (recall that $x \in \mathbb{R}^n$ and s is binary variable). By our assumption of column rank deficient, we know that α is in the column space of A. Thus there exists a vector $b \in \mathbb{R}^n$ such that $\alpha = Ab$. With this bias vector b, we prove our theorem as follows.

$$\begin{aligned}
\mathcal{E}(f, W) &= \int_x \sum_{s=0}^{1}(f(Ax + \alpha s) - g_s(x))^2 p(x)dx \\
&= \int_x (f(Ax) - g_0(x))^2 p(x)dx + \int_x (f(Ax + \alpha) - g_1(x))^2 p(x)dx \\
&= \int_x (f(Ax) - g_0(x))^2 p(x)dx + \int_x (f(A(x + b)) - g_1(x))^2 p(x)dx \\
&= \int_x \big((f(Ax) - g_0(x))^2 p(x) + (f(Ax) - g_1(x - b))^2 p(x - b)\big)\, dx \\
&\geq \int_x \frac{p(x)p(x - b)}{p(x) + p(x - b)} \left(g_0(x) - g_1(x - b)\right)^2 dx = \mathcal{D}_{-b}(g_0, g_1) \geq \min_c \mathcal{D}_c(g_0, g_1)
\end{aligned}$$

Note the inequality of the integral in the last row holds because of the optimum of the following quadratic forms,

$$\begin{aligned}
a_0(t - b_0)^2 + a_1(t - b_1)^2 &= (a_0 + a_1)t^2 - 2(a_0b_0 + a_1b_1)t + (a_0b_0^2 + a_1b_1^2) \\
&= (a_0 + a_1)(t - \frac{a_0b_0 + a_1b_1}{a_0 + a_1})^2 + (a_0b_0^2 + a_1b_1^2) - \frac{(a_0b_0 + a_1b_1)^2}{a_0 + a_1} \\
&\geq \frac{a_0a_1(b_0^2 - 2b_0b_1 + b_1^2)}{a_0 + a_1} = \frac{a_0a_1}{a_0 + a_1}(b_0 - b_1)^2
\end{aligned}$$

■

Remark. The above theorem can be extended to apply when the side variable is inserted at higher layers of the neural network. Specifically, we can view the deep model as two parts: 1) An encoder E which encodes the input x into an intermediate representation $E(x)$; and 2) a predictor F which takes the encoded representation, $E(x)$, and the variable, s, as input and makes predictions $F(E(x), s)$. Hence, a similar error bound to that in Theorem 1 can be derived if the variable is inserted at a higher layer by substituting x by $E(x)$.

Empirical Test of the Assumption. Our theorem is based on the assumption that the weights of the neural network have a column rank deficiency. To validate this assumption, we conducted a test on the weights of the first layer of our Bert-UNet model.

Specifically, the weight matrix A in the first layer of our model has dimensions of 16×63. Through analysis, we have determined that A is approximately low-rank, with a rank of 12. This conclusion is supported by the fact that the top 12 principal components obtained through PCA analysis explain 95% of the total variance.

Furthermore, we have examined whether each column of A lies within the span of the remaining columns. To assess this, we computed the distance between the i'th column of A and the span of the remaining columns, and subsequently normalized this distance by the norm of the i'th column vector. Across all 16 columns, we found that the normalized distances exhibit a mean of 1.3×10^{-7} and a standard deviation of 2.8×10^{-8}.

Appendix B. Model Architecture and Implementation Details

B.1. Model Details

As we described in the main paper's method section, our backbone BERT-UNet model consists of an encoder and a predictor. The encoder has nine 1-D convolutional layers (Conv-Norm-RReLU) that shrink the features' temporal dimension by 240 times. It is then followed by several bi-directional multi-head self-attention layers (BERT) (Devlin et al., 2018) to aggregate the temporal information at the bottleneck. We use 8 layers, 6 heads with hidden-size of 256, intermediate-size of 512 for self-attention, and the max position embeddings is 2400. The decoder contains 7 layers of 1-D de-convolutional layers (DeConv-Norm-RReLU). We also use a skip connection (Ronneberger et al., 2015) by concatenating the convolutional layers in the encoder to the de-convolutional layers in the predictor. Figure 8 illustrates the overall network architecture. We keep the encoder and predictor architectures the same for all models. All models are implemented using PyTorch. The number of parameters for the Gated BERT-UNet model is 26,821,113 and the model size is 107.28MB. In the training process, we use the Adam optimizer with a learning rate of 2×10^{-4}, and train the model for 500 epochs. Due to the varying input length, we set the batch size for all models to 1 (i.e., one night of respiration signal and the corresponding oxygen time series). The breathing signal is measured at 10 fps (frames per second).

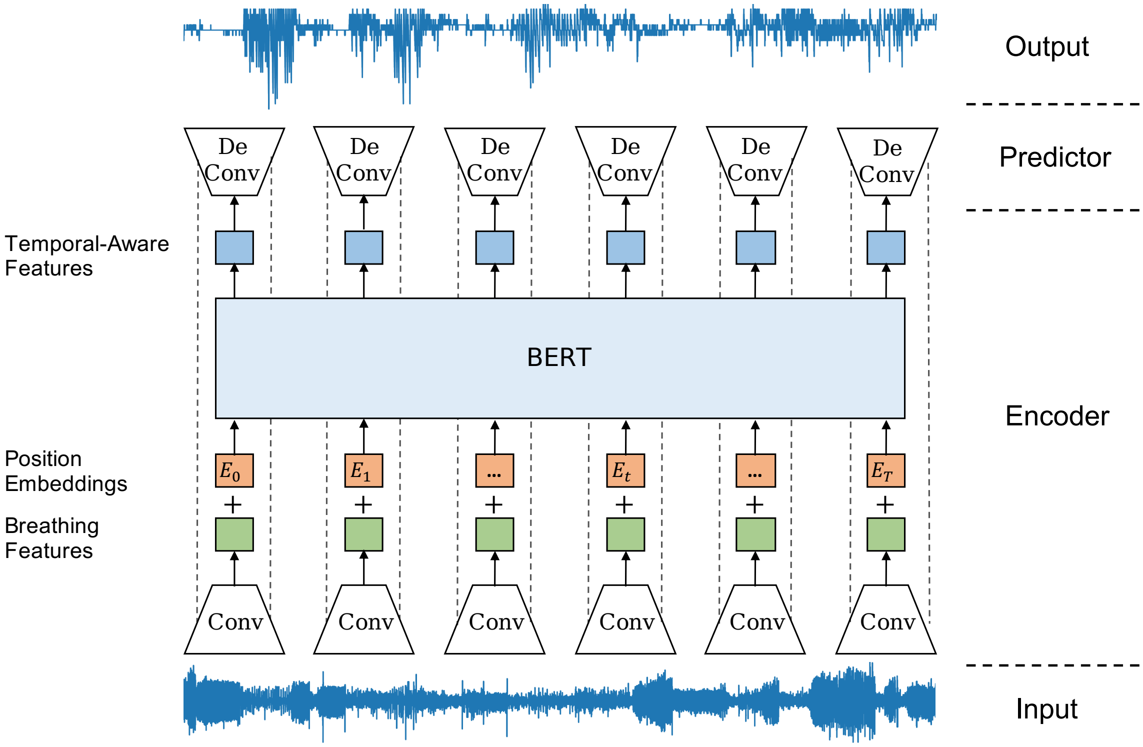
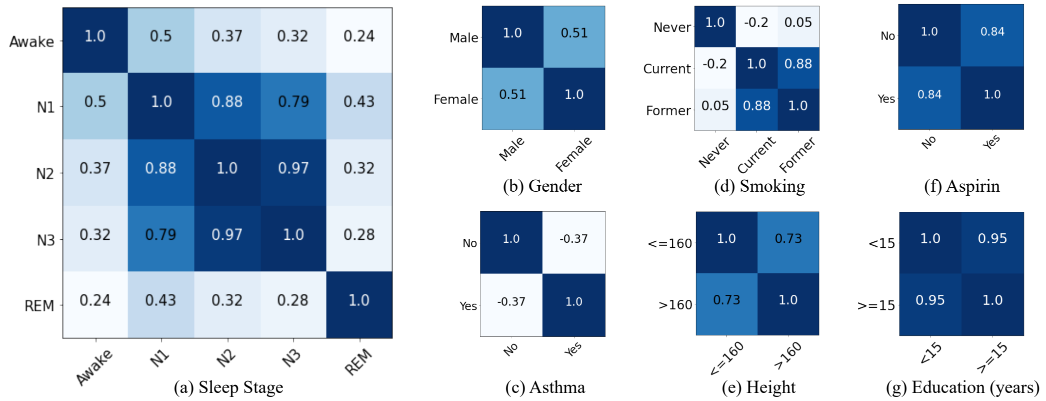


Figure 8: Network architecture for the backbone BERT-UNet model.

B.2. Evaluation Metric Details

As explained in the main paper, the correlation metric is sensitive to the length of segments. In setting of (Mathew et al., 2021), they compute correlation at a length of 240 seconds. Specifically, their data sample contains three cycles, and in each cycle the participant breathes normally for 30-40 seconds, then holds their breath for 30-40 seconds. Therefore, their test samples are around 180-240 seconds. Thus during our evaluation, we divide the data into non-overlapping 240-seconds segments and compute metrics on them.

B.3. Gradient Analysis for Gate Status Identification



(a) Sleep Stage (b) Gender (c) Asthma (d) Smoking (e) Height (f) Aspirin (g) Education (years)

Figure 9: Gradient cosine similarities of different variables.

In our analysis, we tested six variables: (a) Sleep stages: awake, light sleep (N1,N2), deep sleep (N3), rapid eye movement (REM); (b) Gender: male vs. female; (c) Asthma: having asthma or not; (d): Smoking: non-smoker, current smoker, former smoker; (e) Education: less than 15 years vs. greater than 15 years; (f): Height: lower than 160cm or higher than 160cm; (g) Aspirin: taking aspirin or not.

Ultimately, we selected two variables, *sleep stage* and *gender*, as auxiliary variables since they were available in all datasets. We established six status groups: (male, awake), (male, REM), (male, N1+N2+N3), (female, awake), (female, REM), (female, N1+N2+N3). This decision was based on the analysis of gradients shown in Figure 9. Figure 9(a) reveals significant dissimilarity between awake, REM, and NREM sleep stages (N1,N2,N3), but greater similarity within the group of N1,N2,N3. Figure 9(b) displays substantial differences in gradients across genders, consistent with existing medical knowledge that women consume more oxygen during breathing (Ramonatxo, 2003) and have lower oxygen saturation compared to men (Ricart et al., 2008), particularly during exercise (Reybrouck and Fagard, 1999).

Regarding the other variables studied, *Smoking* and *Asthma* were found to have an impact. It is understandable that asthma affects oxygenation since asthma causes difficulty in breathing. As for smoking, medical research indicating (Özdal et al., 2017) that smoking reduces oxygen saturation. As smoking's effects are long-term, regardless of quitting, the oxygen pattern is already altered, which may explain the high gradient similarity between current and former smokers.

Appendix C. More Experiment Results

We include more visualizations of the breathing signals and the corresponding oxygen saturation predicted by the *Gated BERT-UNet* model for the medical datasets: MESA (Figure 12 and Figure 13), MrOS (Figure 10 and Figure 11). We also visualize the model's prediction on unhealthy subjects with various diseases including chronic obstructive pulmonary disease (COPD), asthma, diabetes. In the plots, The background color indicates different sleep stages. The 'dark grey', 'light grey' and 'white' correspond to 'Non-REM sleep' (N1, N2, N3), 'REM' and 'Awake', respectively.

Results on MESA Dataset. The example in Figure 12 shows our model's ability to capture the fluctuations of oxygen saturation. At the same time, the example in Figure 13 shows that our model accurately detects the region of low oxygen saturation, which highlights its usefulness in monitoring patients.

Results on MrOS Dataset. From the zoomed-in regions (a) and (b) in Figure 10, we see that the model exhibits a larger error when the ground truth SpO2 reading is very low. It is mainly caused by the imbalanced labels in the training set since subjects usually experience much less time of having low oxygen level (e.g., below 90%) than having a normal oxygen level between 94% to 100%. Figure 11 shows another example of the dynamics. As shown in the zoomed range, oxygen fluctuations correlate with one's sleep stage: in 'REM' (colored by light gray), the oxygen fluctuates drastically while in 'Non-REM sleep' (colored by dark gray), the oxygen is much more stable and fluctuates in a small range. Our model makes accurate predictions since it leverages the sleep stage information.

Results for Relevant Diseases. As we mentioned earlier, using breathing as the input to the neural network model allows us to both train and test on large respiration dataset from past sleep studies. Since these datasets contain diverse people with a variety of diseases, it makes it possible to check how the model generalizes to unhealthy individuals. Particularly we are interested in diseases that interact with oxygen saturation including pulmonary diseases such as chronic obstructive pulmonary disease (COPD), chronic bronchitis, asthma, emphysema, and others like diabetes and coronary heart disease.

Diabetes is a disease in which the patient's blood sugar levels are too high. Research has shown that diabetes is a risk factor for severe nocturnal Hypoxemia in obese patients Lecube et al. (2009). Further diabetic patients tend to have 3% to 10% lower lung volumes than adults without the disease. Figure 14 shows an example of a diabetic patient who has an oxygen level that keeps oscillating between 85% and 95%. From the zoomed-in region, we can see our model captures the oscillating oxygen dynamics and accurately predicts the oxygen values.

Chronic obstructive pulmonary disease (COPD) refers to a chronic inflammatory lung disease that obstructs airflow from the lungs. Severe COPD can cause hypoxia, an extremely low oxygen level. Figure 15 shows an example of a COPD patient who experiences several oxygen droppings during the REM period (indicated by the orange box). As we can see, our model successfully predicts the events of oxygen dropping.

Chronic bronchitis refers to long-term inflammation of the bronchi. Chronic bronchitis patients can have shortness of breath which affects oxygen levels. Figure 16 is an example of a chronic bronchitis patient whose oxygen level keeps osculating between normal and low during the night. Our model captures the trend well.

Asthma is a condition in which a person's airways narrow, swell, and produce extra mucus. An asthma patient's oxygen levels can be irregular due to the breathing difficulty caused by the disease. As shown in the example in Figure 17, the patient has a normal oxygen level for most of the time, but the oxygen occasionally drops to low levels. Our model works well on detecting those abnormal oxygen levels from the person's respiration.

Emphysema refers to a lung condition in which the air sacs in the person's lung are damaged. Patients with emphysema usually have breathing issues that affect oxygen saturation. Figure 18 shows an example of an emphysema patient. The patient experiences a long period of low oxygen during sleep (as highlighted by the orange box). Our model successfully predicts such unusual oxygen dynamics.

Coronary heart disease develops when the arteries of the heart are too narrow to deliver enough oxygen-rich blood to the heart. A deficiency in providing oxygen-rich blood can lead one's oxygen saturation to deviate from normal level. Figure 19 is an example of a person with coronary heart disease who experiences several severe oxygen drops during sleep. Our model accurately tracks their oxygen level and detects the oxygen reduction events.

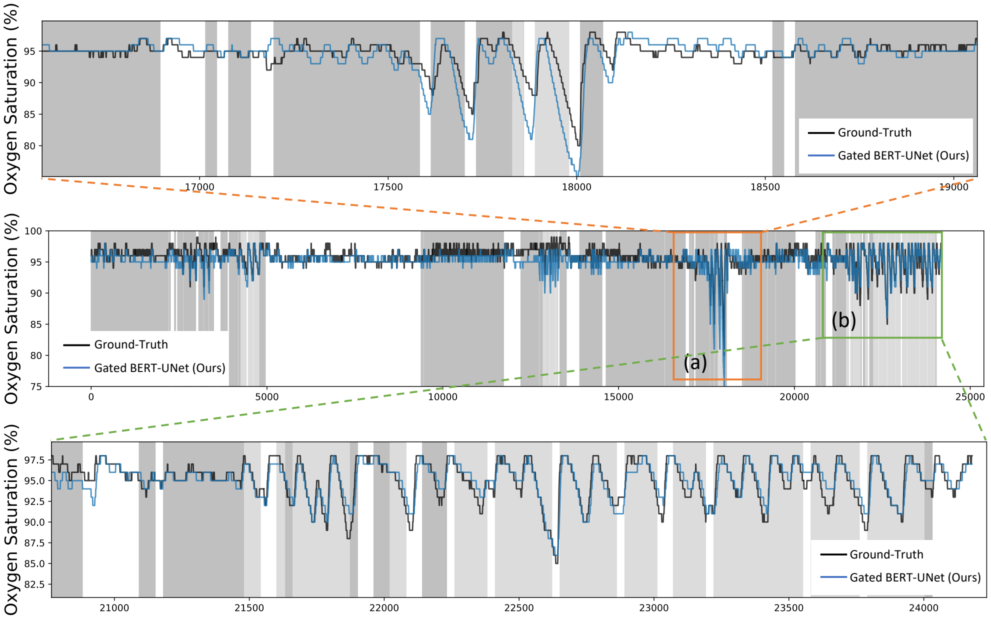


Figure 10: *MrOS* Example 1.

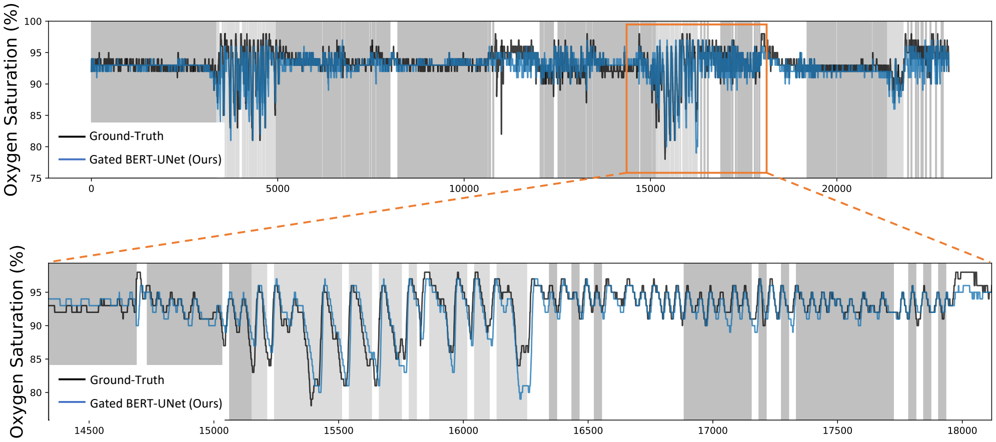


Figure 11: *MrOS* Example 2.

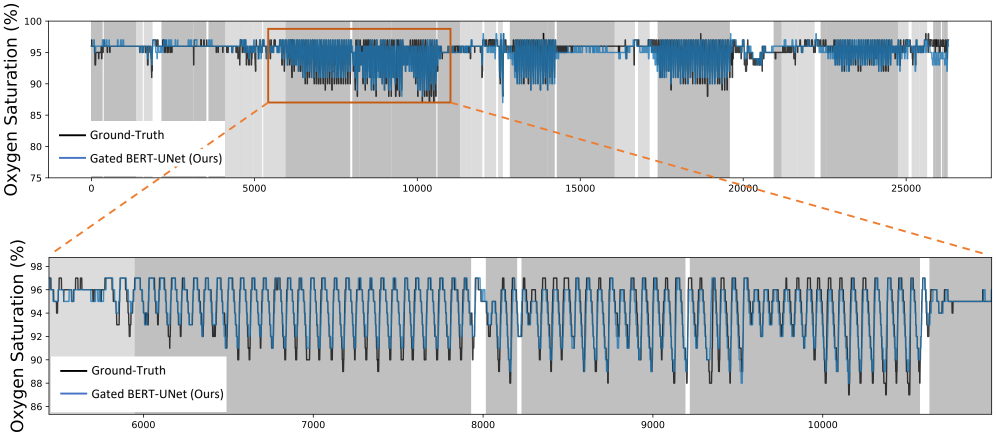


Figure 12: *MESA* Example 1.

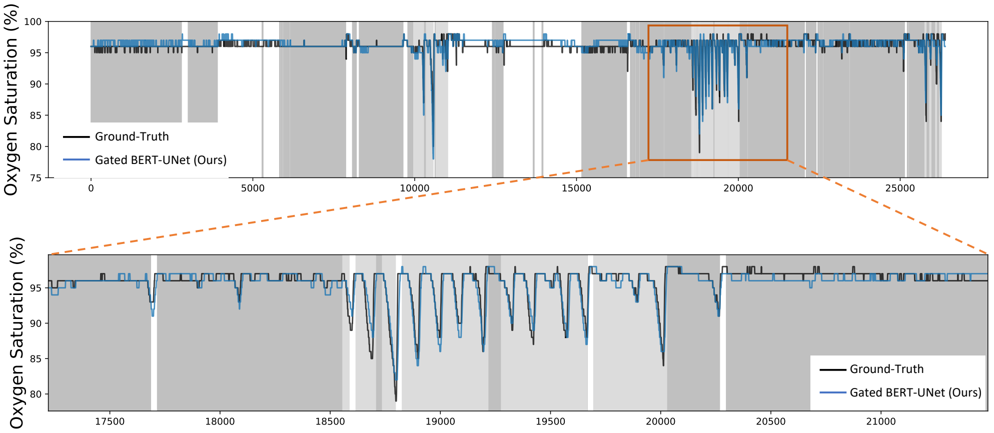


Figure 13: *MESA* Example 2.

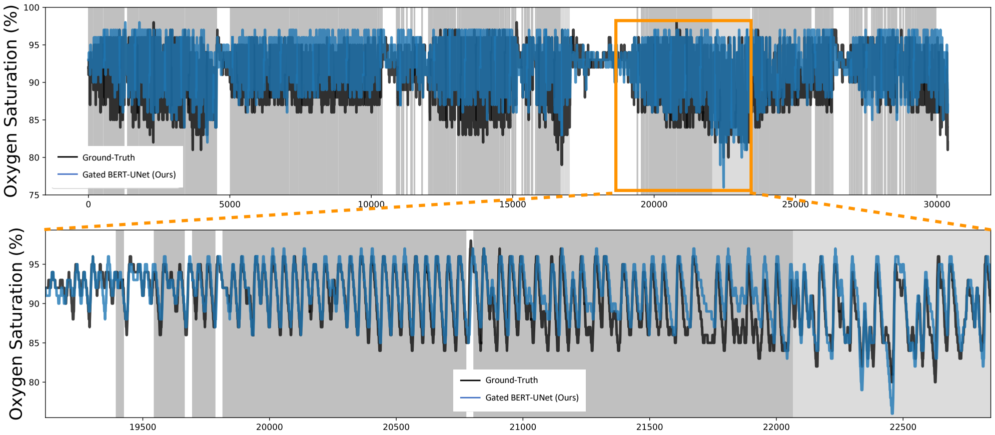


Figure 14: Diabetes Patient Example.

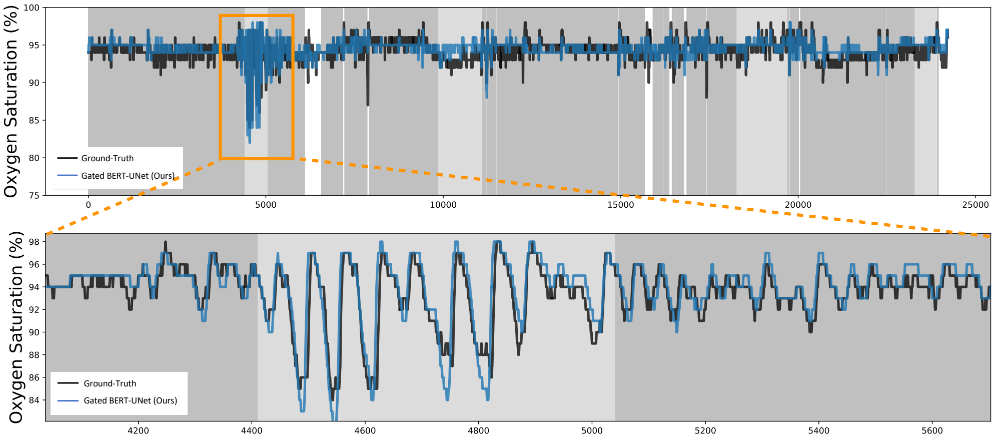


Figure 15: Chronic Obstructive Pulmonary Disease (COPD) Patient Example.

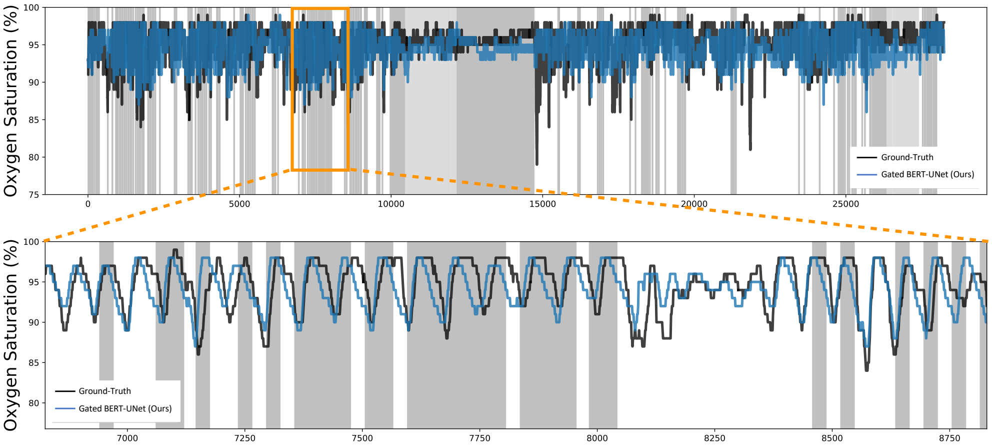


Figure 16: Chronic Bronchitis Patient Example.

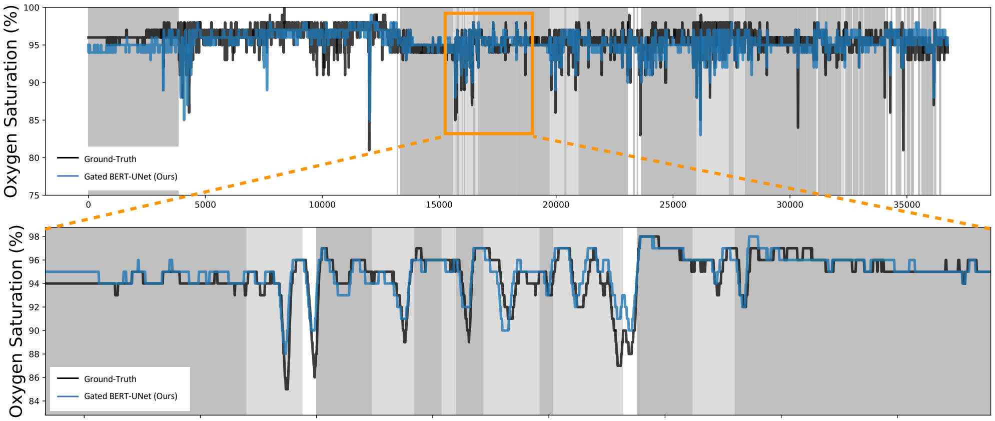


Figure 17: Asthma Patient Example.

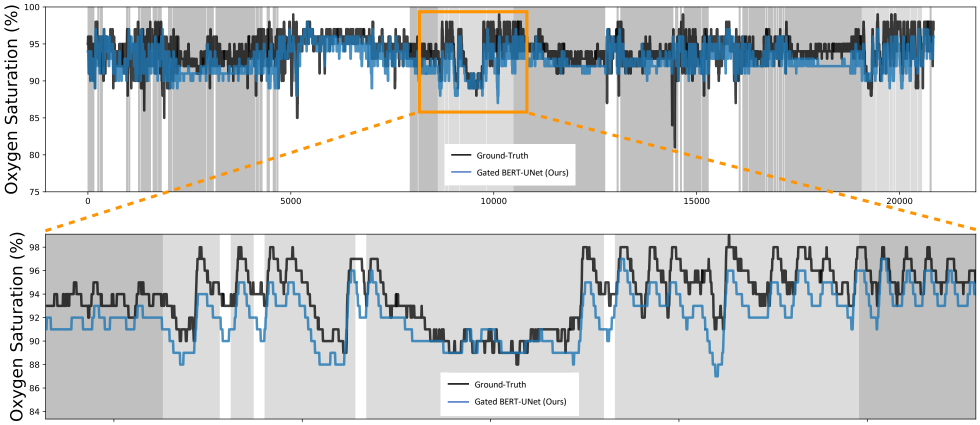


Figure 18: Emphysema Patient Example.

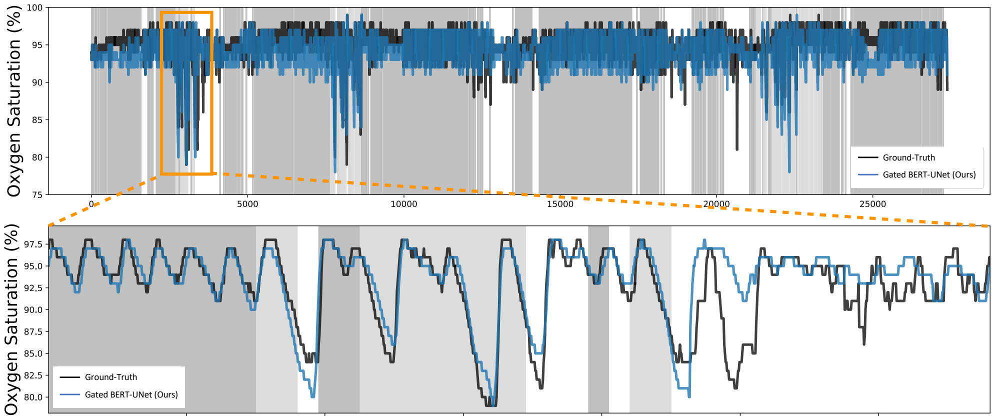


Figure 19: Coronary Heart Disease Patient Example.