Medical & Biological Engineering & Computing (2025) 63:1481–1495
https://doi.org/10.1007/s11517-025-03285-2

ORIGINAL ARTICLE

Interpretable machine learning models for COPD ease of breathing estimation

Thomas T. Kok[1] · John Morales[2] · Dirk Deschrijver[1] · Dolores Blanco-Almazán[3,4,5] · Willemijn Groenendaal[2] · David Ruttens[6] · Christophe Smeets[6] · Vojkan Mihajlović[2] · Femke Ongenae[1] · Sofie Van Hoecke[1]

Received: 13 June 2024 / Accepted: 31 December 2024 / Published online: 14 January 2025

Abstract
Chronic obstructive pulmonary disease (COPD) is a leading cause of death worldwide and greatly reduces the quality of life. Utilizing remote monitoring has been shown to improve quality of life and reduce exacerbations, but remains an ongoing area of research. We introduce a novel method for estimating changes in ease of breathing for COPD patients, using obstructed breathing data collected via wearables. Physiological signals were recorded, including respiratory airflow, acceleration, audio, and bio-impedance. By comparing patient-specific measurements, this approach enables non-intrusive remote monitoring. We analyze the influence of signal selection, window parameters, feature engineering, and classification models on predictive performance, finding that acceleration signals are most effective, complemented by audio signals. The best model achieves an F1-score of 0.83. To facilitate clinical adoption, we incorporate interpretability by designing novel saliency map methods, highlighting important aspects of the signals. We adapt local explainability techniques to time series and introduce a novel imputation method for periodic signals, improving faithfulness to the data and interpretability.

Keywords Chronic obstructive pulmonary disease (COPD) · Respiratory monitoring · Interpretability · Machine learning · Time series classification

1 Introduction

Chronic obstructive pulmonary disease (COPD) is an irreversible lung disease characterized by airflow obstruction, which makes breathing difficult. COPD is responsible for more than 5% of deaths worldwide and ranks as the third most common cause of death [2]. Patients with COPD also experience a significant reduction in quality of life [23]. Besides the effects on patients' health, COPD also carries an economic burden of nearly $50 billion annually in the United States alone [1].

Despite the widespread impact of COPD, current methods for disease monitoring remain limited or inconvenient for patients. Diagnosis and monitoring of COPD progression typically involve lung function tests, which measure pulmonary function, such as spirometry, long volume testing, and diffusing capacity tests [20]. These tests require medical professionals and the necessary equipment, making continuous monitoring challenging. Remote home monitoring has been shown to improve the quality of life for COPD patients [26] and represents a promising area of research [42]. Another promising indicator of disease progression is the ease of breathing, as COPD obstructs airflow and makes breathing more difficult [38].

This study introduces a novel method for predicting changes in ease of breathing. Our method uses data from a cohort of COPD patients whose breathing was artificially obstructed in a controlled study. The collected data consists of airflow and multiple wearable biomedical signals recorded under varying levels of inspiration obstruction. By utilizing features derived from signals recorded at different levels of

✉ Thomas T. Kok
thomas.kok@ugent.be

1 IDLab, Ghent University-Imec, Technologiepark-Zwijnaarde 126, Zwijnaarde, Belgium

2 Imec Netherlands, HTC 31, Eindhoven, Netherlands

3 Universitat Politècnica de Catalunya, Barcelona, Spain

4 Institute for Bioengineering of Catalonia, Barcelona, Spain

5 Biomedical Research Networking Center in Bioengineering, Biomaterials and Nanomedicine, Barcelona, Spain

6 Ziekenhuis Oost-Limburg, Genk, Belgium

obstruction, we can estimate potential changes in ease of breathing.

We also explore how the selection of wearable biomedical signals influences model performance. This provides useful insights into the most relevant signal modalities for estimating breathing ability, and their potential relevance for other use cases.

In medical and healthcare applications, it is crucial to incorporate local interpretability into machine learning pipelines [39]. Without providing explanations alongside predictions, the underlying reasoning of the model cannot be understood, investigated, and validated. This lack of transparency can reduce trust in the model, which is particularly important when making health-related decisions. Physicians may be hesitant to adopt such models without it, as they remain responsible [13, 43]. It is therefore essential to provide understandable explanations to support the decision-making process. This allows physicians to combine their expertise with the model's reasoning to make informed decisions.

Saliency maps, a technique originating in the image domain [40], are often used to explain time series-based classification models. This involves highlighting important areas of the signal associated with specific labels. We designed a new method for imputing occluded segments to improve the quality of saliency map-based explanations for time series, applied to respiratory signal data.

The remainder of this paper is structured as follows: Sect. 2 reviews the literature and related work. Section 3 outlines the data collection, preprocessing, and problem definition. Section 4 details the model and all parameters for evaluation. Section 5 presents the results of our experiments. Section 6 discusses interpretability and introduces our new imputation method, along with techniques to address window dependence. Section 7 discusses the findings and their implications. Finally, Sect. 8 concludes the paper and summarizes key contributions.

2 Related work

Although previous work explores similar problem settings, they do not directly focus on the same problem definition. There is a large body of research on the application of machine learning to respiratory problems, including COPD [18, 25]. Several studies focus on diagnosing whether a patient has COPD [16, 41]. Other studies aim to predict short-term exacerbations using respiratory sounds [19], meteorological data [9], or a variety of physiological measurements [4]. Another study predicts worrisome events based on heart rate and oxygen saturation [33]. These studies differ from this one, as our research integrates sound, acceleration, and bio-impedance measurements to estimate ease of breathing, rather than a smaller set of measurements to predict exacerbations or worrisome events. Initial work has been done on providing interpretable results, such as feature importance for predicting exacerbations [28]. Our approach extends upon previous work by applying local explanations directly to the underlying biomedical signals.

We published previous work based on the same data collection used in this paper. The first explores processing techniques for bio-impedance signals and evaluates the linearity of bio-impedance with volume [10], and proposes markers to evaluate breathing under inspiratory loads, combining bio-impedance with myographic signals [11]. The second defines a comparator model for estimating ease of breathing [27], which this work expands upon by including new data collection, extending feature sets, working with complementary signal modalities, and integrating explainability into the model.

Previous studies have explored generating saliency maps for time series-based models. These methods occlude and impute time series segments or signal timesteps with simple operations [22, 34] or dynamic masks and perturbations [15] to estimate their importance. We extend upon these studies for our problem setting.

3 Data

The study included 66 patients and was performed at Ziekenhuis Oost-Limburg (ZOL), Genk, Belgium. The study was conducted in two stages: the first in 2019 with 50 COPD patients, and the second in 2021 with 16 COPD patients. Both stages followed the same collection protocol and test procedures. Due to COVID-19 safety protocols, only airflow inhalations were measured during the second stage to minimize potential contaminations.

Inspiration was artificially modified to study respiratory changes in COPD patients under controlled conditions. A loading protocol was performed to induce changes in breathing, following previous work [30]. Inspiratory loads were applied in 12% increments from 0 to 60% of maximal inspiratory pressure, which was measured at the start of the test [3], to obtain a range of breathing obstruction levels. Data was collected for a minimum of thirty breaths per load, with at least two minutes of rest between loads.

In addition to airflow, several wearable signals were collected: three-dimensional acceleration using accelerometers located at the parasternal and diaphragm (lower intercostal spaces); audio signals with microphones positioned on the back at the left and right lung zones, and the trachea; and bio-impedance with a tetrapolar configuration on the midaxillary line and symmetrical to the midsternal line. All signals besides bio-impedance were sampled at 10,000Hz, while bio-impedance was sampled at 16Hz.

Table 1 The used preprocessing steps for each signal type

Signal type		Preprocessing steps
Respiratory flow		Decimate to 200Hz 4th order Butterworth low-pass filter between 0.1 and 40Hz
Bio-impedance		Cubic interpolate to 200Hz 4th order Butterworth filter between 0.05 and 2Hz
Audio (all)	Respiratory pattern	Decimate to 200Hz 4th order Butterworth filter between 0.05 and 2Hz
	High-frequency	-
Accelerometer (all)	Respiratory pattern	Decimate to 200Hz 4th order Butterworth filter between 0.1 and 5Hz
		Construct the three dimensions as a multivariate signal
	MMG	Decimate to 200Hz 8th order Butterworth filter between 5 and 40Hz Construct the three dimensions as a multivariate signal

Two systems were used to record the physiological signals. The first system was a standard wired acquisition system (MP150, Biopac Systems, Inc., Goleta, CA, USA), to record the accelerometer data and audio signals, and Biopac, with a pneumotach transducer (TSD107B, Biopac Systems, Inc.) connected to a differential amplifier (DA100C, Biopac Systems, Inc.) to measure the airflow. The second system was a low-power wearable device (imec, Eindhoven, the Netherlands) to record the bio-impedance signal. It used an injecting current of 100 uAp-p at 80 kHz with a sampling frequency of 16 Hz. The audio signals were recorded using three microphones (TSD108, Biopac Systems, Inc) with a frequency response of 35–3500 Hz.

Two patients had missing, saturated, or incorrectly measured signals and were excluded from analysis. The final dataset therefore consists of 64 patients.

3.1 Preprocessing

The signals in the dataset contain noise from various sources, such as patients moving and electrical signal interference. Different preprocessing steps are applied to each signal type to remove this noise, based on previous work [11, 27] and empirically adjusted for the data and problem setting. Table 1 provides an overview of the preprocessing steps for each signal type.

The audio and accelerometer signals are preprocessed with two separate methods to extract information from different frequency ranges. For both signals, preprocessing separates lower frequencies, from which a respiratory-like pattern is extracted, from higher frequencies. The high-frequency bands from the accelerometer signal can be used to extract the mechanomyogram (MMG) signal.

Figure 1 shows an example window of 30 s of each signal after preprocessing, highlighting respiratory patterns captured in the lower frequencies.

Measurements were paused by some patients who experienced breathing difficulties, and these moments were excluded from the dataset. As the pneumotach is removed during these breaks to facilitate easier breathing, this lack of airflow can be detected. When the flow signal is consistently flat during exhalation due to COVID-19 protocols, this should not be excluded. To address this, we use a rolling window to calculate the moving standard deviation of the signal with a window size of 6 s, ensuring at least one full breath is included. We flag parts of the signal as inactive if the standard deviation drops below a predefined, empirically determined threshold of 0.05. Whenever windows are sampled, this is from parts of the signal that are not flagged as inactive.

3.2 Target definition

The objective is to construct a model capable of detecting changes in ease of breathing for patients. To achieve this, at least two instances are required: one as a baseline ($\mathbf{x_1}$), and one or more to evaluate changes in ease of breathing ($\mathbf{x_2}$).

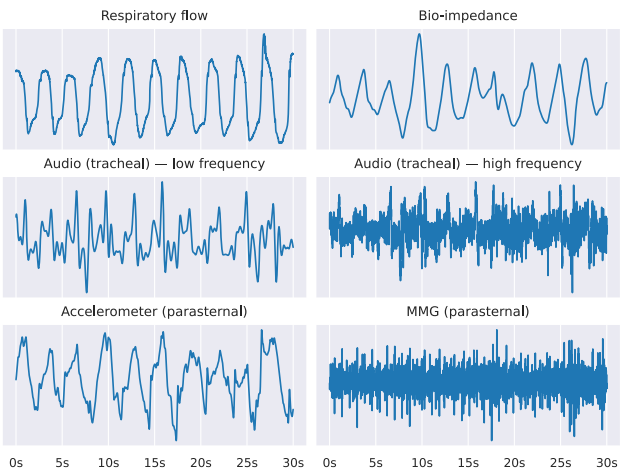


Fig. 1 The same 30 s window for all signals

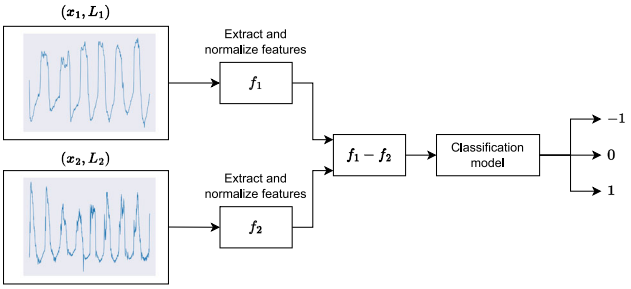


Fig. 2 The comparative, feature-based approach

We extract windows from the full signals to allow comparing two instances with the same load, to classify as no change in ease of breathing.

We define a tertiary classification, where $y \in \{-1, 0, 1\}$: -1 represents a decrease in inspiratory load (easier breathing), 0 represents no change in load, and 1 represents an increase in inspiratory load (more difficult breathing).

To generalize the problem, we define a threshold value τ to compare the windows $\mathbf{x_1}$ and $\mathbf{x_2}$. With this problem definition, the method can be used for longitudinal monitoring, where the latest status can be compared to the baseline with a predefined threshold τ.

The target is defined as follows:

$$y = \begin{cases} -1, & \text{if } L_1 - L_2 \geq \tau \\ 1, & \text{if } L_2 - L_1 \geq \tau \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where L_1 and L_2 are the two inspiratory loads applied for $\mathbf{x_1}$ and $\mathbf{x_2}$, and τ is set to 12% for our specific problem setting.

4 Methods

We define a feature-based comparative approach to address the problem using a machine learning model. For both windows—$\mathbf{x_1}$ and $\mathbf{x_2}$—we extract the same set of features, generating feature sets $\mathbf{f_1}$ and $\mathbf{f_2}$. These feature values are standardized (mean of 0, standard deviation of 1). $\mathbf{f_2}$ is subtracted from $\mathbf{f_1}$ to calculate the final set of features as input for the classification model that predicts the target, y. Figure 2 illustrates our approach.

We examined alternative methods for combining the two feature sets, such as concatenation, normalized feature ratios, and adding baseline features for additional context. However, none of these approaches improved performance.

4.1 Experimental setup

We experimentally compare all variables, such as the feature set and the combination of signal modalities used, to comprehensively investigate their impact on the model. This allows us to optimize the model and empirically explore the effects of these variables. The variables considered are listed in Table 2.

Every configuration is evaluated using the weighted F1-score. Each experimental setup and evaluation use the same seeds, to ensure consistent comparisons.

The dataset is split up into a training (80%) and a test set (20%). All explorative experiments and hyperparameter tuning are performed with 10-fold cross-validation on the training set, and the final experiments are evaluated on the test set. The dataset and cross-validation splits are balanced on the GOLD score [45], which represents disease severity. Each fold contains the same ratio of patients from each group: GOLD I (mild), GOLD II (moderate), and GOLD III-IV (severe).

4.2 Feature engineering

We compare different feature sets and feature selection setups. As defined in Table 3, the feature sets are: a basic feature set, an extended feature set of signal-specific features [6, 12, 32], and a maximal feature set of all available features from various time series, signal, and audio feature extraction packages. To extract the features, implementations were used from *tsfel* [6], *tsfresh* [14], *seglearn* [12], and *librosa* [32].

We use PowerSHAP for feature selection, as it is both effective and efficient compared to other state-of-the-art methods [44]. PowerSHAP assumes that informative features will have a larger impact on the prediction than random features. Artificial random features are introduced to the data, several models are trained on the updated data, and SHAP (SHapley Additive exPlanations) values [31] are calculated. SHAP is a widely used method for local interpretability, as it provides theoretical guarantees. SHAP values represent the

Table 2 An overview of all variables explored

Experimental step	Variables
Feature engineering	Extracted set of features use of feature selection
Signal modalities	Singular signals combinations of wearable signals
Window configuration	Window size window alignment number of windows
Classification	Classification model

Table 3 The basic and extended feature set. (BioZ: Bio-impedance, Acc.: Acceleration, Resp.: Low frequency)

Feature	Flow	BioZ	Audio		Acc.	
			Low	High	Resp	MMG
Basic feature set						
Mean, median	✓	✓	✓	✓	✓	✓
Standard deviation, variance	✓	✓	✓	✓	✓	✓
Root mean square	✓	✓	✓	✓	✓	✓
Minimum, maximum, absolute maximum	✓	✓	✓	✓	✓	✓
Skewness	✓	✓	✓	✓	✓	✓
Kurtosis	✓	✓	✓	✓	✓	✓
Extended feature set						
Time-domain						
Zero-crossing rate	✓	✓	✓	✓	✓	✓
Absolute energy	✓	✓	✓	✓	✓	✓
Autocorrelation	✓	✓	✓	✓	✓	✓
Binned, permutation entropy	✓	✓	✓	✓	✓	✓
Frequency-domain						
Fourier entropy	✓	✓	✓	✓	✓	✓
MFCC (mel-frequency cepstral coefficients)				✓		✓
Spectral bandwidth, centroid, contrast, flatness, roll-off				✓		✓

importance of each feature for each individual prediction. PowerSHAP compares the distribution of these values for each feature to those of random features, to find the most relevant ones.

For the extended and maximal feature sets, we compare the results with and without feature selection. Feature selection is not applied to the basic feature set due to its small size. We expect that this strong feature selection method allows effective use of the maximal feature set. Although this feature set would ordinarily be too large, the feature set can be reduced to a smaller set of relevant features with effective feature selection.

4.3 Signal modalities

The collected data consists of several physiological signals from different modalities. Determining which signal modalities are most relevant, and which are complementary to which other modalities, is relevant both for optimizing the model and for exploring different respiratory signals.

To compare the signal modalities fairly, the experimental setup is kept consistent for all signals. We evaluate all combinations of wearable signals to identify which signals are complementary and to optimize the model. The airflow signal is excluded from all combinations to focus on wearable signal modalities. Any combination of two to five signals from the following signals is included: acceleration (parasternal, high and low frequency); acceleration (diaphragm, high and low frequency); audio (tracheal, high and low frequency); audio (left lung, high and low frequency); audio (right lung, high and low frequency); bio-impedance.

4.4 Window configuration

In previous experimental steps, windows are 30 s long, as in previous work [27]. However, the window size could influence the performance of the model, and tuning it might give more insights into its impact on the model performance. Since respiratory rates generally range between 12 and 15 breaths per minute, a window shorter than 3 s would not contain enough information. At the same time, a window of several minutes would exclude the majority of the dataset as most signals might not contain two long enough uninterrupted periods of breathing. The window size was empirically chosen to maximize model performance while minimizing discarded data points.

Additionally, the size of the window can be defined by the number of breaths instead of seconds, by using the periodicity of the data. This generates windows that are more aligned with the respiratory patterns.

These breaths are defined by detecting peaks in the respiratory flow using the `scipy` implementation, which finds all local maxima and minima with a topographic prominence of at least 1. A window of one breath is defined from one peak to the subsequent peak.

The effect of the window size was evaluated for both options. For the first one, the window size was incrementally increased in steps of 5 s. For the second one, the window size was incrementally increased in steps of one respiratory

Table 4 The mean F1-score and standard deviation for each combination of signal and feature set, from all 10 folds

Signal	Basic	Extended		Maximal	
		No selection	+ Selection	No selection	+ Selection
Flow	0.627 ±.05	0.718 ±.03	0.724 ±.04	0.733 ±.04	0.766 ±.04
Acc. (parasternal)	0.653 ±.05	0.639 ±.04	0.688 ±.04	0.682 ±.03	0.775 ±.04
Acc. (diaphragm)	0.603 ±.06	0.595 ±.06	0.651 ±.07	0.647 ±.05	0.760 ±.04
Audio (tracheal)	0.611 ±.05	0.607 ±.06	0.636 ±.06	0.613 ±.06	0.701 ±.06
Audio (left lung)	0.535 ±.06	0.544 ±.07	0.579 ±.06	0.583 ±.05	0.646 ±.04
Audio (right lung)	0.529 ±.07	0.577 ±.06	0.567 ±.07	0.586 ±.06	0.633 ±.05
Bio-impedance	0.459 ±.09	0.471 ±.06	0.431 ±.07	0.504 ±.06	0.565 ±.04

cycle. Both approaches were evaluated on performance and data availability.

It is also possible to increase the number of windows extracted for each patient. In the previous experimental steps, a single comparison between two randomly placed non-overlapping windows is generated for each pair of loads for a patient. Instead of generating two windows for a single comparison, the non-overlapping requirement can be dropped to generate n windows. All n^2 combinations can then be used to train the model.

We evaluate the impact of increasing the number of windows by generating up to ten comparisons for each target. The number of windows in the test set remains unchanged at a single comparison, for fair evaluation.

4.5 Classification

After extracting all the features from the windowed instances, their difference $f_1 - f_2$ is used as input for the classification model. For the experiments of Table 2, a random forest model is used, as it is known to provide good results in medical studies [17], and has proven effective in previous work [27]. However, other models could potentially improve results. Therefore, we evaluated the following classification models alongside the random forest classifier: CatBoost (gradient boosting tree); a support vector machine (SVM) with radial basis function (RBF) kernel; and a logistic regression classifier. These models are frequently used in related research and represent a diverse range of models. Deep learning models have also been shown to be effective for the classification of respiratory signals [5, 37]. However, they rely on larger datasets than what is available and are thus excluded from this study.

5 Results

5.1 Feature engineering

Table 4 shows the results of the experiments comparing different feature sets for each available signal, using 30 s non-overlapping windows with a random forest classification model. Introducing a larger feature set and applying strong feature selection are both beneficial to the performance of the model. The maximal feature set with strong feature selection achieves the highest performance for each signal.

Comparing F1-scores for the maximal feature set with and without feature selection shows improvements for all signals. Paired t-tests ($\alpha = 0.05$) show these are significant with $p = 0.01$ (flow, tracheal audio, bio-impedance) and $p < 0.001$ (accelerometer, left lung audio, right lung audio). The maximal feature set also shows significant improvements over the extended feature set with $p = 0.03$ (flow), $p = 0.01$ (left lung audio), and $p < 0.001$ (accelerometer, right lung audio, bio-impedance). The only exception is tracheal audio, with $p = 0.15$.

Table 5 The mean scores for several performance measures for each signal, using the maximal feature set with strong feature selection

Signal	B. Acc	MCC	Prec.$_{y=1}$	Rec.$_{y=1}$	Spec.$_{y=1}$
Flow	0.75	0.63	0.82	0.80	0.90
Acc. (parasternal)	0.78	0.68	0.77	0.82	0.87
Acc. (diapgragm)	0.77	0.66	0.76	0.79	0.86
Audio (tracheal)	0.70	0.56	0.72	0.72	0.85
Audio (left lung)	0.64	0.47	0.68	0.68	0.83
Audio (right lung)	0.63	0.45	0.67	0.65	0.83
Bio-impedance	0.57	0.36	0.62	0.60	0.80

B. Acc. (Balanced Accuracy) and MCC (Matthew's Coefficient Correlation) are calculated using the weighted average. Prec. (Precision), Rec. (Recall / Sensitivity / TPR), and Spec. (Specificity / TNR) are all calculated for an increase in applied load, or an increased difficulty in breathing (where $y = 1$)

5.2 Signals

Tables 4 and 5 summarize the results for individual signals. The parasternal accelerometer signal produces the best overall results, followed closely by the diaphragm accelerometer. Among the three audio-based signals, tracheal audio stands out with the highest performance, while bio-impedance performs the worst.

Although the parasternal accelerometer signal provides better overall performance, the respiratory flow shows better results when considering only a potential increase in load, or higher difficulty of breathing.

A McNemar's test ($\alpha = 0.05$) confirms the accelerometer signal is more effective than the respiratory flow for predicting all three classes ($p = 0.02$). However, the respiratory flow signal is more effective when predicting a potential increase in difficulty of breathing, though the difference is not significant ($p = 0.37$).

Table 6 shows the best results for combinations of each possible number of signals, excluding respiratory airflow to focus on less obtrusive wearable signals. Maximal features with feature selection are used.

The best-performing combinations of signals consistently include the accelerometer signals. These signals achieve the best performance when used individually, with additional modalities resulting in further improvements. The audio-based signals, combined with the accelerometer, result in the best-performing models, outperforming only accelerometer signals.

Paired t-tests ($\alpha = 0.05$) show that the combination of two accelerometer signals significantly outperforms the combination of the parasternal accelerometer and tracheal audio ($p = 0.04$) and all other combinations, except the parasternal accelerometer and right lung audio ($p = 0.14$).

The best combinations of three signals all include both accelerometer signals with an audio signal. Comparisons between tracheal and right lung audio ($p = 0.88$) or tracheal and left lung audio ($p = 0.83$) show no significant difference between the different audio signals. The inclusion of the audio signal shows a small improvement over the two accelerometer signals, but this improvement is not significant ($p = 0.42$).

5.3 Window configurations

Figure 3 shows the impact of window size on the performance of the model and the data availability. Two models are evaluated: the first based on the airflow signal; the second based on the optimal combination of wearable signals (parasternal acceleration, diaphragm acceleration, and tracheal audio).

Table 6 The most effective combinations of signals (over 10 folds), for each potential number of signals

Number of signals			
2		3	
Signals	F1	Signals	F1
Acc. (parasternal) Acc. (diaphragm)	0.813 ± 0.03	Acc. (parasternal) Acc. (diaphragm) Audio (tracheal)	0.821 ± 0.04
Acc. (parasternal) Audio (right lung)	0.792 ± 0.06	Acc. (parasternal) Acc. (diaphragm) Audio (right lung)	0.819 ± 0.05
Acc. (parasternal) Audio (tracheal)	0.789 ± 0.04	Acc. (parasternal) Acc. (diaphragm) Audio (left lung)	0.818 ± 0.05
4		5	
Acc. (parasternal) Acc. (diaphragm) Audio (tracheal) Audio (right lung)	0.819 ± 0.04	Acc. (parasternal) Acc. (diaphragm) Audio (tracheal) Audio (right lung) Bio-impedance	0.820 ± 0.04
Acc. (parasternal) Acc. (diaphragm) Audio (right lung) Bio-impedance	0.819 ± 0.04	Acc. (parasternal) Acc. (diaphragm) Audio (tracheal) Audio (left lung) Bio-impedance	0.814 ± 0.04
Acc. (parasternal) Acc. (diaphragm) Audio (tracheal) Bio-impedance	0.813 ± 0.03	Acc. (parasternal) Acc. (diaphragm) Audio (tracheal) Audio (right lung) Audio (left lung)	0.812 ± 0.04

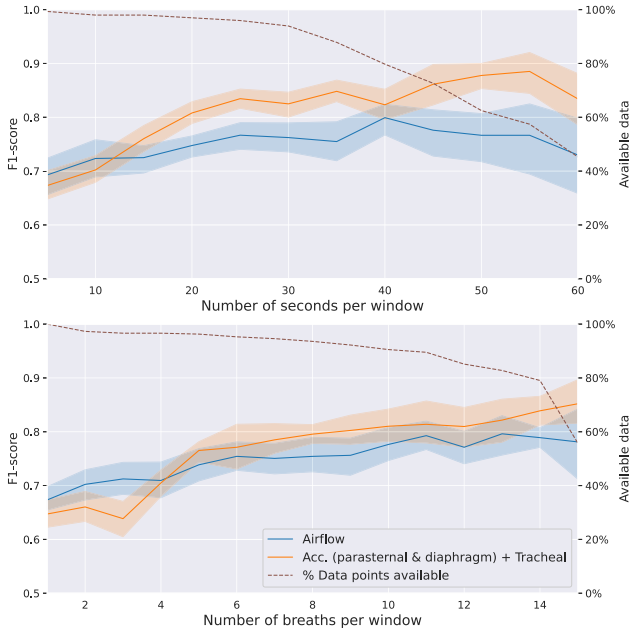


Fig. 3 The F1-score and percentage of available data points, for differing window sizes

Two different methods of defining window sizes are compared: numbers of seconds, and number of breaths.

For windows defined by number of seconds, performance increases with window size up to 40 s for the airflow signal and 55 s for the multivariate input. The percentage of available data points decreases from 30 s onward. When defining windows by breathing cycles, performance improves with longer windows as well, with no visible decrease. This may be due to this method more consistently capturing the information in the signals, or the optimal number of breaths potentially being out of scope.

Figure 4 shows the results of increased dataset size using overlapped windows. The overlap between windows decreases performance for a single comparison. However, increasing the number of comparisons to three or more improves performance, and surpasses the baseline of a single non-overlapping comparison. Beyond this, the performance plateaus around 0.80, as the overlap between windows increases when more is extracted.

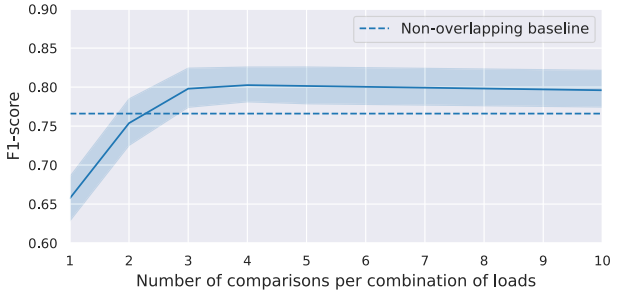


Fig. 4 The F1-score by number of comparisons in training

5.4 Classification model

Table 7 shows the results for each classification model, trained on the configuration as found in the previous experiments (the two acceleration signals and the tracheal audio signal, maximal feature set with feature selection, and three overlapping 30 s windows). The random forest model achieves the second-best performance and is surpassed only by the CatBoost gradient-boosted tree model. This improvement in performance with boosted trees is often observed in many problem domains [7].

The SVM classifier and logistic regression model both underperform compared to the two tree-based models. This suggests the data contains complex information not captured by the RBF input space or the more linear input space of the logistic regression model.

6 Interpretability

In medical contexts, it is important that interpretable models are self-explanatory and do not require further clarification. Where some features like mean, amplitude, and frequency might still be intuitive to physicians, other features like skewness and kurtosis might not. Saliency maps [40], which highlight the relevant areas of the original signal for the model's prediction, are therefore a promising approach to interpretability. Because a visualization of the original signal is intuitive, this approach allows both clinicians and machine learning experts to interpret the model outcomes.

For the remainder of this section, all figures and experiments use the respiratory flow signal. Using a single signal allows a clearer examination of the methods and saliency maps, but these principles can be applied to any signal, or combination of signals by concatenating these signals.

6.1 Background on time-based saliency maps

Saliency maps were originally developed for image processing, highlighting parts of the image most important for classification or object detection. Various studies have adapted saliency maps for time series data [8, 15, 21, 22, 34], typically using SHAP [31] to calculate time-based importance values.

Table 7 The F1-score for each model on the test set

Classification model	F1-score
Random forest	0.803
Boosted tree (CatBoost)	0.832
Support vector machine (RBF)	0.786
Logistic regression	0.724

Algorithm 1 Saliency map generation

GIVEN model M, input instances $\mathbf{x_1}, \mathbf{x_2}$, imputation method I, number of segments n
$L \leftarrow \text{LENGTH}(\mathbf{x_1})/n$
$S \leftarrow []$
for all $\mathbf{x}' \in \{\mathbf{x_1}, \mathbf{x_2}\}$ **do**
  for all $k \in 1..n$ **do**
    $k_0 \leftarrow k \cdot L$
    $k_1 \leftarrow (k+1) \cdot L$
    $S \leftarrow S + \mathbf{x}'[k_0 : k_1]$
  end for
end for
$V \leftarrow \text{SHAP}(M, S, I)$
return V

Approaches to generate saliency maps for time series generally divide the time series into n segments, treating each segment as a separate feature. These features are then used as input for SHAP calculations, where different combinations of segments are occluded by removing and imputing these segments. The imputation is done with a constant value, linear interpolation, noise, or sampling from a predefined background dataset, usually a subsample from the training set. To calculate the importance values, many partially occluded samples are drawn, for which the classification probabilities are calculated. These are then used as input for the Kernel SHAP method, as formalized in Algorithm 1.

Figure 5 shows an example saliency map generated using this approach for our problem setting. From the saliency maps, it can be inferred that the classification is primarily based on the stable, deeper breathing present in L_1, as both the amplitudes and associated importance values contrast with those present in L_2. The same reasoning is applicable to the slope of inhalation, where the steepest inhalations in L_1 are marked as important, contrasting with the more gradual slopes present in L_2.

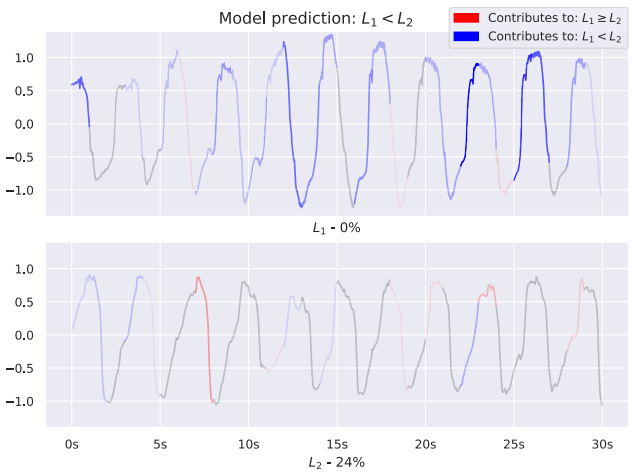


Fig. 5 An example of a generated time-based saliency map

Although this approach is effective, it has two key issues that we aim to address. The first is that the imputation methods often fail to accurately represent the original distribution, due to out-of-distribution sampling or bias within the distribution. Previous work has shown that out-of-distribution sampling negatively affects explanation quality [24]. For instance, the background sampling method replaces occluded segments with random samples from the background dataset, which often do not match the original respiratory pattern and thus generate an out-of-distribution instance. Figure 6 shows an example with different imputation methods all generating out-of-distribution instances.

The second issue is the dependence of the results on the placement and size of the segments, as local structures can vary in scale and potentially overlap at the edges of two sequential segments.

6.2 Improving the imputation method

To address the limitations of existing imputation methods, we introduce *periodic background sampling*. This method updates background sampling [31], where out-of-distribution sampling is avoided, while sampling from a distribution representing the entire dataset.

Instead of selecting random samples at the same x_{start} and x_{end}, which are randomly situated points in a periodic respiratory signal, we use domain knowledge to select more appropriate points. We select x_{start} and x_{end} to be aligned to the same point within the respiratory pattern based on the relative distance between four markers: start of exhalation,

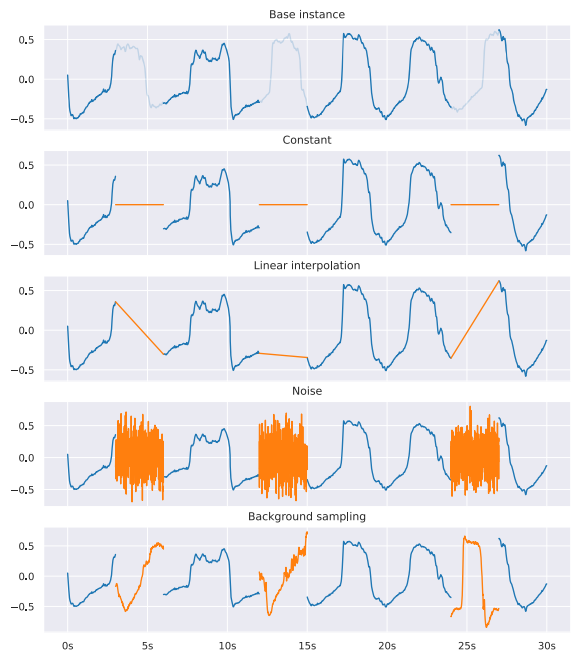


Fig. 6 Imputation methods applied to an example instance

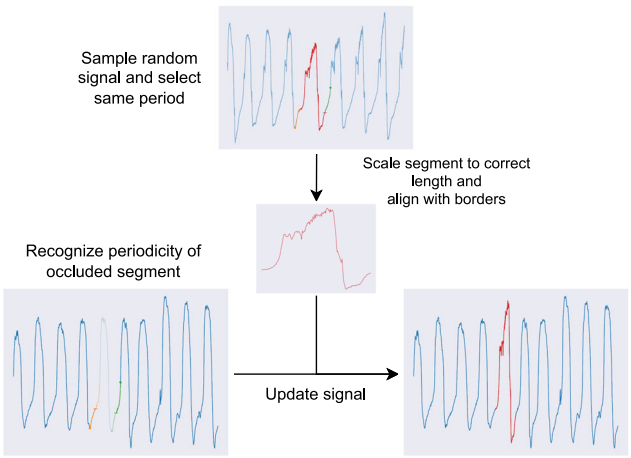


Fig. 7 Example periodic background sampling imputation

peak of exhalation, start of inhalation, and peak of inhalation. For example, if the occluded segment begins shortly after the start of an inhalation, so will the sampled background window. Besides this alignment, the sampling remains completely random. The background sample is then scaled to match the gap in the original time series. Figure 7 shows a schematic overview of the method, with an example imputation.

B.1. Distribution representativeness

The updated instances, after imputation of occluded segments, should optimally follow two properties:

1. They should be in-distribution relative to the original data, as sampling out-of-distribution instances has been shown to negatively impact the quality of explanations [24].
2. They should be representative for the full distribution, without bias towards a subset of the original data. This ensures the feature attribution method fully explores the effects of occluding a segment. If this is not the case, the information is biased and does not allow correct feature attribution calculations [31].

To evaluate whether imputed instances are representative of the original distribution, we use of principal component analysis (PCA). PCA converts the features extracted from the instances to a two-dimensional representation. This allows us to compare the original distribution of the dataset with the distribution of imputed instances. PCA is fitted on the feature space of the training dataset, and projected onto its first two principal components.

Figure 8 shows the distribution of the original dataset, compared to the distribution after imputing 5 out of 10 segments for each instance. The figure is zoomed in for clarity, but over 99% of the data remains visible. Periodic background sampling is the only imputation method that preserves the original distribution; the other three methods show a bias towards a specific direction, or towards a subset of the original distribution.

To evaluate whether the imputations faithfully represent the original distribution, we consider individual data points.

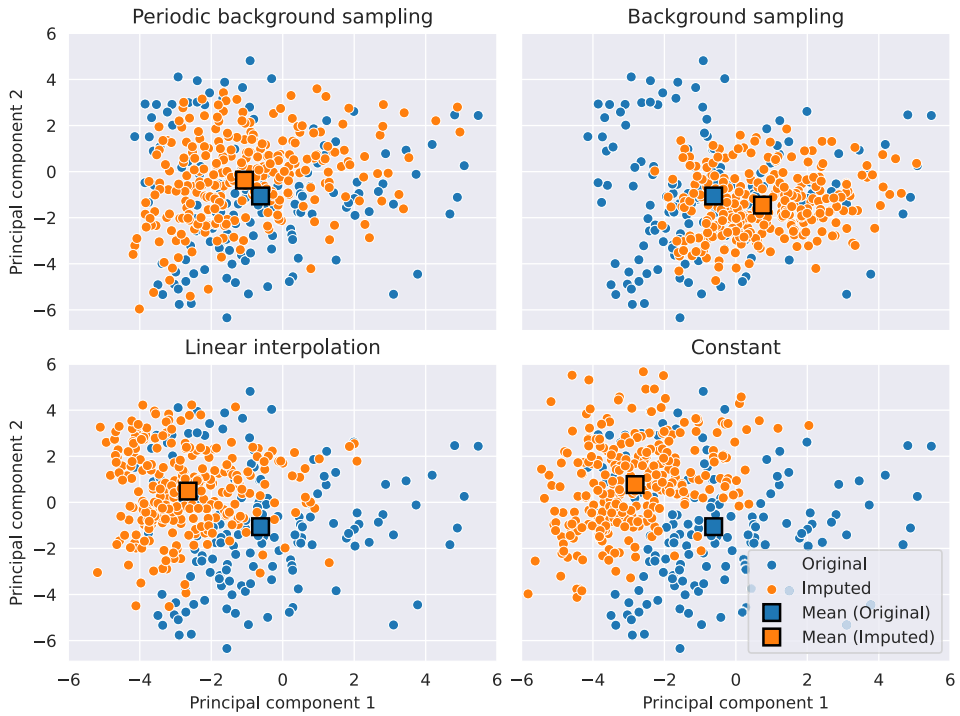


Fig. 8 The 2-dimensional PCA transformation of the original dataset, compared with generated instances

Table 8 The adjusted Wasserstein distance between directions towards the original distribution, and the imputations

Imputation method	Wasserstein distance
Periodic background sampling	0.804 ± 0.58
Background sampling	0.993 ± 0.53
Linear interpolation	0.889 ± 0.48
Constant	0.889 ± 0.66

We expect instances on the edge of the distribution to move inwards, and instances at the center to move outwards.

We represent the directions from original instances towards their imputed versions as angles, as well as the directions from original instances to all other original instances. Ideally, these two distributions should be as close as possible, indicating a faithful representation of the original data.

For comparison, we employ the Wasserstein distance [36]:

$$W(U, V) = \inf_{\gamma \in \prod(U,V)} \mathbb{E}_{(u,v)\sim\gamma} |u - v|$$

This metric is ideal as it accounts for differences in distribution and distances between directions. Comparing directions requires special consideration, as the angle space is circular (where 0 and 2π are equivalent). To account for this, we modify the Wasserstein distance calculation. Instead of directly calculating the distance between the two sets of directions, we first calculate the distance between all directions in both sets. These are then compared and averaged to all distances between the directions in each individual set, using the Wasserstein distance.

We run the following method with $k = 1$ and $m = 100$, for n instances.

Given instance $\mathbf{x}$, the dataset $\mathbf{X}$ and imputation method i:

1. For $\mathbf{x}' \in \mathbf{X}$: Calculate the angle between $\mathbf{x}$ and $\mathbf{x}'$
2. Generate $\mathbf{M}$ consisting of m instances with k segments occluded and imputed using i
3. For $\mathbf{m}' \in \mathbf{M}$: Calculate the angle between $\mathbf{x}$ and $\mathbf{m}'$
4. Calculate the distance between the two sets of angles, using the Wasserstein distance

The results are shown in Table 8. The periodic background sampling method achieves a lower average adjusted Wasserstein distance than the other three methods, whereas the regular background sampling method has the highest average distance.

B.2. Saliency maps and importance values

When calculating SHAP values using the mentioned imputation methods, it is possible for imputations to be too similar to the original segment. For example, linearly interpolating a segment that already was close to linear, causing the occlusion of the segment to have no impact on the prediction. These segments can effectively not be evaluated, causing the explanation to miss potentially relevant information.

Table 9 shows the percentage of importance values that are exactly zero for each imputation method, where k is the number of segments. The ratio of zero-values is roughly similar for all methods when using larger segments ($k = 5$ or $k = 10$), but diverges with smaller segments ($k = 20$ or $k = 30$). In particular, linear interpolation and constant imputation return the highest ratio of zero-values with smaller segments. This is likely due to these imputation methods being simple and deterministic, unlike the two background sampling methods. Smaller segments in periodic time series will more closely approximate either their linear interpolation or a constant value.

6.3 Removing the window size dependence

Our current approach to calculate saliency maps splits up the instance into k segments, and calculates the importance of each segment. However, as shown in previous work [34], this creates a dependency on the size and placement of the segments. Smaller segments may fail to capture relevant larger structures, whereas larger segments may overlook important details. Furthermore, the arbitrary placement of segment boundaries may split up important details, making it impossible to understand their relevance.

To address these limitations, we propose a solution that incorporates multiple segmentations of the time series to calculate importance values and averages them timestep-wise over different segmentations. With this approach, we mitigate the problem of segments that are either too narrow or too broad, enabling the final saliency map to show details at various sizes.

Table 9 The ratio of SHAP values equal to 0 for each imputation method, where k is the number of segments used

Imputation method	Zero-ratio			
	k = 5	k = 10	k = 20	k = 30
Periodic background sampling	0.297	0.303	0.414	0.312
Background sampling	0.350	0.317	0.458	0.300
Linear interpolation	0.297	0.268	0.577	0.547
Constant	0.321	0.325	0.513	0.441

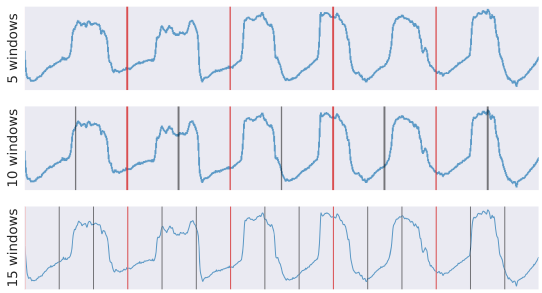
Careful consideration must still be given to the choice of the segmentations. Intuitively, one might select values such as $k = 5$ or 10, but that would allow suboptimal segments. As illustrated in Fig. 9a, using multiples of k results in shared boundaries, missing important details around these boundaries. Instead, we use only prime numbers (e.g., 7, 11, 17), ensuring unique segment placements for each k as illustrated in Fig. 9b. This approach eliminates issues caused by suboptimal window placements.

Examples comparing saliency maps generated with different methods can be found on our project website at: https://predict.idlab.ugent.be/copd/.

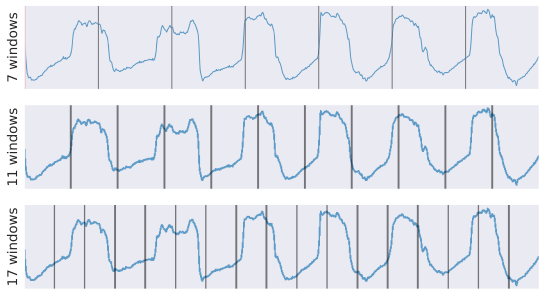
7 Discussion

The results show that developing a model with wearable signals to estimate ease of breathing is feasible. The empirical exploration of different aspects of the model helps to understand which signal modalities improve model performance, both individually and in combination.

The accelerometer is the best-performing signal and slightly outperforms the airflow signal. Although the airflow signal can be considered the gold standard for respiratory information, the parasternal accelerometer signal may capture more information about the patient useful for estimating ease of breathing. This could be due to the combination of the low-frequency periodic information and the high-frequency muscle vibration information. However, the airflow signal outperforms the accelerometer when predicting increases in inspiratory load.



(a) Using multiple intuitive number of windows



(b) Using multiple prime number of windows

Fig. 9 The different splits of an instance into different window sizes, where red edges indicate presence in another split

In contrast, the bio-impedance signal performs worse than the other signals, despite its correlation with respiratory patterns. This suggests that the measured bio-impedance signal contains less relevant information about the ease of breathing than the other signals. This limitation may be due to the inherent characteristic of the signal modality, but it could also result from the lower sampling frequency (16Hz as opposed to 10,000Hz). Future studies could collect high-frequency bio-impedance signals to validate these findings.

When comparing different window sizes, the highest F1-score was achieved with a window size of 55 s. This score was also higher than any window of any number of breaths. However, up until a larger number of breaths, the data can be used more efficiently, and it is possible that defining the window in breaths could be the most effective once more data points are obtained. It may also be more effective with a smaller dataset, as it retains a larger percentage of the data while achieving a similar performance.

Our models enable remote monitoring of ease of breathing for COPD patients, as the wearable measurements can potentially be taken at home without professional oversight. With regular measurements and comparisons to a baseline, and a final F1-score of 0.83, ease of breathing can be reliably estimated.

In practice, clinical integration of this model would involve defining a baseline with one or more initial measurements, which future measurements can be compared to. These can be performed at home by the patient themselves with minimal discomfort, in case of symptoms, or at any frequency the physician deems necessary.

Persistent changes observed over multiple measurements can be flagged for review. All relevant measurements along with the model's estimations and the interpretable saliency maps, can be forwarded to medical professionals.

Using multiple measurements as baselines could increase the reliability of the results by reducing the reliance on a single baseline measurement. Future measurements can be compared to these baselines, providing more extensive insights into the progression of the patient over time.

Some limitations of the current research should be noted. The data used in this study involves artificial obstructions, representing ease of breathing in increments of 12% of the maximal inspiratory pressure. Although the model effectively classifies the differences between these loads, these differences might be smaller or manifest differently in real-world settings. Real-world settings may introduce additional challenges, as data is often less reliable than in the clinical environments. Collecting data with different levels of breathing obstruction and settings would allow the training of more reliable models.

As a result, it remains unclear how well the model will generalize to real-world patients. The potential for generalization is likely improved by the comparative model being used, as differential features are less sensitive to distribution shifts than absolute features. If generalization remains suboptimal, it would be beneficial to fine-tune the model on real-world data.

Challenges related to the data collection arose. Parts of the signal are missing for some patients, as they took short breaks during the collection. Additionally, exhalations are missing for a subset of patients due to COVID-19 safety protocols. It is possible that guaranteeing uninterrupted measurements may improve model performance, but this is not always feasible in practice for remote monitoring and our approach has shown to be robust enough to reach an effective performance even with this missing data.

Although larger datasets are beneficial, collecting additional data is often challenging. Our dataset includes 66 patients, comparable to or larger than those in previous studies on similar applications (n = 16–62) [19, 29, 33, 37]. The results indicate that the size of our dataset is sufficient for training models that generalize well, with performance stabilizing with around 60% of the data used. Although additional data could be beneficial, the size of the dataset does not pose a limitation for this study.

The effectiveness of our approach is dependent on the feature extraction and selection. To maximize performance, we employed an extensive feature extraction process, drawing features from four-time series packages. This was paired with PowerSHAP, a robust and competitive feature selection method validated in prior studies and empirically confirmed in our use case. Our results show that the feature set enables effective modeling. Nonetheless, incremental improvements may be possible through more advanced feature extraction or selection techniques.

The proposed imputation method was shown to improve the representativeness of the original data distribution and to reduce the ratio of importance values equal to zero. These results highlight the effectiveness of periodic background sampling, though further explanation would be useful. Previous work [35] suggests comparing calculated importance values to estimated true importance values. However, this estimation involves removing features, which is exactly what is being evaluated. The sequential nature of time series data further complicates the use of similar evaluation methods.

Future work will address these challenges by collecting data from the same patients at different points in time. This will allow for longitudinal validation of the method, with observed changes in respiratory status reflecting real-world changes.

Deep learning techniques can also be evaluated when more data is available. Neural network models have shown to be effective on other respiratory problem settings [37] and could be compared to the current approach.

It is important that explanations are understandable to physicians and patients. Human-centered evaluation should be explored, to evaluate the effectiveness of the explanations with physicians.

8 Conclusion

In this paper, we defined a model to detect changes in ease of breathing for COPD patients and empirically examined multiple aspects of the model. The results of these experiments demonstrated that the accelerometer is the most effective wearable signal modality for this problem setting, with the parasternal accelerometer signal achieving the highest performance. When multiple signal modalities were considered, including audio recordings further improved model performance. These results highlight that an effective combination of minimally invasive signals can be acquired with a wearable device, opening new possibilities for remote monitoring of COPD.

The model was further enhanced with interpretability using time-based saliency maps and periodic background sampling, a novel imputation method for periodic signals. This method more accurately represents the original distribution and is less likely to miss relevant information. Additionally, we removed any dependence on window size and segmentation.

Funding This work was partially supported by the Flemish Government (AI Research Program).

Data availability Due to the sensitivity of the data, no supporting data for this article is publicly available.

Declarations

Ethics approval The study protocol was approved by the ZOL medical ethics committee (reference 18/0047U) and followed the World Medical Association's Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects. All patients provided written informed consent before participation.

Conflict of interest The authors declare no competing interests.

References

1. American Lung Association. COPD trends brief: burden. https://www.lung.org/research/trends-in-lung-disease/copd-trends-brief/copd-burden. Accessed 04 Jun 2023
2. World Health Organisation. global health estimates 2019: deaths by cause, age, sex by country and by region, 2000-2019. https://www.who.int/data/global-health-estimates. Accessed 04 Jul 2023

3. (2002) ATS/ERS statement on respiratory muscle testing. Am J Respir Crit Care Med 166:518–624
4. Al Rajeh A, Hurst J (2016) Monitoring of physiological parameters to predict exacerbations of chronic obstructive pulmonary disease (COPD): a systematic review. J Clin Med 5:108
5. Altan G, Kutlu Y, Allahverdi N (2020) Deep learning on computerized analysis of chronic obstructive pulmonary disease. IEEE J Biomed Health Inform 24:1344–1350
6. Barandas M, Folgado D, Fernandes L, Santos S, Abreu M, Bota P, Liu H, Schultz T, Gamboa H (2020) TSFEL: time series feature extraction library. SoftwareX 11:100456
7. Bentéjac C, Csörgő A, Martínez-Muñoz G (2021) A comparative analysis of gradient boosting algorithms. Artif Intell Rev 54(3):1937–1967
8. Bento Ja, Saleiro P, Cruz AF, Figueiredo MA, Bizarro P (2021) Timeshap: explaining recurrent models through sequence perturbations. In: SIGKDD conference on knowledge discovery & data mining, pp 2565–2573
9. Bhowmik RT, Most SP (2022) A personalized respiratory disease exacerbation prediction technique based on a novel spatio-temporal machine learning architecture and local environmental sensor networks. Electronics 11:2562
10. Blanco-Almazán D, Groenendaal W, Catthoor F, Jané R (2019) Wearable bioimpedance measurement for respiratory monitoring during inspiratory loading. IEEE Access 7:89487–89496
11. Blanco-Almazán D, Groenendaal W, Lozana Garcia M, Lijnen L, Smeets C, Ruttens D, Catthoor F, Jané R (2020) Combining bioimpedance and myographic signals for the assessment of COPD during loaded breathing. IEEE Trans Biomed Eng 68(1):298–307
12. Burns DM, Whyne CM (2018) Seglearn: a python package for learning sequences and time series. J Mach Learn Res 19:1–7
13. Bussone A, Stumpf S, O'Sullivan D (2015) The role of explanations on trust and reliance in clinical decision support systems. In: ICHI, pp 160–169
14. Christ M, Braun N, Neuffer J, Kempa-Liehr A (2018) Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). Neurocomputing 307:72–77
15. Crabbe J, van der Schaar M (2021) Explaining time series predictions with dynamic masks. In: International conference on machine learning
16. Das N, Topalovic M, Janssens W (2018) Artificial intelligence in diagnosis of obstructive lung disease. Curr Opin Pulm Med 24:117–123
17. Emanet N, Öz H, Bayram N, Delen D (2014) A comparative analysis of machine learning methods for classification type decision problems in healthcare. Decision Analytics 1:6
18. Exarchos KP, Aggelopoulou A, Oikonomou A, Biniskou T, Beli V, Antoniadou E, Kostikas K (2022) Review of artificial intelligence techniques in chronic obstructive lung disease. IEEE J Biomed Health Inform 26:2331–2338
19. Fernandez-Granero MA, Sanchez-Morillo D, Leon-Jimenez A (2018) An artificial intelligence approach to early predict symptom-based exacerbations of COPD. Biotechnology & Biotechnological Equipment 32:778–784
20. Gold WM, Koth LL (2015) Pulmonary function testing. Murray and Nadel's Textbook of Respiratory Medicine pp 407–435.e18
21. Guidotti R, Monreale A, Spinnato F, Pedreschi D, Giannotti F (2020) Explaining any time series classifier. In: IEEE International Conference on Cognitive Machine Intelligence (CogMI)
22. Guillemé M, Masson V, Rozé L, Termier A (2019) Agnostic local explanation for time series classification. IEEE International Conference on Tools with Artificial Intelligence (ICTAI)
23. Halbert RJ, Natoli JL, Gano A, Badamgarav E, Buist AS, Mannino DM (2006) Global burden of COPD: systematic review and meta-analysis. Eur Respir J 28:523–532
24. Hase P, Xie H, Bansal M (2021) The out-of-distribution problem in explainability and search methods for feature importance explanations. In: Advances in neural information processing systems, vol 34, pp 3650–3666
25. Kaplan A, Cao H, FitzGerald JM, Iannotti N, Yang E, Kocks JW, Kostikas K, Price D, Reddel HK, Tsiligianni I, Vogelmeier CF, Pfister P, Mastoridis P (2021) Artificial intelligence/machine learning in respiratory medicine and potential role in asthma and COPD diagnosis. J Allergy Clin Immunol Pract 9:2255–2261
26. Koff P, Jones R, Cashman J, Voelkel N, Vandivier R (2009) Proactive integrated care improves quality of life in patients with COPD. Eur Respir J
27. Kok TT, Groenendaal W, Blanco-Almazán D, Lijnen L, Smeets C, Ruttens D, Morales J, Dhaene T, Ongenae F, Van Hoecke S, Deschrijver D (2023) Comparator model for detecting changes in the ease of breathing of COPD patients. In: International joint conference on artificial intelligence, 6th international workshop on knowledge discovery from healthcare data
28. Kor CT, Li YR, Lin PR, Lin SH, Wang BY, Lin CH (2022) Explainable machine learning model for predicting first-time acute exacerbation in patients with chronic obstructive pulmonary disease. J Pers Med 12
29. Lanclus M, Clukers J, Van Holsbeke C, Vos W, Leemans G, Holbrechts B, Barboza K, De Backer W, De Backer J (2019) Machine learning algorithms utilizing functional respiratory imaging may predict COPD exacerbations. Acad Radiol 26(9):1191–1199
30. Lozano-García M, Sarlabous L, Moxham J, Rafferty GF, Torres A, Jané R, Jolley CJ (2018) Surface mechanomyography and electromyography provide non-invasive indices of inspiratory muscle force and activation in healthy subjects. Sci Rep 8
31. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Advances in neural information processing systems, vol 30, pp 4768–4777
32. McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, Nieto O (2015) librosa: audio and music signal analysis in python. In: Proceedings of the 14th Python in science conference
33. Merone M, Pedone C, Capasso G, Incalzi RA, Soda P (2017) A decision support system for tele-monitoring COPD-related worrisome events. IEEE J Biomed Health Inform 21:296–302
34. Mujkanovic F, Doskoč V, Schirneck M, Schäfer P, Friedrich T (2020) timeXplain - a framework for explaining the predictions of time series classifiers. arXiv:2007.07606 [cs, stat]
35. Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, Schlötterer J, van Keulen M, Seifert C (2022) From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable ai. ACM Comput Surv 55:1–42
36. Panaretos VM, Zemel Y (2019) Statistical aspects of Wasserstein distances. Annual Rev Stat Appl 6:405–431
37. Pastor-Serrano O, Lathouwers D, Perkó Z (2021) A semi-supervised autoencoder framework for joint generation and classification of breathing. Comput Methods Programs Biomed 209:106312
38. Pauwels RA, Buist AS, Calverley PM, Jenkins CR, Hurd SS (2001) GOLD Scientific Committee: global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease. NHLBI/WHO global initiative for chronic obstructive lung disease (GOLD) workshop summary. Am J Respir Crit Care Med 163:1256–1276
39. Shortliffe EH, Sepúlveda MJ (2018) Clinical decision support in the era of artificial intelligence. JAMA 320:2199–2200
40. Simonyan K, Vedaldi A, Zisserman A (2014) Deep inside convolutional networks: visualising image classification models and saliency maps. In: Workshop at international conference on learning representations

41. Spathis D, Vlamos P (2019) Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. Health Informatics J 25:811–827
42. Tomasic I, Tomasic N, Trobec R, Krpan M, Kelava T (2018) Continuous remote monitoring of COPD patients—justification and explanation of the requirements and a survey of the available technologies. Med Biol Eng Comput 56:547–569
43. Vellido A (2020) The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Appl 32:1–15
44. Verhaeghe J, Van Der Donckt J, Ongenae F, Van Hoecke S (2022) PowerSHAP: a power-full shapley feature selection method. In: ECML PKDD, pp 71–87
45. Vogelmeier CF, Criner GJ, Martinez FJ, Anzueto A, Barnes PJ, Bourbeau J, Celli BR, Chen R, Decramer M, Fabbri LM et al (2017) Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report. GOLD executive summary. Am J Respir Crit Care Med 195:557–582

Thomas T. Kok is a PhD Candidate in Computer Science Engineering at the PreDiCT team of IDLab, Ghent University-imec. His research interests include medical AI, time series, and interpretability.

John Morales received the PhD degree in Biomedical Data Processing at KU Leuven, and is an R&D Engineer at Imec Netherlands. His research interests include biomedical signal processing and analysis.

Dirk Deschrijver is an Associate Professor at IDLab, Ghent University-imec in SUMO Lab. His research interests include time series, predictive analytics and machine learning in health care, energy and manufacturing.

Dolores Blanco-Almazán received the PhD degree at Universitat Politècnica de Catalunya and was Postdoctoral Researcher at the Institute for Bioengineering of Catalonia. Her research interests include biomedical engineering.

Willemijn Groenendaal received the PhD degree at TU Eindhoven, and is currently Principal Member of Technical Staff at Imec. Her research interests include monitoring technology and algorithms in the clinical domain.

David Ruttens received the PhD degree at KU Leuven, and is head of the department of respiratory medicine at Ziekenhuis Oost Limburg, consultant at UZLeuven and Assistant Professor at UHasselt.

Christophe Smeets received the PhD degree at UHasselt, and is Scientific Research Coordinator at Ziekenhuis Oost Limburg. His research interests include clinical trials, wearable technology, and remote monitoring.

Vojkan Mihajlović received the PhD degree at the University of Twente, and is Principal Member of Technical Staff at Imec. His research interests include neuromodulation and functionally selective stimulation.

Femke Ongenae is Assistant Professor at IDLab, Ghent University-imec, and co-leads the PreDiCT team and the KnoWS team. Her research interests include semantic reasoning, hybrid AI, eHealth and data analytics.

Sofie Van Hoecke is Associate Professor at IDLab, Ghent University-imec, and leads the PreDiCT team. Her research interests include multimodal machine learning and dashboards for predictive maintenance and healthcare.