Controlled Breathing Effect on Respiration Quality Assessment Using Machine Learning Approaches

Andrea Rozo[1], Jeroen Buil[2], Jonathan Moeyersons[1], John Morales[1], Roberto Garcia van der Westen[2], Lien Lijnen[3], Christophe Smeets[4], Sjors Jantzen[5], Valerie Monpellier [5], David Ruttens[3,4]
Chris Van Hoof[6], Sabine Van Huffel[1], Willemijn Groenendaal[2], Carolina Varon[1,7]

[1] KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Belgium; [2] imec, Nederland/Holst Centre, Netherlands; [3] Universiteit Hasselt, Belgium; [4] Ziekenhuis Oost-Limburg, Belgium
[5] Nederlandse Obesitas Kliniek, Netherlands; [6] imec OnePlanet, Netherlands; [7] Service Chimie-Physique, Université libre de Bruxelles, Belgium

Abstract

Thoracic bio-impedance (BioZ) measurements have been proposed as an alternative for respiratory monitoring. Given the ambulatory nature of this modality, it is more prone to noise sources. In this study, two pre-trained machine learning models were used to classify BioZ signals into clean and noisy classes. The models were trained on data from patients suffering from chronic obstructive pulmonary disease, and their performance was evaluated on data from patients undergoing bariatric surgery. Additionally, transfer learning (TL) was used to optimize the models for the new patient cohort. Lastly, the effect of different breathing patterns on the performance of the machine learning models was studied. Results showed that the models performed accurately when applying them to another patient population and their performance was improved by TL. However, different imposed respiratory frequencies were found to affect the performance of the models.

1. Introduction

Chronic respiratory diseases are among the leading causes of death worldwide, with chronic obstructive pulmonary disease (COPD) and asthma being the most common illnesses. The diagnosis and monitoring of these diseases are carried out by devices that measure the pulmonary function. The traditional method for this is spirometry, which is a maximum breathing test to determine the ventilatory capacity of the lungs. When used in a clinical environment, it is a safe, practical and reproducible method. However, it does not allow continuous measurements and the lung function test should be performed by a trained operator [1]. Moreover, the use of a face mask or mouth piece to perform the test could result on altered breathing pattern of the subjects [2].

More comfortable and noninvasive modalities, such as thoracic bio-impedance (BioZ), have been proposed as alternatives to extract respiratory information. BioZ devices measure changes in the electrical impedance of the subject's thorax by injecting a high-frequency low-amplitude sinusoidal current. The impedance changes are mainly due to the lungs' air volume variations while breathing. This technique does not alter the breathing pattern of the subject and allows longer periods of analysis, since it can be applied with a wearable device [3].

However, given that this is a wearable monitoring technique, BioZ signals are more prone to noise sources, which result in the presence of artefacts and prevent the extraction of reliable features. Some of these artefacts are easily removed through filtering, but others, such as motion artefacts due to the displacement of the electrodes or stretching and wrinkling of the skin around the electrode, are still difficult to manage. One starting point to address these artefacts is to identify and remove segments of poor quality. This is investigated in [4], where a machine learning approach is proposed to detect contaminated segments in BioZ signals. Two models were presented, one convolutional neural network (CNN), and one support vector machine (SVM) classifier. These models were designed for a dataset of COPD patients.

The present study evaluates the performance of the pre-trained machine learning models in [4] when applied on patients undergoing bariatric surgery. Additionally, transfer learning (TL) was used to optimize the models for the new patient cohort. Lastly, the effect of different breathing patterns on the performance of the models was analyzed.

2. Methodology

2.1. Data

The dataset used to train the models consisted of the respiratory recordings of 47 COPD patients of the Ziekenhuis

ISSN: 2325-887X DOI: 10.22489/CinC.2021.220

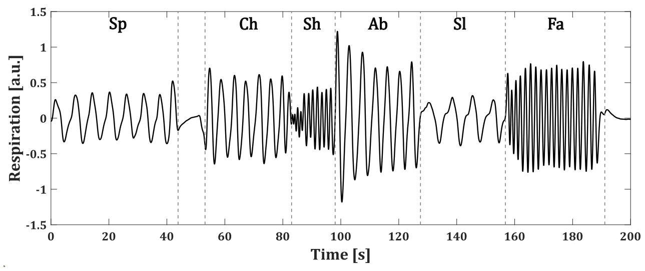

Figure 1. Controlled breathing protocol followed by a BS patient. Each of the breathing types are specified as: spontaneous (Sp), chest (Ch), shallow (Sh), abdominal (Ab), slow (Sl) and fast (Fa).

Oost-Limburg (Belgium). Each patient was equipped with a wearable device to measure the BioZ, as well as a traditional wired acquisition system, which measures respiratory airflow with an airflow transducer, used as gold standard. More details about this dataset can be found in [4,5].

The dataset used to test the models consisted of 72 respiratory recordings of 20 patients who underwent bariatric surgery (BS), and were in treatment at the Nederlandse Obesitas Kliniek (Netherlands). 14 patients were female and the mean BMI at inclusion was 42.5 ± 3.4 kg/m^2. The respiration of each patient was measured using a wearable device (BioZ), and a standard spirometer (gold standard). The wearable device has a sampling frequency of 16 Hz, while the spirometer uses a sampling frequency of 1000 Hz.

Each patient performed a controlled breathing protocol during the recording of their respiration. The protocol consisted of one minute of spontaneous breathing (Sp), followed by a period of breath holding and then five blocks of thirty seconds of chest (Ch), shallow (Sh), abdominal (Ab), slow (Sl) and fast (Fa) breathing. The pacing of the different breathing types was left to the patient's comfort. Figure 1 shows an example of the protocol followed by a patient.

2.2. Pre-Processing

Two pre-processing steps were performed in both datasets. Firstly, the signals were band-pass filtered using a Butterworth filter with cutoff frequencies at 0.05 Hz and 0.70 Hz, removing the baseline oscillations and high frequency content not related to breathing. Secondly, the signals were segmented. Each recording of the COPD dataset was divided into non-overlapping one-minute segments as described in [4], obtaining 1896 segments in total. For the segmentation of the recordings of the BS dataset, the first step was to synchronize the BioZ signals with the gold standard using the lag of the maximum cross-correlation between them to shift the BioZ signal. The recordings were then divided into thirty-second segments with and without an overlap of 15 seconds excluding the breath holding period, obtaining in total 2916 segments.

As the goal of this study was to observe the effect of different respiratory patterns on the performance of the classifiers, the segments of the BS dataset were grouped according to each of the six different types of breathing.

2.3. Labeling

The recordings from each dataset were labeled by four independent annotators with experience in biomedical signal analysis. The labeling was carried out using the graphical user interface (GUI) and the five classes defined in [4]. The classes 1 (Excellent signal quality), 2 (Good signal quality), 3 (Average signal quality) and 4 (Bad signal quality) refer to the quality of the BioZ signal with respect to the gold standard. The class 5 (Bad reference quality) is reserved for the cases where the gold standard is of bad quality due to acquisition problems, motion artefacts or signal saturation.

The classes for the BioZ signals were binarized, considering 1 and 2 as clean (1) and 3 and 4 as noisy (-1). Majority voting among annotators was performed to create a single label per signal. The segments in which the majority voting resulted in label 5 and the ones in which no majority voting was achieved were removed from further analysis. This resulted in a total of 1471 and 2298 segments for the COPD and the BS datasets, respectively.

2.4. Classification

The first classification model is a feature-based SVM classifier. Each input signal was normalized by subtracting its mean and dividing by its standard deviation before being used. Afterwards, features from the auto-correlation function (ACF) and from the power spectral density (PSD) of the signals were computed. These features were computed for the whole 60-seconds segment and for non-overlapping 15-seconds sub-segments. The five more informative features were selected using the minimum redundance, maximum relevance (MRMR) algorithm. These features were: the first peak of the ACF of the whole segment, the lower bound of the bandwidth of the whole segment, the standard deviation of the first peak of the ACF of the sub-segments, the mean normalized power of the sub-segments and the mean bandwidth of the sub-segments.

For the BS dataset, the segments were 30 seconds long and were then divided into 15-second sub-segments with a 10-second overlap.

The second classifier is a 1-dimensional CNN. Each input signal was first normalized by subtracting its mean and dividing by its standard deviation. The network was composed by two blocks of convolutional layers, having 10 filters in the first block and five in the second. Then, these feature maps were passed to a global average pooling layer, which generated a single feature vector. Finally, this vector was fed to a fully-connected output layer with a softmax activation function.

2.5. Transfer Learning

When solving a classification problem using machine learning algorithms, one of the main assumptions is that the training and testing data are in the same feature space and have the same distribution. However, in real applications, these assumptions do not always hold. One way to tackle this issue is to use TL [6]. The principle behind TL is to solve a new classification problem by using the solution from a similar problem as a starting point. In this way, less training data for the new classification problem is required to get a robust solution.

In this study, TL was used to optimize each of the classifiers for the BS dataset. For the SVM, the transfer learning approach described in [7] was used. This approach modifies the objective function of the SVM, minimizing the classification error of the new data and the dissimilarity between the original and the adapted SVM classifiers. One assumption of this approach is that the features used with the original data for the original classifier also describe the new data for the adapted classifier. TL was applied to the original SVM model for each breathing type.

In the case of the CNN, the same principle used in [8] was applied. In this approach the same weights of the trained CNN are reused on the adapted model. The weights of all layers besides the classification layer are fixed and then the classification layer is retrained. After, a fine-tuning (FT) step is added, in which all the weights of the adapted model are retrained. The retraining is done imposing a low learning rate and a small number of epochs to not modify significantly the effect of the TL. The only assumption of this approach is that the signals from the training and testing datasets should have the same sampling frequency. TL was applied to the original CNN to obtain an adapted model for each breathing type, followed by FT to obtain the second adapted model.

2.6. Performance evaluation

The models, before and after TL, were tested with a cross-validation approach. The division was done by taking all the segments of the 70 % of the recordings of the BS dataset for the retraining set for TL and the remaining 30 % in the actual test set. In order to assess the generalization capability of the models, this division was done 10 times at random. The same splits were used for all models.

Two performance metrics were used to evaluate and compare the performance of the models: the area under the curve (AUC) and the Cohen's kappa coefficient (κ), which measures the agreement between two raters. In this case, the two raters were the four annotators (majority voting) and the classification model.

Significant differences between the models before and after TL were evaluated with a Wilcoxon signed rank test (significant if $p < 0.05$) to assess the utility of TL for the BS dataset.

3. Results and Discussion

The total number of segments after labeling and removing the ones with bad reference quality, for each dataset, is presented in Table 1.

Table 1. Overview of the datasets, indicating the number of segments in each class. The suffix corresponds to the type of breathing imposed during the respiratory protocol.

Group	Clean (1)	Noisy (-1)	Total
COPD	1118	353	1471
BS-Sp	202	240	442
BS-Ch	181	317	498
BS-Sh	181	305	486
BS-Ab	208	256	464
BS-Sl	83	216	299
BS-Fa	26	83	109

The classifiers were trained with all the available segments from the COPD dataset, and tested in each of the sub-groups from the BS dataset. The results for the SVM classifier before and after TL are presented in Table 2. Results are indicated as mean ± standard deviation.

Observing the performance of the original model, it was found that for the fast (Fa) and slow (Sl) breathing groups, it was worse than for the other breathing types. This behavior was also noted after applying TL.

It was found that the performance after applying TL was consistently better than for the original model for all the sub-groups, with a significant improved agreement on the labeling for the chest (Ch) breathing. It could be noted that for almost every breathing type, the model after TL had less variability in the results, which could indicate that TL improves the generalization capabilities of the SVM classifier.

Table 2. Performance of the SVM models (before and after TL) for each of the sub-groups of the BS dataset. The results are presented as mean ± standard deviation.

	SVM - original		SVM - TL	
Group	AUC (%)	κ	AUC (%)	κ
Sp	95.06	0.78	97.07	0.79
	±3.11	±0.10	±1.95	±0.07
Ch	89.85	0.71	93.57	0.73
	±4.96	±0.11	±1.69	±0.06
Sh	84.42	0.56	92.29	0.62
	±6.76	±0.14	±4.88	±0.10
Ab	90.26	0.64	94.42	0.64
	±3.52	±0.09	±2.75	±0.09
Sl	80.32	0.51	85.22	0.53
	±2.83	±0.11	±5.66	±0.21
Fa	72.02	0.36	74.30	0.35
	±13.18	±0.23	±13.25	±0.22

The results for the CNN models are shown in Table 3. As was the case with the SVM, with the CNN the performance of the original model for Fa and Sl groups was worse than for the other groups, even after applying TL and FT.

The performance of the CNN after TL only showed significant improvements in the agreement of labeling for Sp, Sh and Ab breathing, compared to the original model. In contrast, after FT the general performance significantly improved for Sh and the agreement in labeling showed significant improvements also for Sp and Sl.

It was noticed that the values of the standard deviation of the AUCs and κ for the models after TL and FT remained close to the values of the original model. This could suggest that the generalization capability of the original CNN is good, and is not much affected by the use of a new dataset. Additionally, it was seen that the values of σ of the CNN models were lower than the ones of the original SVM, and comparable to the ones of the SVM after TL. This could mean that the CNN generalizes better than the original SVM, and that TL improves the generalization capability of the latter making it closer to the one of the CNN.

The probabilities of being classified as clean, given by the output of the classifier layer of the CNN, were also observed. They showed a consistent increment when using TL and FT for all breathing types, except for the fast breathing.

Table 3. Performance of the CNN models (before and after TL and FT) for each of the sub-groups of the BS dataset. The results are presented as mean ± standard deviation.

	CNN - original		CNN - TL		CNN - FT	
Group	AUC (%)	κ	AUC (%)	κ	AUC (%)	κ
Sp	94.73	0.71	94.66	0.74	94.76	0.77
	±1.47	±0.05	±1.51	±0.05	±1.63	±0.04
Ch	93.30	0.66	93.27	0.64	93.80	0.69
	±1.54	±0.08	±1.60	±0.08	±1.28	±0.05
Sh	89.64	0.52	89.60	0.56	90.00	0.62
	±3.12	±0.09	±2.98	±0.09	±2.83	±0.07
Ab	92.82	0.65	92.76	0.68	92.95	0.67
	±1.74	±0.05	±1.96	±0.06	±2.12	±0.08
Sl	85.48	0.42	85.38	0.44	85.34	0.46
	±5.72	±0.12	±5.59	±0.12	±5.63	±0.10
Fa	82.80	0.30	79.57	0.40	79.67	0.40
	±10.30	±0.24	±14.61	±0.30	±13.92	±0.26

It is worth noting that the distributions of clean and noisy segments of both datasets (COPD and BS) were not the same. The COPD dataset presented a majority of clean segments, while for the BS dataset there were more noisy segments. Additionally, the fact that the COPD data contained mainly spontaneous breathing, could have an effect in the performance of the models when applied to signals with different breathing rates. Also, the total number of segments in the groups of Sl and Fa breathing are lower compared to the other types of breathing. This could induce a bias when assessing the performance of the models in these groups. As future work, it is proposed to study the effect of balancing both datasets when training and testing the classifiers, in terms of classes and in terms of breathing frequencies, as well as to study some data augmentation techniques when using the CNNs.

4. Conclusions

The results presented in this study show that both classifiers for the quality assessment of respiratory signals performed accurately when tested in a patient cohort different to the one in which they were trained. In general, for both classifiers before and after TL, the results showed a lower performance for slow and fast breathing. This suggests that the features used by the classifiers are more affected by changes in the signal's morphology combined with the imposed breathing rates that differ most from the spontaneous breathing.

Acknowledgements

KU Leuven STADIUS acknowledges the financial support of imec. This research received funding from the Flemish Government (AI Research Program). S.V.H., A.R., J.M. and J.M. are affiliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium. C.V. acknowledges the financial support of ESA, BELSPO.

References

[1] Johns DP, et al. Diagnosis and early detection of COPD using spirometry. JTD 2014;6(11):1557–1569.

[2] Askanazi J, Silverberg PA, Foster RJ, Hyman AI, Milic-Emili J, Kinney JM. Effects of respiratory apparatus on breathing pattern. Journal of Applied Physiology 4 1980; 48(4):577–580.

[3] Blanco-Almazan D, et al. Wearable bioimpedance measurement for respiratory monitoring during inspiratory loading. IEEE Access 2019;7:89487–89496.

[4] Moeyersons J, et al. Artefact detection in impedance pneumography signals: A machine learning approach. Sensors 2021;21(8):1–17.

[5] Blanco-Almazan D, et al. Combining bioimpedance and myographic signals for the assessment of COPD during loaded breathing. IEEE TBME 2021;68(1):298–307.

[6] Weiss K, et al. A survey of transfer learning, volume 3. Springer International Publishing, 2016. ISBN 4053701600.

[7] De Cooman T, et al. Personalizing heart rate-based seizure detection using supervised SVM transfer learning. Frontiers in Neurology 2020;11(February):1–13.

[8] Nanni L, et al. Comparison of transfer learning and conventional machine learning applied to structural brain MRI for the early diagnosis and prognosis of Alzheimer's disease. Frontiers in Neurology 2020;11(November):1–15.

Address for correspondence:

Andrea Rozo
ESAT/STADIUS/KU Leuven
Kasteelpark Arenberg 10, bus 2446, 3001 Leuven, Belgium.
ca.rozo2200@gmail.com