Feasibility Study of Spirometric Parameter Estimation from Exhaled Breath Sounds Using Time–Frequency Representations and Deep Learning

A. Salvador-Navarro[1][a], J. De la Torre-Cruz[1][b], A. Muñoz-Montoro [1][c], P. Revuelta-Sanz[2][d], J. M. Cruz-Molina[3], R. P. Paiva[4][e] and F. J. Canadas-Quesada[1][f]

[1]*Department of Telecommunication Engineering, University of Jaen, Campus Cientifico-Tecnologico de Linares, Avda. de la Universidad, s/n, Linares (Jaen), 23700, Spain*

[2]*Department of Computer Science, University of Oviedo, Campus de Gijon s/n, Gijon (Asturias), 33203, Spain*

[3]*Pulmonology Department, Hospital Universitario San Agustin de Linares, Avda. San Cristobal s/n, Linares (Jaen), 23700, Spain*

[4]*University of Coimbra, CISUC/LASI – Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering, Coimbra, Portugal*

salvador@ujaen.es

Keywords: Spirometry, FVC, FEV_1, PEF, Exhaled Breath Sounds, Time–Frequency Representation, Deep Learning.

Abstract: Respiratory diseases are a major global health concern, requiring accessible and reliable tools for lung function assessment. Although spirometry remains the gold standard, its use is often limited by equipment availability and patient cooperation. Recent advances in machine learning (ML) enable the estimation of spirometric parameters from respiratory sounds, offering a non-invasive and low-cost alternative. This work investigates the feasibility of estimating forced vital capacity (FVC), forced expiratory volume in one second (FEV_1) and peak expiratory flow (PEF) from exhaled breath sounds. Several time–frequency (TF) representations and configurations are benchmarked to determine the most effective approaches, including hybrid combinations that integrate complementary information. A deep learning (DL) framework based on pre-trained convolutional neural networks (CNNs) is developed to automatically extract features and perform parameter regression. Preliminary results confirmed the feasibility of estimating spirometric parameters from respiratory sounds, obtaining $R^2 = 0.51$ for FVC using STFT, $R^2 = 0.38$ for FEV_1 using Mel-spectrograms and $R^2 = 0.47$ for PEF using CQT, encouraging further research in this emerging and clinically relevant field.

1 INTRODUCTION

Respiratory diseases rank as the third leading cause of death worldwide, accounting for approximately 4 million deaths each year and affecting more than 450 million people, according to the Global Burden of Disease (GBD) study (Momtazmanesh et al. (2023)). The European Respiratory Society (ERS) and the American Thoracic Society (ATS) classify these conditions into three categories (Stanojevic et al. (2022)): obstructive diseases characterized by airflow limitation due to airway narrowing or blockage such as, asthma and Chronic Obstructive Pulmonary Disease (COPD); restrictive diseases which limit lung expansion and reduce overall lung volume including pulmonary fibrosis and neuromuscular disorders; and mixed patterns, where obstruction and restriction coexist, typically in advanced respiratory disorders. All three categories commonly result in breathlessness, reduced physical capacity, and progressive pulmonar deterioration (O'Donnell et al. (2007)).

Spirometry is the gold standard for assessing pulmonary function measuring key spirometric parameters (Bailey (2012)), including the forced expiratory volume in the first second (FEV_1), the forced vital capacity (FVC) and their ratio (FEV_1/FVC), which is particularly useful for identifying obstructive patterns. In addition, the peak expiratory flow (PEF), which measures the highest flow rate achieved during

[a] https://orcid.org/0009-0005-9636-2237
[b] https://orcid.org/0000-0002-6291-4698
[c] https://orcid.org/0000-0001-9518-8955
[d] https://orcid.org/0000-0002-5150-1046
[e] https://orcid.org/0000-0003-3215-3960
[f] https://orcid.org/0000-0002-3873-6078

a forced exhalation, is closely linked to the diagnosis of obstructive diseases such as asthma (Ponce et al. (2023)), since it tends to decrease when airway narrowing or obstruction is present. Despite its clinical relevance, access to spirometry is often limited, particularly in rural and low-resource settings (Masekela et al. (2019)). Even in advanced healthcare systems, spirometry has its limitations, mainly because it requires patient cooperation that can be especially challenging for children, older adults, and people with disabilities, making more difficult the early diagnosis.

Spirometry devices have evolved considerably since Hutchinson's first water-sealed spirometer in 1844 (Spriggs (1977)). The introduction of flow-sensing spirometers based on turbine represented a major technological advance. Modern devices with integrated microprocessors are compact, portable and capable of accurate real-time pulmonary measurements (De Jongh (2008)). Recent approaches have emerged to estimate spirometric parameters from respiratory sounds using machine learning (ML). Although publicly available datasets are still limited, some studies have begun to explore dual-modality recordings. For instance, Gupta et al. (Gupta et al. (2024)) collected data from 137 participants, combining acoustic signals acquired with a digital stethoscope during spirometry testing to classify respiratory patterns. In another study, Xu et al. (Xu et al. (2022)) compiled 309 cough recordings from 133 participants using a smartphone, aiming to predict spirometric parameters obtained independently from forced expiratory manoeuvres. Beyond these datasets, novel methods have been proposed, including SpeechSpiro (Vatanparvar et al. (2021)), which applies a hybrid convolutional neural network (CNN) and long short-term memory (LSTM) architecture to speech recordings for mobile health tracking and approaches leveraging cough acoustics, either through handcrafted acoustic features (Xu et al. (2022)) and deep learning (DL) pipelines that transform coughs into spectrograms and apply a residual neural network combined with patient data and a support vector regressor (SVR) to estimate spirometric parameters (Xu et al. (2023)).

In this work, we explore the feasibility of estimating spirometric parameters from exhaled breath sounds by evaluating a benchmark of different time–frequency (TF) representations to identify the most suitable TF approach for this biomedical task. We further investigate whether hybrid approaches, obtained by combining multiple TF representations, provide significant improvements over individual methods. Beyond this comparative analysis, we propose a DL–based framework for estimating spirometric parameters from forced exhaled sounds. The framework leverages pre-trained CNN architectures to automatically extract features from TF structures, which are subsequently used to predict three spirometric parameters: FVC, FEV_1 and PEF, allowing for a detailed assessment of which TF representation performs best for each individual parameter.

The organization of the paper is as follows. Section 2 describes the set of TF representations used in the study. Section 3 presents the DL feature framework, structured into pre-processing, feature extraction, regression and training. Section 4 reports the evaluation, covering the dataset, setup, metrics and the experimental results. Finally, Section 5 describes the conclusions and future work.

2 TIME-FREQUENCY REPRESENTATIONS

This section briefly reviews the mathematical and signal processing background underlying the TF structures evaluated in this work. In this section, we consider a sound recording input signal $x[n]$ sampled at a rate of f_s Hz.

2.1 STFT Spectrogram

The spectrogram is the most used TF representation, depicts the energy distribution of signals such as respiratory sounds across time and frequency. It is computed via the Short-Time Fourier Transform (STFT), where the spectrogram $\mathbf{X}_c \in \mathbb{C}^{K\times L}$ with K frequency bins and L time frames is obtained by evaluating each coefficient $\mathbf{X}_c(k,l)$ as follows:

$$\mathbf{X}_c(k,l) = \sum_{n=0}^{N-1} x[(l-1)\cdot J + n]\cdot w[n]\, e^{-j2\pi kn/N} \quad (1)$$

where $w[n]$ denotes the N-sample analysis window using a hop size J (in samples), where $k = [0,\ldots,K-1]$ and $l = [1,\ldots,L]$. In most sound processing tasks, such as detection, classification, or estimation, it is common to use only the magnitude spectrogram $\mathbf{X} = |\mathbf{X}_c| \in {\mathbb{R}_+}^{K\times L}$.

Since the STFT spectrogram provides a fixed linear frequency resolution, it can be suboptimal for low-frequency respiratory components where relative resolution is desirable. As a result, its performance may degrade when analyzing auscultated respiratory sounds in noisy environments (Das et al. (2020)).

2.2 Mel-Spectrogram

The Mel-spectrogram attempts to replicate the human auditory perception. It allocates higher resolution to the low spectral band and vice versa, consistent with the nonlinear frequency sensitivity of the human cochlea. This is achieved by mapping the linear frequency f onto the Mel scale $M(f)$, which reflects the approximately logarithmic rather than linear perception of pitch, as described in Eq. (2).

$$M(f) = 1127 \cdot \log\left(1 + \frac{f(\text{Hz})}{700}\right) \tag{2}$$

The Mel-spectrogram is computed by analyzing the energies from a set of Mel filter bank (MFB). Specifically, the energy $E(c,l)$ extracted from the c-th filter in the l-th frame can be obtained according to Eq. (3),

$$E(c,l) = \log_{10}\left(\sum_{k=0}^{\frac{N}{2}-1} V(c,k) \cdot |\mathbf{X}_c(k,l)|\right), c = 1,...,C \tag{3}$$

where $V(c,k)$ represents the normalized response of the filter bank distributed along the Mel scale, C is the number of Mel filters and $\mathbf{X}_c$ refers to the STFT spectrogram (see Section 2.1) using a N-sample window.

2.3 Constant-Q Transform

The Constant-Q Transform (CQT) is particularly well-suited for signals with non-uniform spectral content, such as musical or physiological signals (Purwins et al. (2019)). Like the Mel-spectrogram, it employs a logarithmic frequency scale, which can provide a more accurate representation of respiratory sounds given their typically non-uniform frequency distribution. The CQT can be expressed mathematically as:

$$X^{\text{CQT}}(t,k) = \frac{1}{N_k} \sum_{n=0}^{N_k-1} x[(t-1)J + n]\, w[n,k]\, e^{-j\frac{2\pi kn}{N_k}} \tag{4}$$

where $x[\cdot]$ denotes the input signal, J the hop size, and k the frequency-bin index in the CQT domain. The TF atoms $a_k(n)$ are defined as:

$$a_k(n) = \frac{1}{N_k}\,\omega\left(\frac{n}{N_k}\right) e^{-j2\pi n\frac{f_k}{f_s}} \tag{5}$$

where f_k is the center frequency of the k-th bin, $\omega(\cdot)$ is the window function and $N_k \in \mathbb{Z}^+$ is the window length, which is inversely proportional to the f_k.

The quality factor Q is defined as:

$$Q = \frac{f_k}{\Delta f_k} \tag{6}$$

where Δf_k represents the -3 dB bandwidth of the frequency response of the atom $a_k(n)$. The f_k are logarithmically distributed as:

$$f_k = f_1\, 2^{\frac{k-1}{b}} \tag{7}$$

with f_1 being the f_k of the lowest bin and b the number of bins per octave. The parameter b controls the trade-off between time and frequency resolution in the CQT.

Although both CQT and Mel-spectrogram employ a logarithmic frequency scale, the CQT offers adaptive spectral resolution that could be suitable to learn more details in low-frequency regions where adventitious respiratory sounds, such as wheezes, are located.

2.4 Cochleagram

The cochleagram (COCH) is a TF built on the gammatone filter, which models how the human cochlea perceives sounds by frequency (Valero and Alias (2012); Patterson et al. (1987)). Unlike filters with evenly spaced bands, it uses a non-uniform resolution, specifically, low frequencies are analyzed with narrow filters, while higher frequencies use wider ones, matching how our hearing actually works, obtaining improved robustness to noise and acoustic variations. The COCH is computed using a bank of gammatone filters (Das et al. (2020); Chen et al. (2014)), where the impulse response of each filter, $g(t)$ is shown in Eq. (8),

$$g(t) = t^{o-1} \cdot e^{-2\pi b(f_c)t} \cdot \cos(2\pi f_c t), \quad t > 0 \tag{8}$$

In this context, o represents the filter order, while $b(f_c)$ denotes the exponential decay coefficient corresponding to the center frequency (f_c) in hertz, which determines the filter's bandwidth (Chen et al. (2014)). The center frequencies are distributed uniformly along the equivalent rectangular bandwidth (ERB) scale, as expressed in Eq. (10). In this work, the filter order was fixed at $o = 4$, a configuration that has previously demonstrated effectiveness in biomedical signal analysis; for example, in the classification of adventitious respiratory sounds (Mang et al. (2023)).

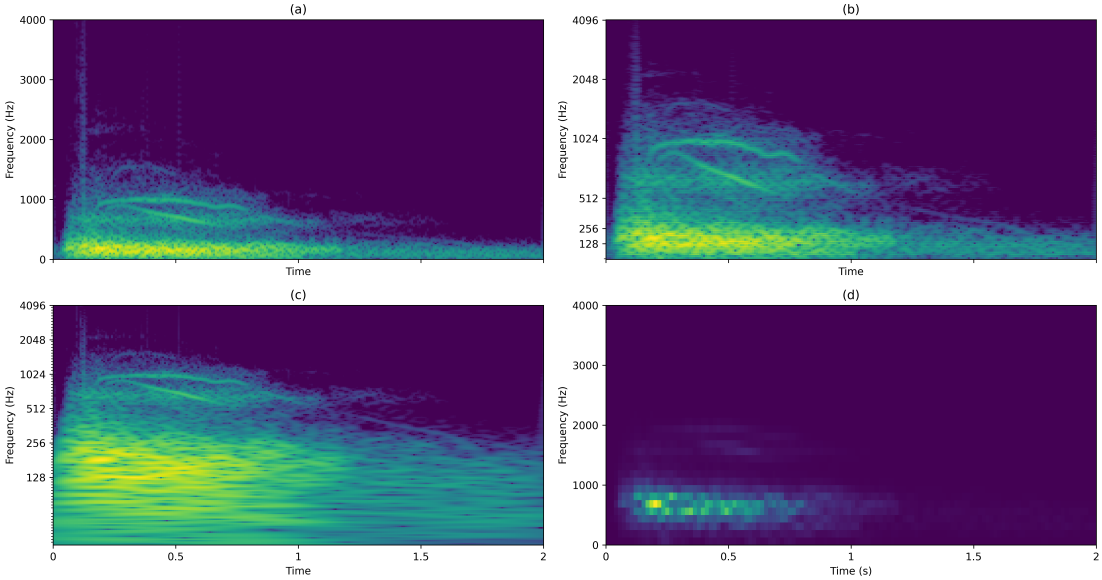

Figure 1: Magnitude spectrograms of a 2-second forced exhalation obtained using different TF representations: (a) STFT, (b) Mel-spectrogram, (c) CQT and (d) COCH.

$$b(f_c) = 1.019 \cdot \mathrm{ERB}(f_c) \tag{9}$$

$$\mathrm{ERB}(f_c) = 24.7\left(4.37\frac{f_c}{1000} + 1\right) \tag{10}$$

Afterwards, the $x[n]$ is processed through each filter $g(t)$. The resulting signals are divided into overlapping frames of length N with a hop size of J. The COCH representation is then derived by calculating the frame-wise power across all channels, following the methodology presented in (Chen et al. (2014)).

3 DEEP FEATURE LEARNING

This work presents a DL-based framework for estimating FVC, FEV_1 and PEF from acoustic signals recorded during forced exhalations. The approach follows a three-stage pipeline: (i) pre-processing, transforming raw sound into TF representations; (ii) feature extraction, deriving relevant descriptors from these representations; and (iii) regression, mapping the extracted features to spirometric outcomes. A schematic of this workflow is provided in Figure 2.

3.1 Pre-Processing

Each input respiratory signal $x[n]$ was zero-padded to a maximum duration of 6-seconds to standardize signal lengths across recordings. The waveform was then normalized to reduce amplitude variability. From the normalized signal, multiple TF representations $\mathbf{X} \in \mathbb{R}^{F\times T}$ were computed, including the STFT, Mel-spectrogram, CQT and COCH. Here, F denotes the number of frequency bins, which may be linearly or nonlinearly spaced depending on the TF representation, and T denotes the number of time frames, capturing spectral variations in energy content over time.

To ensure compatibility with used CNN architectures (He et al. (2016)), each TF representation was resized to 256×256, standardizing ratios, reducing computational cost and promoting a consistent feature distribution throughout the dataset, which improves training stability and generalization (Deng et al. (2009); Howard et al. (2017)). Finally, each single-channel spectrogram was replicated across three channels to create RGB-like inputs that are compatible with pre-trained network architectures.

In addition to the individual TF representations, a Combined-TF representation was generated by concatenating the TF configurations that achieved the best performance. The fusion was performed on the single-channel feature maps $\mathbf{X}_i \in \mathbb{R}^{F\times T}$ corresponding to each selected configuration. For each sample, the tensors were concatenated along the frequency axis to form a unified spectral–temporal representation $\mathbf{X}_{\mathrm{comb}}$, defined as:

$$\mathbf{X}_{\mathrm{comb}} = \mathrm{concat}(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M), \tag{11}$$

where M denotes the number of TF representations. The concatenated tensor was subsequently resized to a fixed spatial resolution of 256×256 to ensure dimensional consistency across samples. In a similar manner as previously mentioned, this composite map was replicated across three identical channels to

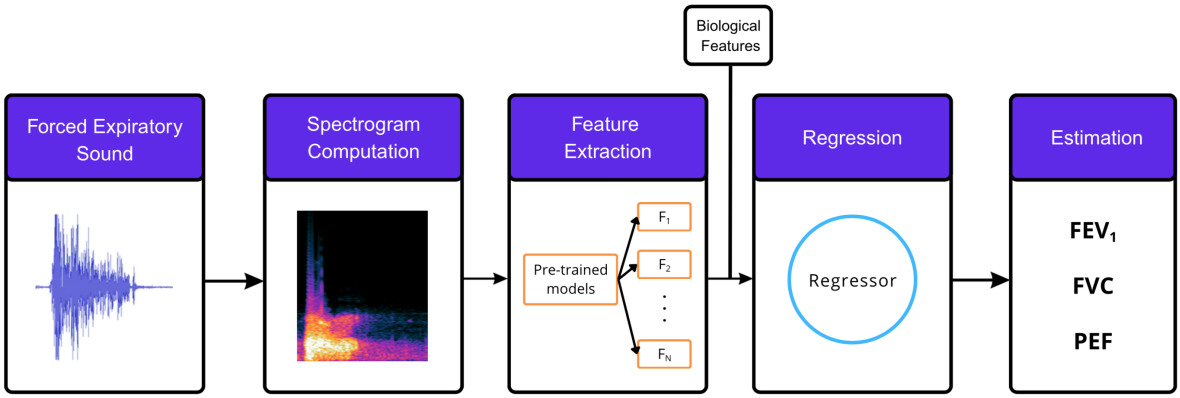

Figure 2: Flowchart of the proposed method.

match the input requirements of the CNN architectures employed in the experiments.

3.2 Feature Extraction

Discriminative features are automatically extracted from the TF representations computed in the previous stage using a Global Average Pooling (GAP) layer (Lee et al. (2016)). Instead of preserving the full spatial structure of these representations, GAP computes the mean activation within each channel, producing a fixed-length vector that summarizes the most relevant patterns (Lee et al. (2016)). This approach provides a strong inductive bias for regression tasks by mapping spatial features into a single descriptor per channel, drastically reducing the parameter space and facilitating the prediction of continuous variables (Zagoruyko and Komodakis (2016)).

Mathematically, the GAP operation (Dogan (2023)) is defined in Eq. 12, where $\mathbf{Z}$ denotes the output tensor of the last convolutional block and $Z_{ft}^{(k)}$ its activation at spatial location (f,t) for channel k. F' and T' correspond to the reduced frequency and time dimensions produced by the CNN architecture, and K is the total number of channels (e.g., $K = 2048$ for ResNet (He et al. (2016))). The output is a channel-wise average that condenses the activation patterns into a robust, fixed-length representation (Dogan (2023)). After obtaining the GAP features, Z-score normalization (Simonyan and Zisserman (2014)) is applied to ensure that all dimensions contribute proportionally to the learning process and to improve convergence across CNN architectures.

$$f_{GAP}^{(k)}(\mathbf{Z}) = \frac{1}{F' \cdot T'} \sum_{f=1}^{F'} \sum_{t=1}^{T'} Z_{ft}^{(k)}, \quad k = 1, \ldots, K \quad (12)$$

Finally, the feature vector is augmented with four auxiliary biological attributes: gender, age, height and weight. Incorporating these descriptors provides additional context that captures inter-individual variability and has been shown to improve model generalization and predictive stability (Xu et al. (2023)).

3.3 Regressor

To predict the spirometric parameters, the extracted features are mapped to a fully connected regression layer with three linear units, each corresponding to one parameter. This formulation allows the model to jointly optimize feature learning and the estimation of all spirometric indices in an end-to-end manner. Using a simple regression layer, instead of deeper or more complex dense structures, helps mitigate overfitting and ensures stable predictions across the three outputs (Bishop and Nasrabadi (2006)).

3.4 Training

During the training stage, a 5-fold cross-validation strategy was employed to ensure a robust evaluation of the model's generalization capability (Kim (2009)). The folds were partitioned by participant, ensuring that no signals from the same participant appeared in more than one fold. In each fold, one partition was reserved as the test set, while the remaining four were divided into training (80%) and validation (20%) subsets. This process was repeated until each fold had served once as the test set. The final performance indicators were obtained by averaging the results across all folds, thus reducing the variance of the estimates.

In this work, the ResNet50 architecture (He et al. (2016)) was selected as the backbone network, given its proven performance in biomedical signal analysis and its balance between depth and computational cost (Xu et al. (2023)). The model was optimized using the Adam algorithm, with ReLU activation func-

tions applied throughout the convolutional and dense layers. Training was conducted with a learning rate of 0.001, a batch size of 16 and a maximum of 100 epochs. An early stopping was incorporated with a patience of 20 epochs applied in the validation set.

4 EVALUATION

The evaluation process is described in this section by means of a dedicated spirometric dataset detailed in Section 4.1. The remainder of this section is structured by the setup (Section 4.2), the metrics (Section 4.3) and the experimental results (Section 4.4).

4.1 Dataset

Given the lack of publicly available labeled databases containing respiratory sounds of forced exhalations, a dedicated labeled spirometric database is being developed at the University of Jaén (UJA). For the present work, a subset of this dataset has been used, comprising recordings from 35 healthy participants (17 men and 18 women) aged between 18 and 65 years. The recordings were obtained under controlled laboratory conditions at the Superior Polytechnic School of Linares (EPSL), ensuring high-quality and consistent signal acquisition. Each participant performed the forced exhalations maneuver, during which the acoustic signal was recorded using a Thinklabs ONE digital stethoscope (Thinklabs (2025)), while spirometric parameters such as FVC, FEV_1, PEF, among others were simultaneously measured and stored using a Contec SP-10 portable turbine spirometer (Contec (2025)). Following the criterion of our medical expert, the stethoscope was placed alternately on the left and right sides of the neck for each participant, capturing three valid recordings per side, resulting in a total of 210 forced exhalation sound signals. All sound recordings are mono and were sampled at $f_s = 16$ kHz. The database was developed following rigorous ethical and technical standards according to ERS/ATS guidelines. Ethical approval was obtained from the Human Research Ethics Committee (CEIH) of the University of Jaén (UJA) and from the Regional Government of Andalusia (Junta de Andalucía) for the collection of data from patients with and without respiratory diseases, respectively.

Finally, a denoising process (Fitzgerald (2010)) was applied to each recording to reduce transient acoustic interferences, such as those caused by stethoscope friction or movement, as well as cardiac sounds captured by the stethoscope. All such transient components were treated as percussive sounds, with the aim of attenuating them to improve the estimation of spirometric parameters. For that purpose, a window length of approximately 168 ms was used, as we empirically observed that this length showed a satisfactory suppression of short-duration events while preserving most of the energy associated with forced exhalation sounds.

4.2 Setup

Several TF representations were evaluated using different parameter configurations to assess their impact on performance: (i) STFT was computed using Hamming windows of length $N = [256, 512, 1024, 2048]$ samples with the corresponding hop size $J = [128, 256, 512, 1024]$ samples; (ii) Mel-spectrograms employed $C = [32, 64, 128, 256]$ filters, frequency ranges $f_{\min} = [0, 50, 100]$ Hz and $f_{\max} = [4000, 8000]$ Hz; (iii) CQT was generated using minimum frequencies $f_{\min} = [0, 50, 100]$ Hz, a total of logarithmically spaced bins $n_{\mathrm{b}} = [48, 60, 72, 84]$ and $b = 12$ bins per octave; (iv) COCH was computed using gammatone filter banks with 32 or 64 channels, $N = [512, 1024]$ samples, $J = [256, 512]$ samples and $f_{\min} = [0, 50, 100]$ Hz.

4.3 Metrics

Four metrics have been considered to measure the estimation performance of the proposed method: (i) the Absolute Error (*AE*), which quantifies the magnitude of the deviation between the predicted and ground-truth value; (ii) the Mean Absolute Error (*MAE*), which provides the average of the absolute deviations; (iii) the Median Absolute Error (*medAE*), defined as the median of the absolute deviations and therefore less sensitive to outliers than the *MAE*; and (iv) the Coefficient of Determination (R^2), which measures the proportion of variance in the ground-truth values that can be explained by the model predictions. These metrics are formally defined as:

$$AE_i = \left|\hat{y}_i - y_i\right| \tag{13}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{14}$$

$$medAE = \mathrm{median}\left(\{AE_i\}_{i=1}^{n}\right) \tag{15}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{16}$$

where $\hat{y}_i$ denotes the estimated value of the spirometric parameter for the i-th sample, y_i the ground-truth value, $\bar{y}$ the mean of the ground-truth values and n the total number of evaluated recordings.

4.4 Results

The experimental evaluation was conducted to assess system performance under different TF representations. The analysis comprised two main stages. In the first stage (Section 4.4.1), each TF representation was independently analyzed across its parameter configurations to identify the most effective transform for estimating the target spirometric parameters. In the second stage (Section 4.4.2), the best-performing TF representation with its best parameter configuration from the previous stage were combined to examine whether hybrid or multi-representation approaches could further enhance the overall system performance.

4.4.1 Individual TF Representations

Results from the spirometric parameters estimation applying each TF configuration are shown in Figure 3. Higher R^2 values were found to correlate with lower *MAE* and *medAE*, indicating consistent model performance across metrics. Moreover, no single TF representation proves universally optimal across all spirometric parameters; rather, the effectiveness depends on the specific parameter being estimated, as can be seen in Figure 4. For FVC, the STFT with $(N,J) = (512,256)$ samples achieved an $R^2 = 0.51$, suggesting that relevant features are captured across both fine and coarse-scale temporal variations, with a balanced TF resolution that represents short-term fluctuations and overall spectral structure.

FEV_1 proved to be the most challenging parameter to estimate. Its best performance was achieved using the Mel-spectrogram with $C = 32$ filters, $f_{\text{min}} = 50$ Hz and $f_{\text{max}} = 4000$ Hz, resulting in an $R^2 = 0.38$. This fact suggests that relevant information is concentrated in the low-frequency range below 4 kHz. Besides, the use of the lowest number of filters seems to extract more spirometric relevant content, reducing the presence of spurious events that can distort the FEV_1 estimation.

Regarding PEF, the best estimation was obtained using the CQT with $f_{\text{min}} = 50$ Hz and $n_{\text{bins}} = 72$, achieving an $R^2 = 0.47$. This suggests that key information for PEF is concentrated in low and mid frequencies, and that 72 bins provide an optimal tradeoff between frequency resolution and spectral detail.

Finally, the COCH showed suboptimal performance in estimating all spirometric parameters. This limitation can be attributed to the fact that COCH emulates the frequency selectivity of the human auditory system, emphasizing perceptually salient spectral bands, particularly in the mid- and high-frequency ranges, while reducing the respiratory content of low-frequency patterns. In contrast, exhaled breath sounds that reflect airflow intensity and pulmonary function predominantly contain low-frequency energy and slow temporal modulations, which are closely tied to spirometric parameters. By prioritizing perceptual over physiological characteristics, the COCH attenuates the spectral features associated with airflow dynamics, thereby disrupting the correspondence between the captured respiratory signal and the underlying acoustic mechanisms of exhaled breath production. As a result, this perceptual bias constrains its capacity to accurately estimate spirometric parameters.

4.4.2 Combined TF Representations

The estimation performance of the three best TF representations (and their best configurations) from the previous stage are combined to analyze their joint effect in this context. As a result, four combinations are explored: STFT with Mel-spectrogram (SM), STFT with CQT (SC), Mel-spectrogram with CQT (MC) and the joint combination of STFT, Mel-spectrogram and CQT (SMC). The results are shown in Figure 5.

The results indicate that no single combination outperforms the best individual representation for each spirometric parameter. For FVC, the SM combination matches the performance of STFT alone ($R^2 = 0.51$), while SC and MC underperform slightly and SMC reaches 0.45. Similarly, for FEV_1, all combinations achieve modest R^2 values around 0.29–0.30, reflecting the intrinsic difficulty of estimating this parameter from exhaled breath sounds. Focusing on PEF, combinations that include CQT, such as MC and SMC, achieve R^2 values of 0.47 and 0.45, highlighting the inclusion of CQT provides complementary spectral information and helps compensate for the limitations of the other representations.

Taken together, these findings suggest that while combining TF representations does not necessarily improve the estimation of a specific parameter beyond the best single representation, the joint use of STFT, Mel-spectrogram and CQT provides the most robust overall representation, balancing the strengths and weaknesses of each approach and yielding stable predictions across all spirometric parameters, reporting a clear correlation between exhaled breath sounds and spirometric measurements. Although the obtained performance is limited by the small dataset and the complexity of the models, which require sufficient samples and variability, this preliminary study confirms the feasibility of this approach and opens a promising line of research in non-invasive, sound-based assessment of lung function based on exhaled pulmonary sounds.

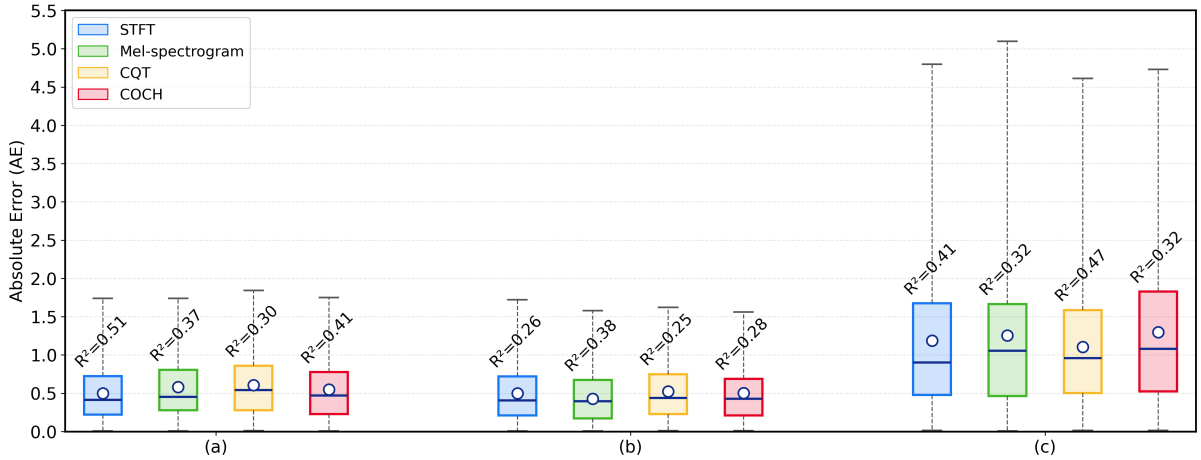

Figure 3: Performance metrics across different TF representations (Blue = STFT, green = Mel-spectrogram, yellow = CQT and red = COCH) for (a) FVC, (b) FEV_1 and (c) PEF. In each boxplot, the white dot indicates the *MAE*, the blue line represents the *medAE* and the metric R^2 is shown above each box.

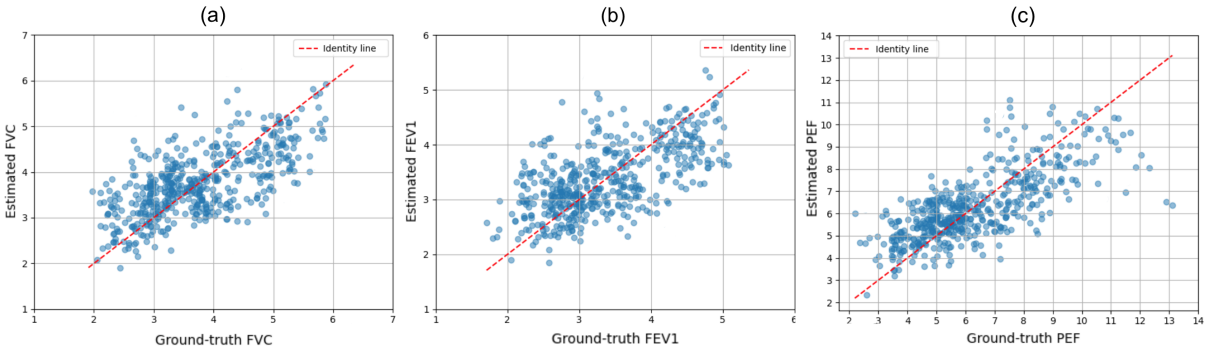

Figure 4: Comparison between ground-truth and estimated spirometric parameters using the optimal TF representation for each. Subplots correspond to (a) FVC, (b) FEV_1 and (c) PEF, using STFT, Mel-spectrogram and CQT, respectively.

5 CONCLUSIONS AND FUTURE WORK

In this work, we have demonstrated the feasibility of estimating spirometric parameters from forced-exhalation breath sounds, highlighting the potential of this novel and clinically relevant approach. To the best of our knowledge, this is the first study that analyzed the correlation between forced exhaled sound signals and the estimation of spirometric parameters.

Results indicate that TF representations such as STFT, Mel-spectrogram and CQT capture complementary information, and that their combination provides robust predictions across spirometric parameters, despite no single representation being optimal for all variables. The findings suggest that FVC estimation relies on the detailed spectral structure of the signal, FEV_1 is mainly determined by low-frequency components below 4 kHz, with lower spectral resolution providing a more stable representation and PEF depends on low- and mid-frequency bands, benefiting from representations that balance frequency resolution and spectral detail, such as CQT. Overall, these results confirm that different spectral representations emphasize distinct aspects of the expiratory signal and that combining them helps balance their strengths and weaknesses, leading to more reliable predictions across all spirometric parameters.

Future work will be focused on three directions. First, the limited size of the current spirometric sound database is the main limitation of this work, as the reduced number of recordings constrains model performance and generalization. Future efforts will therefore expand the database by increasing both the number of cases and the diversity of participants, including individuals with respiratory pathologies, to improve training and achieve more robust validation across broader populations. Second, a systematic

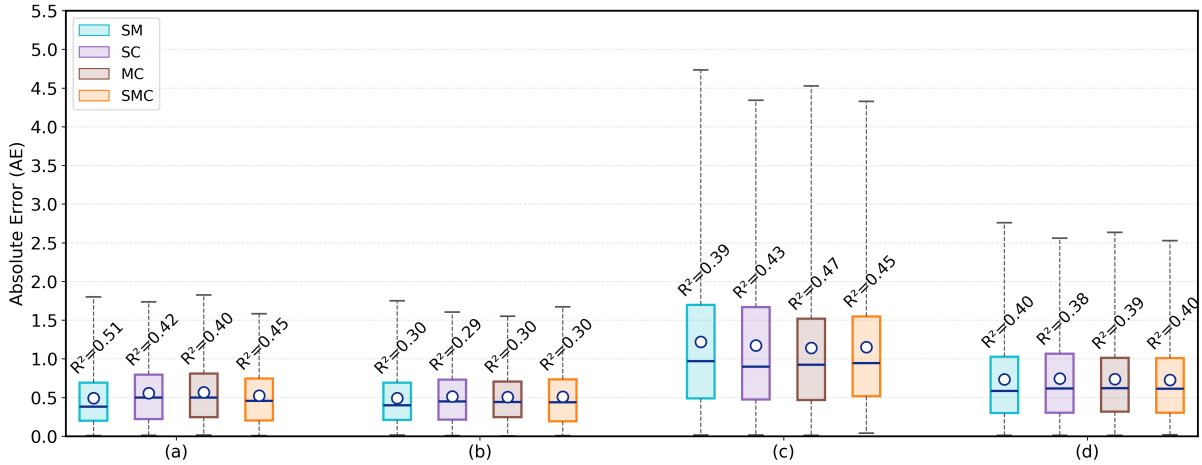

Figure 5: Performance metrics across combined TF representations (Cyan = SM, purple = SC, brown = MC and orange = SMC) for (a) FVC, (b) FEV_1, (c) PEF and (d) Overall. In each boxplot, the white dot indicates the *MAE*, the blue line represents the *medAE* and the metric R^2 is shown above each box.

analysis of manually engineered acoustic features using ML techniques, together with the incorporation of additional handcrafted acoustic features, simpler classical regressors, and various lower-complexity networks, will be conducted to provide efficient and interpretable alternatives to DL models. Third, the use of Physics-Informed Neural Networks (PINNs) could incorporate physiological pulmonary constraints into the estimation process, potentially enhancing the accuracy and reliability of predicted spirometric parameters.

FUNDING

This work was supported by the Grant PID2023-146520OB-{C21,C22} funded by MICIU/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by "ERDF/EU", by the "European Union" or by the "European Union NextGenerationEU/PRTR".

REFERENCES

Bailey, K. L. (2012). The importance of the assessment of pulmonary function in copd. *Medical Clinics of North America*, 96(4):745–752.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Chen, J., Wang, Y., and Wang, D. (2014). A feature study for classification-based speech separation at low signal-to-noise ratios. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1993–2002.

Contec (2025). Spirometer contec sp-10: https://www.gambarter.com. Last accessed: 2025-10-21.

Das, S., Pal, S., and Mitra, M. (2020). Acoustic feature based unsupervised approach of heart sound event detection. *Computers in biology and medicine*, 126:103990.

De Jongh, F. (2008). Spirometers. *Breathe*, 4(3):251–254.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Dogan, Y. (2023). A new global pooling method for deep neural networks: Global average of top-k max-pooling. *Traitement du signal*, 40(2):577–587.

Fitzgerald, D. (2010). Harmonic/percussive separation using median filtering.

Gupta, S., Agrawal, M., and Deepak, D. (2024). Correlating spirometry findings with auscultation sounds for diagnosis of respiratory diseases. *Biomedical Signal Processing and Control*, 87:105347.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 53(11):3735–3745.

Lee, C.-Y., Gallagher, P. W., and Tu, Z. (2016). Generalizing pooling functions in convolutional neural net-

works: Mixed, gated, and tree. In *Artificial intelligence and statistics*, pages 464–472. PMLR.

Mang, L. D., Cañadas-Quesada, F. J., Carabias-Orti, J. J., Combarro, E. F., and Ranilla, J. (2023). Cochleogram-based adventitious sounds classification using convolutional neural networks. *Biomedical Signal Processing and Control*, 82:104555.

Masekela, R., Zurba, L., and Gray, D. (2019). Dealing with access to spirometry in africa: a commentary on challenges and solutions. *International journal of environmental research and public health*, 16(1):62.

Momtazmanesh, S., Moghaddam, S. S., Ghamari, S.-H., Rad, E. M., Rezaei, N., Shobeiri, P., Aali, A., Abbasi-Kangevari, M., Abbasi-Kangevari, Z., Abdelmasseh, M., et al. (2023). Global burden of chronic respiratory diseases and risk factors, 1990–2019: an update from the global burden of disease study 2019. *EClinicalMedicine*, 59.

O'Donnell, D. E., Banzett, R. B., Carrieri-Kohlman, V., Casaburi, R., Davenport, P. W., Gandevia, S. C., Gelb, A. F., Mahler, D. A., and Webb, K. A. (2007). Pathophysiology of dyspnea in chronic obstructive pulmonary disease: a roundtable. *Proceedings of the American Thoracic Society*, 4(2):145–168.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., Rice, P., et al. (1987). An efficient auditory filterbank based on the gammatone function. In *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, volume 2.

Ponce, M. C., Sankari, A., and Sharma, S. (2023). Pulmonary function tests. In *StatPearls [internet]*. StatPearls publishing.

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., and Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 13(2):206–219.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Spriggs, E. A. (1977). John hutchinson, the inventor of the spirometer–his north country background, life in london, and scientific achievements. *Medical history*, 21(4):357–364.

Stanojevic, S., Kaminsky, D. A., Miller, M. R., Thompson, B., Aliverti, A., Barjaktarevic, I., Cooper, B. G., Culver, B., Derom, E., Hall, G. L., et al. (2022). Ers/ats technical standard on interpretive strategies for routine lung function tests. *European Respiratory Journal*, 60(1).

Thinklabs (2025). Thinklabs one digital stethoscope: https://www.thinklabs.com. Last accessed: 2025-10-21.

Valero, X. and Alias, F. (2012). Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. *IEEE transactions on multimedia*, 14(6):1684–1689.

Vatanparvar, K., Nathan, V., Nemati, E., Rahman, M. M., McCaffrey, D., Kuang, J., and Gao, J. A. (2021). Speechspiro: Lung function assessment from speech pattern as an alternative to spirometry for mobile health tracking. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 7237–7243. IEEE.

Xu, W., He, G., Pan, C., Shen, D., Zhang, N., Jiang, P., Liu, F., and Chen, J. (2022). A forced cough sound based pulmonary function assessment method by using machine learning. *Frontiers in Public Health*, 10:1015876.

Xu, W., He, G., Shen, D., Xu, B., Jiang, P., Liu, F., Lou, X., Guo, L., and Ma, L. (2023). A noval pulmonary function evaluation method based on resnet50+ svr model and cough. *Scientific Reports*, 13(1):22065.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. *arXiv preprint arXiv:1605.07146*.