Automated Diagnosis of Respiratory Diseases from Respiratory Sounds: A Reproducibility Case Study

Diogo Pessoa[1][a], João Garcia[1], Juan De La Torre-Cruz[2][b], Francisco Cañadas-Quesada[2][c]
and Rui Pedro Paiva[1][d]

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

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

{dpessoa, ruipedro}@dei.uc.pt, joaodavidgarcia2001@gmail.com, {jtorre, fcanadas}@ujaen.es

Keywords: Respiratory Diseases, Respiratory Sound, Machine-Learning, Differential Diagnosis, Reproducibility.

Abstract: The automatic diagnosis of respiratory diseases using lung sound recordings has attracted growing attention due to advances in machine learning and the increasing availability of open-access respiratory databases. However, many studies in the field report near-perfect results that are difficult to reproduce and seldom translate to real-world clinical contexts and applications. In this work, we present a reproducibility case study in which we replicate a published deep learning model for pulmonary disease classification based on convolutional and recurrent neural networks, by reproducing the original methodology and correcting its methodological flaws—most notably, the presence of data leakage arising from patient overlap between training and testing sets—we demonstrate that previously reported results were overly optimistic. By enforcing patient-level data separation, we observed a significant drop in the model's performance, suggesting limited generalization. This study highlights the importance of transparent and reproducible research practices, rigorous experimental evaluation setups, and the development of cross-database and domain-adaptive models to ensure clinically reliable and generalizable computer-aided diagnostic systems for respiratory sound analysis.

1 INTRODUCTION

Respiratory diseases remain one of the leading causes of morbidity and mortality worldwide, accounting for a substantial burden on healthcare systems and significantly lowering patients' quality of life. In fact, these diseases represent one of the biggest health concerns around the globe, causing one-sixth of all deaths worldwide (World Health Organization (WHO), 2024). Conditions such as chronic obstructive pulmonary disease (COPD), asthma, bronchiectasis, and lower respiratory tract infections are prevalent across all age groups and frequently lead to hospital admissions and long-term disability. Therefore, early detection and continuous monitoring of respiratory pathologies are crucial to enable timely clinical intervention and management, ultimately improving patient outcomes (Marques et al., 2013).

Despite clinicians using multiple techniques to diagnose subjects with respiratory diseases, including spirometry, medical imaging, and laboratory screening, traditional auscultation remains widely used, largely because of its ease of use, non-invasiveness, and accessibility. Through auscultation, it is possible to extract relevant information on the physiological context of the lungs and upper airways and analyze their acoustic patterns. While having many positive aspects, auscultation also presents significant drawbacks, as it is inherently subjective and depends on the clinician's experience, hearing acuity, and environmental conditions (Pessoa, 2025; Pessoa et al., 2022). Moreover, the lack of a quantitative, reproducible assessment limits its reliability, especially for longitudinal monitoring and large-scale screening. These limitations have motivated the development of computer-aided auscultation systems that combine acoustic signal processing with machine learning to provide objective, automated evaluation of respiratory sounds.

Recent advances in artificial intelligence (AI) and deep learning (DL) have further accelerated progress

[a] https://orcid.org/0000-0002-7783-7488
[b] https://orcid.org/0000-0002-6291-4698
[c] https://orcid.org/0000-0002-3873-6078
[d] https://orcid.org/0000-0003-3215-3960

in this field. By learning discriminative patterns directly from data, AI-based models can effectively capture subtle acoustic features associated with specific pathologies. Audio-based respiratory disease differential diagnosis, or screening, can be formulated as a classification task, with respiratory sounds as input and a categorical prediction of the trained respiratory conditions as output (Xia et al., 2022).

In the literature, this pathology-driven classification has been addressed by either classifying complete respiratory-sound audio recordings or solely classifying chunks of complete recordings, such as isolated breathing cycles (namely, using the RSD (Rocha et al., 2018; Rocha et al., 2019)). Despite the promising results reported in recent years, many studies on the automated diagnosis of respiratory diseases still suffer from significant methodological limitations that compromise the validity and reproducibility of their findings. One major problem is the improper splitting of data between sets, with recordings from the same patient appearing in both the training and test sets, a phenomenon known as data leakage. This approach leads models to focus on individual acoustic patterns rather than specific features related to the diseases themselves. This, in turn, will lead to overly optimistic results that do not translate well when the models are tested in new patients.

Another methodological concern arises from using data augmentation on both the training and test sets (as in (García-Ordás et al., 2020)), which can further exacerbate leakage and artificially inflate evaluation metrics. There are also biases that arise from including only specific patient groups without proper reasoning, which might influence the results due to the significant heterogeneity between subjects. Moreover, in general, many studies lack clear details about the methodologies used, making it very difficult for the research community to replicate and validate the findings accurately.

All these inconsistencies emphasize the importance of rigorous data splitting methods, patient-independent, transparent, and reproducible experimental processes, and openly shared implementations. In this article, our primary goal is to develop a case study in which we identify a methodologically flawed article and reproduce its methodology, including its flaws and the corresponding rectifications.

2 MATERIALS AND METHODS

In this section, we present the methodological steps associated with replicating the paper “Recognition of pulmonary diseases from lung sounds using convolutional neural networks and long short-term memory" (Fraiwan et al., 2021a). In particular, our analysis will focus on the impact of data leakage problems and patient selection.

2.1 Dataset

The paper combines two datasets, the Respiratory Sound Database (RSD) (Rocha et al., 2018; Rocha et al., 2019) and a dataset containing locally recorded stethoscope lung sounds acquired at King Abdullah University Hospital (KAUH), in Jordan (Fraiwan et al., 2021b). Table 1 summarizes the demographic information for each dataset, categorized by six pulmonary conditions: Healthy/Normal, Asthma, Pneumonia, Bronchiectasis, COPD, and Heart Failure

For both the KAUH and the RSD datasets, demographic details include the number of subjects (categorized by gender), mean age with standard deviation, and the total number of recordings. The KAUH database consists of 301 recordings from 103 subjects, and the RSD contains 920 recordings from 110 subjects.

2.2 Article Overview

In the paper “Recognition of pulmonary diseases from lung sounds using convolutional neural networks and long short-term memory", the authors proposed a two-stage pipeline for the automatic classification of respiratory disease: preprocessing and training/classification. The preprocessing stage aimed to improve signal quality and standardize the data. Signals were resampled to 4 kHz and segmented into 5-second windows. Noise reduction was achieved through one-dimensional wavelet smoothing, and displacement artifacts were removed using robust LOESS regression (rLOESS). Finally, z-score normalization was applied to standardize the signals.

The proposed architecture is presented in Figure 1. It combines convolutional and recurrent layers to classify pulmonary diseases. The CNN layers extract spatial features, while the BDLSTM layers capture and model their temporal dependencies. Besides the proposed model, the authors have also considered using solely a BDLSTM network or a CNN network.

The authors utilized a 10-fold cross-validation scheme for model evaluation. This approach divides the dataset into ten subsets, iteratively using one subset for testing and the remaining nine for training. However, the paper does not explicitly state whether recordings from the same patient were ensured to remain exclusively in either the training or validation sets. This omission raises concerns about po-

Table 1: Demographic information of the subjects.

Dataset	Category	Normal	Asthma	Pneumonia	BRON	COPD	HF
KAUH	Number of subjects	35 (24 M, 11 F)	32 (15 M, 17 F)	5 (3 M, 2 F)	3 (2 M, 1 F)	9 (8 M, 1 F)	19 (10 M, 9 F)
	Age (mean ± SD)	43 ± 20	46 ± 16	56 ± 10	37 ± 27	57 ± 10	59 ± 19
	Number of recordings	110	88	18	6	23	56
RSD	Number of subjects	26 (13 M, 13 F)	1 (1 F)	6 (3 M, 2 F)	13 (6 M, 7 F)	64 (48 M, 16 F)	N/A
	Age (mean ± SD)	12 ± 20	70	62 ± 29	25 ± 21	69 ± 8	N/A
	Number of recordings	135	4	148	116	779	N/A

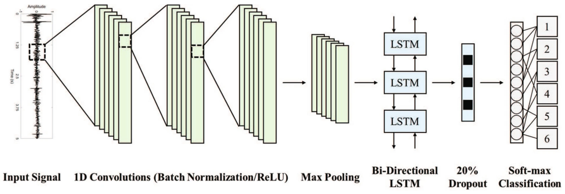


Figure 1: Model architecture (Adapted from (Fraiwan et al., 2021a)).

tential data leakage, where the model may inadvertently learn patient-specific characteristics. The performance of the proposed models was evaluated using metrics such as accuracy, precision, recall, F1-score, and specificity, along with the combined confusion matrix for all folds.

The CNN+BDLSTM model demonstrated very good performance, achieving near-perfect classification rates across all six pulmonary conditions. As shown in Figure 2 (c), the model correctly classified 98.8% of Healthy/Normal cases, 95.6% of Asthma, 98.8% of Pneumonia, COPD with 99.0% accuracy, while Bronchiectasis and Heart Failure achieved 100% accuracy.

While the reported metrics indicate outstanding performance, concerns about potential data leakage from improper handling of patient-level splits during k-fold cross-validation must be addressed. Without ensuring that data from the same patient did not appear in both training and test sets, the results may overestimate the model's generalization capabilities. Additionally, there appears to be cherry-picking, as 10 patients were removed from the RSD without explanation or specification of which patients were removed.

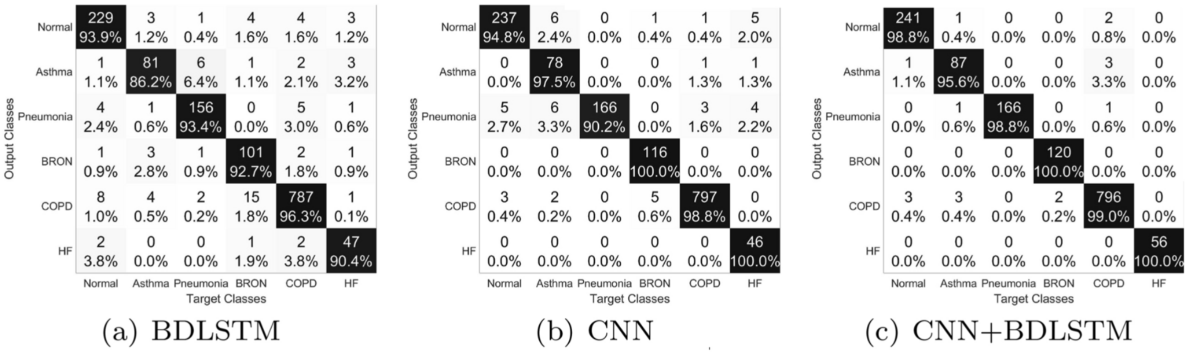
2.3 Article Replication

The replication process consisted of two main phases. The first phase involved replicating the methodology exactly as described in the previous section, with data leakage. Preprocessing techniques, such as resampling, denoising, and data normalization, were implemented, and the CNN + BDLSTM model was trained using identical hyperparameters and architecture. Its performance was evaluated using 10-fold cross-validation, without accounting for patient isolation across folds.

The second phase used patient-level splits to evaluate the impact of data leakage on the model's performance. Changes were made to the cross-validation procedure to ensure that patient data was completely isolated, simulating a scenario without data leakage. This scenario better resembles a realistic one, since the model is normally intended to be deployed in patients different from those whose data were used to train it.

3 RESULTS

As stated in the previous section, the article replication consisted of reproducing the article's results, both with and without accounting for patient isolation during data splitting. The performance metrics for the first phase, without patient isolation, are presented in Table 2. Despite achieving significant accuracy (83.52%) and high recall and precision for specific classes, these values are lower than those reported by the authors. Specifically, lower performance was observed for Bronchiectasis and Pneumonia, which tended to be misclassified as the most represented class, COPD. Figure 3 shows the combined confusion matrix of all folds obtained without patient isolation. The limited information on the class balance in each fold may explain the performance differences between the replication and the original study. COPD is prevalent in the dataset; therefore, the train/test sets of each fold may contain a significant percentage of



	Normal	Asthma	Pneumonia	BRON	COPD	HF
Normal	229 93.9%	3 1.2%	1 0.4%	4 1.6%	4 1.6%	3 1.2%
Asthma	1 1.1%	81 86.2%	6 6.4%	1 1.1%	2 2.1%	3 3.2%
Pneumonia	4 2.4%	1 0.6%	156 93.4%	0 0.0%	5 3.0%	1 0.6%
BRON	1 0.9%	3 2.8%	1 0.9%	101 92.7%	2 1.8%	1 0.9%
COPD	8 1.0%	4 0.5%	2 0.2%	15 1.8%	787 96.3%	1 0.1%
HF	2 3.8%	0 0.0%	0 0.0%	1 1.9%	2 3.8%	47 90.4%

(a) BDLSTM

	Normal	Asthma	Pneumonia	BRON	COPD	HF
Normal	237 94.8%	6 2.4%	0 0.0%	1 0.4%	1 0.4%	5 2.0%
Asthma	0 0.0%	78 97.5%	0 0.0%	0 0.0%	1 1.3%	1 1.3%
Pneumonia	5 2.7%	6 3.3%	166 90.2%	0 0.0%	3 1.6%	4 2.2%
BRON	0 0.0%	0 0.0%	0 0.0%	116 100.0%	0 0.0%	0 0.0%
COPD	3 0.4%	2 0.2%	0 0.0%	5 0.6%	797 98.8%	0 0.0%
HF	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	46 100.0%

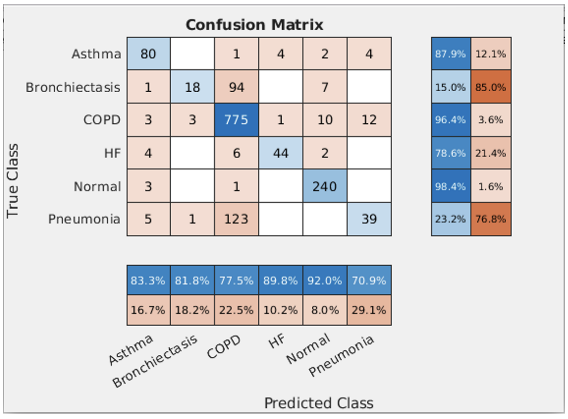
(b) CNN

	Normal	Asthma	Pneumonia	BRON	COPD	HF
Normal	241 98.8%	1 0.4%	0 0.0%	0 0.0%	2 0.8%	0 0.0%
Asthma	1 1.1%	87 95.6%	0 0.0%	0 0.0%	3 3.3%	0 0.0%
Pneumonia	0 0.0%	1 0.6%	166 98.8%	0 0.0%	1 0.6%	0 0.0%
BRON	0 0.0%	0 0.0%	0 0.0%	120 100.0%	0 0.0%	0 0.0%
COPD	3 0.4%	3 0.4%	0 0.0%	2 0.2%	796 99.0%	0 0.0%
HF	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	56 100.0%

(c) CNN+BDLSTM

Figure 2: Confusion matrix results for the different models (Adapted from (Fraiwan et al., 2021a)).

that class, since there is no specification regarding the 10-fold cross-validation stratification.

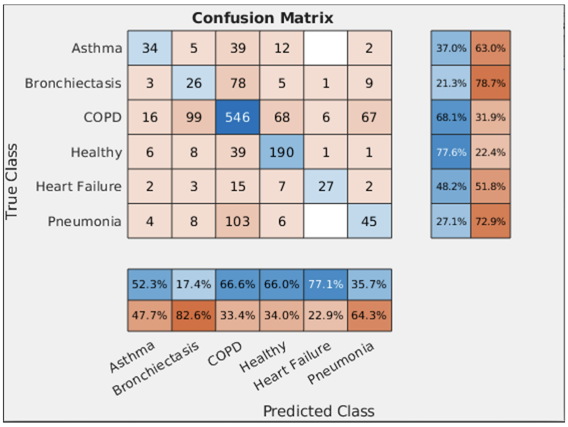


	Asthma	Bronchiectasis	COPD	HF	Normal	Pneumonia		
Asthma	80		1	4	2	4	87.9%	12.1%
Bronchiectasis	1	18	94		7		15.0%	85.0%
COPD	3	3	775	1	10	12	96.4%	3.6%
HF	4		6	44	2		78.6%	21.4%
Normal	3		1		240		98.4%	1.6%
Pneumonia	5	1	123			39	23.2%	76.8%
	83.3%	81.8%	77.5%	89.8%	92.0%	70.9%		
	16.7%	18.2%	22.5%	10.2%	8.0%	29.1%		

Figure 3: First phase results confusion matrix.

The second phase used patient-level splits to evaluate the impact of data leakage on model performance. Changes were made to cross-validation to ensure that data from a given patient was completely isolated, simulating a scenario without data leakage. The performance metrics obtained in this phase are presented in Table 3. Figure 4 shows the corresponding confusion matrix for all folds. When comparing with previous results, where patient data were split separately, we observe a significant impact on model performance, highlighting the importance of eliminating data leakage. In fact, all evaluation metrics dropped considerably compared to the first phase, indicating a more realistic evaluation scenario.

The analysis of the confusion matrix in Figure 4 reveals a higher misclassification rate between similar respiratory pathologies. Bronchiectasis and Pneumonia, for example, were frequently misclassified as COPD, reflecting the similarity and overlap in their acoustic characteristics, which make accurate discrimination more difficult when patient-level isolation is applied. This highlights the importance of preventing data leakage and implementing robust validation methodologies for accurate performance evaluation, leading to a more reliable and clinically relevant approach.



	Asthma	Bronchiectasis	COPD	Healthy	Heart Failure	Pneumonia		
Asthma	34	5	39	12		2	37.0%	63.0%
Bronchiectasis	3	26	78	5	1	9	21.3%	78.7%
COPD	16	99	546	68	6	67	68.1%	31.9%
Healthy	6	8	39	190	1	1	77.6%	22.4%
Heart Failure	2	3	15	7	27	2	48.2%	51.8%
Pneumonia	4	8	103	6		45	27.1%	72.9%
	52.3%	17.4%	66.6%	66.0%	77.1%	35.7%		
	47.7%	82.6%	33.4%	34.0%	22.9%	64.3%		

Figure 4: Second phase results confusion matrix.

4 DISCUSSION AND CONCLUSION

This study presented a reproducibility analysis of a published work on automated respiratory disease classification using lung sound recordings ("Recognition of pulmonary diseases from lung sounds using convolutional neural networks and long short-term memory"). By replicating the original methodology and addressing its methodological shortcomings, we demonstrated the substantial impact that data leakage and patient overlap between training and testing sets can have on model performance, systematically leading to overly optimistic results. When the same patients were present in both subsets, the model achieved higher performance values, consistent with the results reported in the original paper.

Table 2: Performance results of the first phase (considering 10-fold cross-validation without patient-isolated folds - data leakage).

Class	Recall (%)	Precision (%)	Specificity (%)	F1 Score (%)	MCC	Support
Asthma	87.9	83.3	98.4	85.5	0.820	91
Bronchiectasis	15.0	81.8	98.2	25.3	0.230	120
COPD	96.3	77.5	75.6	85.7	0.763	804
Heart Failure (HF)	78.5	89.8	98.5	83.7	0.818	56
Healthy/Normal	98.3	92.0	98.9	95.0	0.915	244
Pneumonia	23.2	70.9	97.6	34.8	0.331	168
Macro Avg	66.5	82.9	94.5	68.4	0.646	-
Global	Accuracy: 83.5%					

Table 3: Performance results of the second phase (considering 10-fold cross-validation with patient-isolated folds - no data leakage).

Class	Recall (%)	Precision (%)	Specificity (%)	F1 Score (%)	MCC	Support
Asthma	37.0	52.3	96.8	43.2	0.368	92
Bronchiectasis	21.3	17.4	95.4	19.1	0.148	122
COPD	68.1	66.6	69.3	67.3	0.578	802
Heart Failure (HF)	48.2	77.1	97.4	59.0	0.539	56
Healthy/Normal	77.6	66.0	93.7	71.4	0.633	245
Pneumonia	27.1	35.7	93.4	30.7	0.228	166
Macro Avg	46.6	52.5	90.9	48.5	0.416	-
Global	Accuracy: 64.5%					

However, when a patient-independent data split was enforced, performance dropped markedly across all evaluation metrics, indicating that the earlier results were overoptimistic and lacked generalization.

These findings emphasize the critical importance of rigorous evaluation protocols in biomedical signal analysis. Ensuring strict patient-level separation, transparent preprocessing descriptions, and standardized validation strategies are essential for developing clinically reliable and generalizable AI models. Furthermore, this case study reinforces the need to openly share datasets, code, and experimental configurations, enabling the research community to validate, compare, and improve upon existing methods.

When reviewing the broader literature on automated diagnosis of respiratory diseases, a pattern similar to that observed in the article replicated in the current study emerges: numerous studies report near-perfect diagnostic accuracy, yet none of these approaches have achieved clinical deployment or widespread use in real-world screening or monitoring contexts. This disconnect raises essential questions about the true generalization ability of existing models and the potential presence of dataset biases — such as differences in recording conditions, distribution of recordings per respiratory condition, microphone types, or demographic distributions. It might also suggest that models may inadvertently learn such patterns/biases instead of genuine pathological features. Consequently, it is crucial to adopt more rigorous validation methodologies, ensure transparent reporting of experimental configurations, and leverage diverse and heterogeneous datasets to assess the robustness of these systems beyond a single benchmark. For instance, one such approach could be to cross-validate the developed classification model on different external databases, such as the BRACETS and RespiratoryDatabase@TR (COPD Severity Analysis) databases (Pessoa et al., 2023; Altan et al., 2017).

Future work in the area should focus on promoting reproducible research practices and developing standardized benchmarks for respiratory sound analysis. By addressing methodological flaws and promoting transparency, the field can advance toward robust, interpretable, and clinically translatable automated systems that assist healthcare professionals in accurately diagnosing and monitoring respiratory diseases. Moreover, future work should also focus on deploying these methods in clinical pilots or settings to better understand their potential for monitoring and diagnosing respiratory diseases.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon Europe Research and Innovation

Programme under Grant Agreement No. 101226783 (MULTIPULM), and national funds from FCT - Foundation for Science and Technology, I.P., within the scope of project CISUC - UID/CEC/00326/2020 with funds from the European Social Fund, through the Regional Operational Program Centro 2020.

REFERENCES

Altan, G., Kutlu, Y., Garbİ, Y., Pekmezcİ, A. Ö., and NURAL, S. (2017). Multimedia Respiratory Database (RespiratoryDatabase@TR): Auscultation Sounds and Chest X-rays. *Natural and Engineering Sciences*, 2(3):59–72.

Fraiwan, M., Fraiwan, L., Alkhodari, M., and Hassanin, O. (2021a). Recognition of pulmonary diseases from lung sounds using convolutional neural networks and long short-term memory. *Journal of Ambient Intelligence and Humanized Computing 2021 13:10*, 13(10):4759–4771.

Fraiwan, M., Fraiwan, L., Khassawneh, B., and Ibnian, A. (2021b). A dataset of lung sounds recorded from the chest wall using an electronic stethoscope. *Data in Brief*, 35:106913.

García-Ordás, M. T., Benítez-Andrades, J. A., García-Rodríguez, I., Benavides, C., and Alaiz-Moretón, H. (2020). Detecting Respiratory Pathologies Using Convolutional Neural Networks and Variational Autoencoders for Unbalancing Data. *Sensors 2020, Vol. 20, Page 1214*, 20(4):1214.

Marques, A., Oliveira, A., and Jácome, C. (2013). Computerized Adventitious Respiratory Sounds as Outcome Measures for Respiratory Therapy: A Systematic Review. *https://home.liebertpub.com/rcare*, 59(5):765–776.

Pessoa, D., Machado Rocha, B., de Carvalho, P., and Paiva, R. P. (2022). Automated respiratory sound analysis. *Wearable Sensing and Intelligent Data Analysis for Respiratory Management*, pages 123–168.

Pessoa, D., Rocha, B. M., Strodthoff, C., Gomes, M., Rodrigues, G., Petmezas, G., Cheimariotis, G. A., Kilintzis, V., Kaimakamis, E., Maglaveras, N., Marques, A., Frerichs, I., Carvalho, P. d., and Paiva, R. P. (2023). BRACETS: Bimodal repository of auscultation coupled with electrical impedance thoracic signals. *Computer Methods and Programs in Biomedicine*, 240:107720.

Pessoa, D. R. M. (2025). Automatic Respiratory Function Analysis using Respiratory Sound and Electrical Impedance Tomography. *Automatic Respiratory Function Analysis using Respiratory Sound and Electrical Impedance Tomography*.

Rocha, B. M., Filos, D., Mendes, L., Serbes, G., Ulukaya, S., Kahya, Y. P., Jakovljevic, N., Turukalo, T. L., Vogiatzis, I. M., Perantoni, E., Kaimakamis, E., Natsiavas, P., Oliveira, A., Jácome, C., Marques, A., Maglaveras, N., Pedro Paiva, R., Chouvarda, I., and De Carvalho, P. (2019). An open access database for the evaluation of respiratory sound classification algorithms. *Physiological Measurement*, 40(3):035001.

Rocha, B. M., Filos, D., Mendes, L., Vogiatzis, I., Perantoni, E., Kaimakamis, E., Natsiavas, P., Oliveira, A., Jácome, C., Marques, A., Paiva, R. P., Chouvarda, I., Carvalho, P., and Maglaveras, N. (2018). A Respiratory Sound Database for the Development of Automated Classification. In *IFMBE Proceedings*, vol. 66, pages 33–37. Springer, Singapore.

World Health Organization (WHO) (2024). The top 10 causes of death.

Xia, T., Han, J., and Mascolo, C. (2022). Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues. *Experimental Biology and Medicine*, 247(22):2053–2061.