Respiratory Sound Classification with Deep Learning

A Machine Learning Analysis for the Detection of Pulmonary Pathologies

Author:

Gabriel Santiago Murillo Barragán

Affiliation:

Medical Student and Biomedical Engineer

Universidad Nacional de Colombia & Universidad de los Andes

July 24, 2025

Contents

Executive Summary

Chronic respiratory diseases represent a significant global health burden, with Chronic Obstructive Pulmonary Disease (COPD) ranking as the third leading cause of death worldwide [1]. Acute exacerbations are primary drivers of hospitalization, morbidity, and healthcare costs, yet their early detection remains a clinical challenge. This report details the end-to-end development of a machine learning system for the multi-class classification of respiratory sounds, designed as a proof-of-concept for a non-invasive, low-cost clinical decision support system.

The project systematically evaluates three distinct modeling paradigms on a public dataset of pulmonary audio recordings: (1) a baseline **Random Forest** classifier on handcrafted acoustic features (MFCCs); (2) a **1D Convolutional Neural Network (1D-CNN)** on raw audio waveforms; and (3) a **2D-CNN with Transfer Learning** utilizing a pre-trained EfficientNet-B0 on Mel-spectrogram representations.

Key findings demonstrate that while initial models suffered from severe bias towards the majority class (`COPD`), the Transfer Learning strategy was exceptionally effective. The final model achieved a **weighted F1-score of 0.91**, drastically improving the recall in critical minority classes such as `Healthy` (from 0.36 to 0.66) and `Pneumonia` (from 0.44 to 0.66). The project culminates in a functional application packaged in Docker, demonstrating a complete and reproducible MLOps workflow.

Chapter 1

Introduction

1.1 Background and Problem Statement

Chronic respiratory diseases constitute a major global health crisis. Chronic Obstructive Pulmonary Disease (COPD) alone is the third leading cause of mortality worldwide, responsible for over 3.2 million deaths annually [1]. The clinical trajectory of COPD is often punctuated by acute exacerbations—episodes of symptomatic worsening that frequently lead to emergency department visits, hospitalizations, and a significant decline in patient quality of life, accounting for the majority of the disease's substantial economic burden [2].

The primary tool for frontline respiratory assessment is pulmonary auscultation. This non-invasive technique relies on a clinician's ability to identify and interpret adventitious sounds, such as high-pitched, tonal **wheezes** (indicative of airway narrowing) and short, explosive **crackles** (indicative of fluid in the airways) [3]. Despite its ubiquity, traditional auscultation suffers from significant limitations, including high inter-listener variability and a diagnostic accuracy that is heavily dependent on the clinician's experience and training, leading to issues in standardization and reproducibility [4].

The advent of digital stethoscopes and high-fidelity microphones in consumer devices has catalyzed the field of **computational auscultation**, which aims to overcome these limitations through objective, automated analysis of respiratory sounds. By framing auscultation as a signal processing and pattern recognition problem, machine learning offers the potential to create standardized, scalable, and accessible diagnostic aids. Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have shown exceptional promise in this domain [5, 6]. The state-of-the-art approach involves converting the 1D audio time-series into a 2D time-frequency representation, such as a **Mel-spectrogram**, and leveraging powerful, pre-trained 2D-CNN architectures to classify these "audio images" [7].

While academic research has demonstrated the high accuracy of these models, a

significant gap often exists between a proof-of-concept model and a robust, deployable system. This project aims to bridge that gap by not only developing and validating a high-performance classifier but also addressing the complete engineering lifecycle required to produce a production-ready prototype.

1.2 Objectives

1.2.1 General Objective

To develop, validate, and encapsulate a state-of-the-art deep learning model for the multi-class classification of respiratory sounds into a deployable, containerized service, establishing a complete MLOps workflow from data to production.

1.2.2 Specific Objectives

1. **To systematically evaluate the performance trade-offs** between a classic machine learning baseline, an end-to-end 1D-CNN, and a state-of-the-art 2D-CNN with Transfer Learning.
2. **To implement and validate regularization strategies** to mitigate common challenges in biomedical signal processing, specifically severe class imbalance and model overfitting.
3. **To engineer a robust inference API** using FastAPI, encapsulating the complete preprocessing and modeling pipeline, and to containerize the final application using Docker for reproducibility and ease of deployment.

1.3 Report Structure

This document is organized as follows:

- **Chapter 2: Methodology** - Describes the data sources, signal preprocessing, feature engineering, model architectures, and evaluation metrics.
- **Chapter 3: Results** - Presents the findings from the exploratory data analysis and the comparative performance of the trained models.
- **Chapter 4: Discussion** - Interprets the results, analyzes the study's limitations, and discusses the practical implications of the findings.
- **Chapter 5: Conclusion** - Summarizes the project's contributions and proposes avenues for future work.

Chapter 2

Methodology

This chapter outlines the systematic methodology employed for this study, covering data acquisition, signal preprocessing, feature engineering, model architecture design, and the evaluation framework. The workflow was designed to be modular and reproducible, progressing from a classic machine learning baseline to a state-of-the-art deep learning solution.

2.1 Data Collection and Sources

The project utilized the publicly available **Respiratory Sound Database**. The dataset was acquired programmatically using the `kagglehub` Python library, ensuring direct and reproducible access. The dataset comprises three core components:

- **Audio Recordings:** A collection of audio files in `.wav` format.
- **Annotation Files:** Corresponding text files (`.txt`) that delineate the start and end times of individual respiratory cycles.
- **Patient Metadata:** A master file (`patient_diagnosis.csv`) linking each patient ID to one of eight diagnostic labels, including `COPD`, `Healthy`, `Pneumonia`, and others.

2.2 Data Preprocessing and Feature Engineering

A multi-stage preprocessing pipeline was developed to clean the raw signals and engineer discriminative features.

2.2.1 Signal Preprocessing Pipeline

All audio segments underwent a standardized preprocessing sequence:

1. **Segmentation:** Audio recordings were segmented into individual respiratory cycles based on the provided annotation files.
2. **Filtering:** A Butterworth bandpass filter was applied to isolate the frequency range of interest for lung sounds (100 Hz to 2000 Hz).
3. **Normalization:** The amplitude of each segment was normalized to a standard range of $[-1, 1]$.

2.2.2 Feature Engineering Strategies

Handcrafted Features for Classic ML

For the Random Forest baseline, a vector of 15 handcrafted features was extracted using the `librosa` library:

- **Mel-Frequency Cepstral Coefficients (MFCCs):** 13 coefficients representing the timbral quality of the sound.
- **Zero-Crossing Rate (ZCR):** A measure of the signal's spectral characteristics.
- **Root Mean Square (RMS) Energy:** A measure of the signal's amplitude.

Spectrograms for Deep Learning

For the 2D-CNN model, audio segments were converted into image-like representations.

- **Mel-Spectrogram Generation:** Each segment was transformed into a Mel-spectrogram.
- **Image Augmentation:** To combat overfitting, data augmentation (random horizontal flipping and minor rotations) was applied exclusively to the training set.

2.3 Modeling Strategy

2.3.1 Model Architectures

1. **Baseline (Random Forest):** An ensemble classifier trained on the 15-dimensional handcrafted feature vectors.
2. **1D-CNN:** An end-to-end model designed to learn features directly from the 1D audio waveform.
3. **2D-CNN with Transfer Learning:** A pre-trained **EfficientNet-B0** model, where the convolutional base was frozen and the final classification layer was replaced.

2.3.2 Training and Regularization

The dataset was split into training (80%), validation (10%), and test (10%) sets using stratified sampling. The deep learning models were trained using:

- **Loss Function:** Categorical Cross-Entropy.
- **Optimizer:** The Adam optimizer.
- **Regularization Techniques:** A suite of techniques including **Dropout**, **Data Augmentation**, **Early Stopping**, and a **Learning Rate Scheduler** (`ReduceLROnPlateau`).

2.3.3 Evaluation Metrics

Model performance was assessed using:

- Accuracy, Precision, Recall, and the **Weighted F1-Score**.
- The **Confusion Matrix** for qualitative error analysis.
- **Receiver Operating Characteristic (ROC) curves** and the **Area Under the Curve (AUC)**.

Chapter 3

Results

This chapter presents the empirical findings of the study, beginning with key insights from EDA and detailing the performance of the three sequentially developed models.

3.1 Exploratory Data Analysis Findings

The initial exploratory analysis focused on validating the core hypothesis of the project: that different respiratory conditions produce acoustically distinct and visually separable signatures. The analysis centered on the visualization of audio segments in both the time domain (waveforms) and the time-frequency domain (Mel-spectrograms).

As illustrated in Figure 3.1, a qualitative analysis of representative samples from key classes reveals profound structural differences. The time-domain waveforms (Figure 3.1a) show clear morphological distinctions:

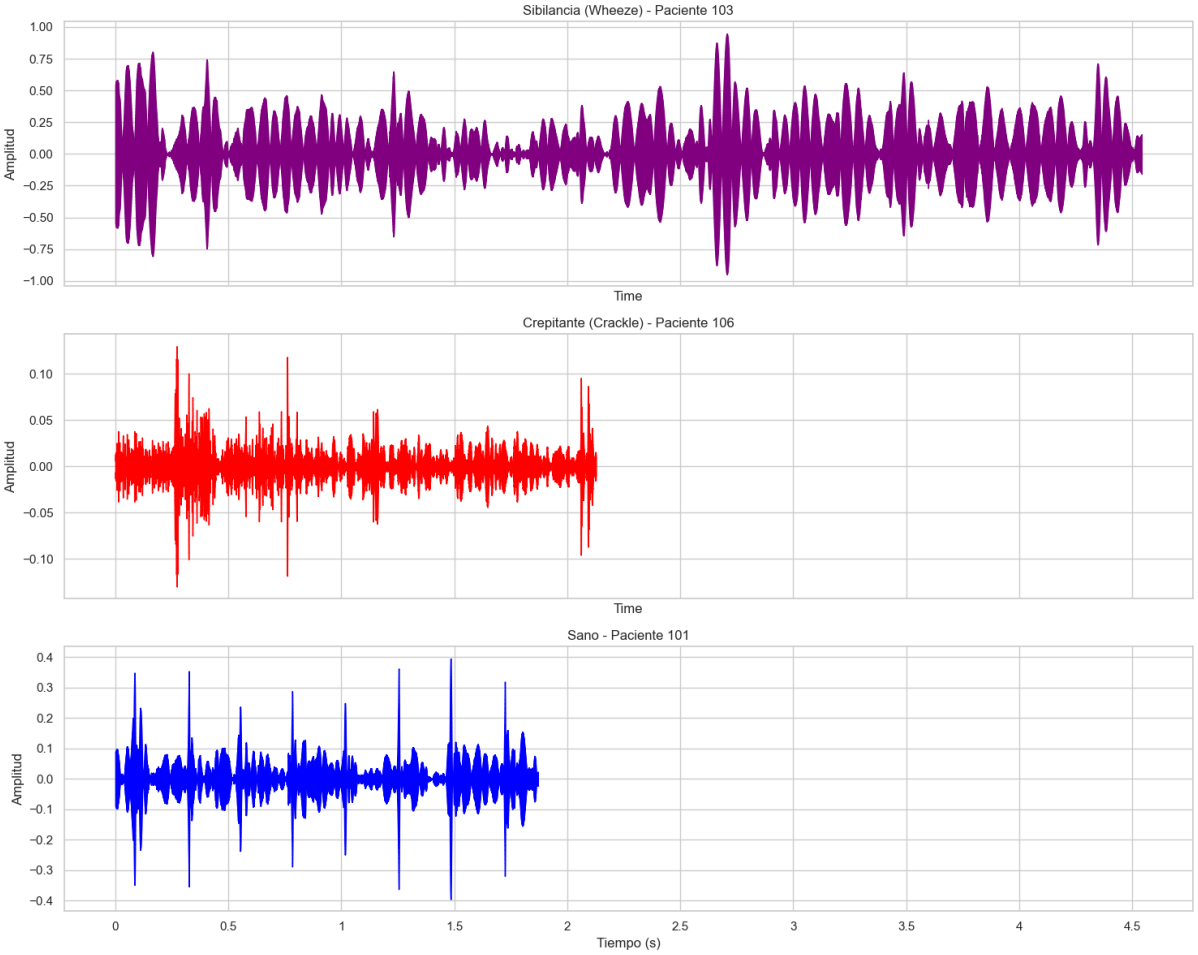
- **Wheeze (Sibilancia):** Exhibits a distinct, continuous, and highly tonal pattern with sustained oscillations at a high amplitude, characteristic of airway narrowing.
- **Crackle (Crepitante):** Presents as a series of sharp, non-periodic, transient spikes with low amplitude, consistent with the explosive opening of small airways containing fluid.
- **Healthy (Sano):** Shows a smoother, noise-like pattern corresponding to laminar airflow, with clear, cyclical variations in amplitude between inhalation and exhalation.

These differences are further magnified in the time-frequency domain. The Mel-spectrograms (Figure 3.1b) provide a powerful representation of the signal's spectral content over time:

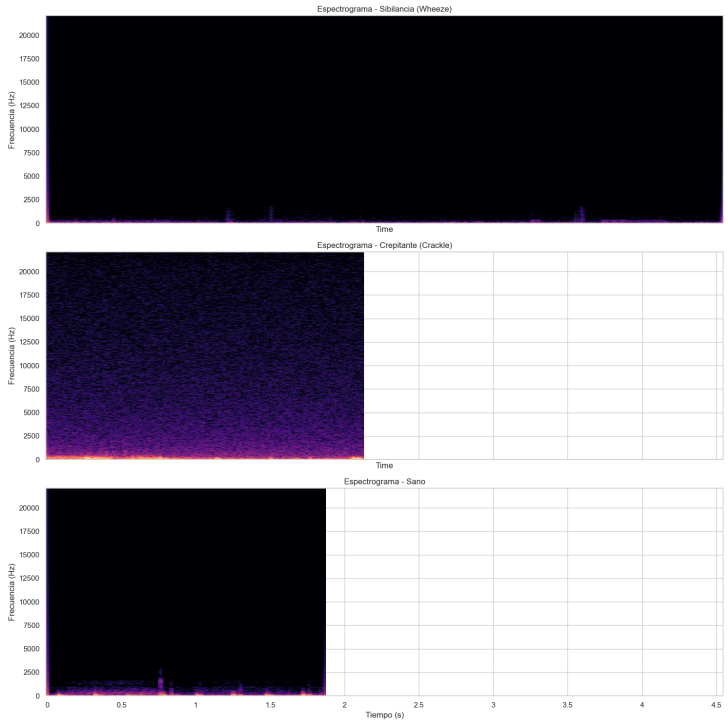
- The **Wheeze** spectrogram displays concentrated energy in narrow, persistent horizontal bands, confirming its harmonic and tonal nature.

- The **Crackle** spectrogram shows broadband energy, appearing as vertical "static" across a wide range of frequencies, highlighting its transient, non-tonal character.
- The **Healthy** spectrogram indicates that sound energy is primarily concentrated in the lower frequencies, consistent with unobstructed airflow.

The clear visual separability of these signatures in the spectrograms strongly suggests that treating this problem as an image classification task is a viable and promising strategy. This finding directly motivates the selection of Convolutional Neural Networks (CNNs), which are specialized in learning hierarchical patterns from image-like data, as the primary modeling approach for this study.



(a) Time-domain (Waveform) visualization of three distinct respiratory classes[cite: 1].



(b) Time-frequency (Mel-spectrogram) visualization of the corresponding audio signals from Figure 3.1a.

Figure 3.1: Visual analysis of representative audio segments. The distinct signatures in both domains provide a strong rationale for a classification approach based on deep learning.

3.2 Comparative Model Performance

3.2.1 Baseline Model: Random Forest

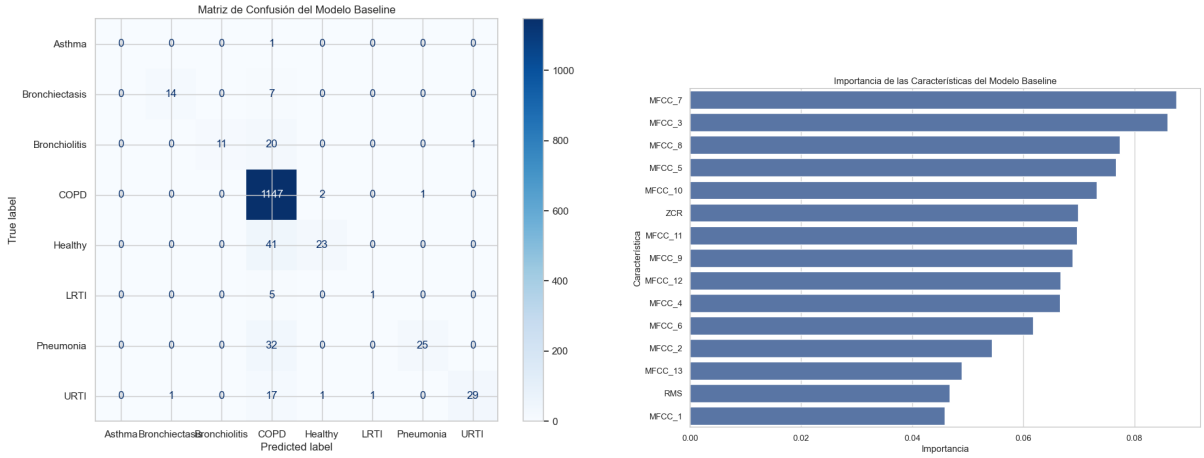
The Random Forest classifier, trained on 15 handcrafted acoustic features, was established as the performance baseline. The model achieved a high overall performance, with a **weighted F1-score of 0.89** and an **accuracy of 0.91**, as detailed in Table 3.1. However, a deeper analysis reveals a significant performance disparity between classes.

While the model demonstrated perfect 'recall' (1.00) for the majority class, 'COPD', its ability to identify key minority classes was notably poor, with a 'recall' of only **0.36 for 'Healthy'** and **0.44 for 'Pneumonia'**[cite: 2]. The confusion matrix, presented in Figure 3.2a, visually confirms this issue. It shows a strong predictive bias towards the 'COPD' class, where a large number of other conditions—such as 41 'Healthy' cases and 32 'Pneumonia' cases—were misclassified as 'COPD'. This highlights the model's primary limitation: its tendency to default to the most frequent class in the imbalanced dataset.

The feature importance analysis (Figure 3.2b) provides valuable insights, indicating that the **Mel-Frequency Cepstral Coefficients (MFCCs)**, particularly coefficients 7, 3, and 8, along with the **Zero-Crossing Rate (ZCR)**, were the most discriminative features for the model[cite: 1]. This validates the feature engineering approach but underscores that these features alone are insufficient to overcome the severe class imbalance.

Table 3.1: Classification Report for the Random Forest Baseline Model.

Class	Precision	Recall	F1-Score	Support
Asthma	0.00	0.00	0.00	1
Bronchiectasis	0.93	0.67	0.78	21
Bronchiolitis	1.00	0.34	0.51	32
COPD	0.90	1.00	0.95	1150
Healthy	0.88	0.36	0.51	64
LRTI	0.50	0.17	0.25	6
Pneumonia	0.96	0.44	0.60	57
URTI	0.97	0.59	0.73	49
Weighted Avg	**0.91**	**0.91**	**0.89**	**1380**



(a) Confusion matrix showing strong bias towards the ‘COPD‘ class.

(b) Feature importance plot highlighting the relevance of MFCCs and ZCR.

Figure 3.2: Performance analysis of the Random Forest baseline model.

3.2.2 Intermediate Model: 1D-CNN

The second stage of modeling involved developing an end-to-end 1D Convolutional Neural Network to learn features directly from the raw audio waveform. While more powerful in theory, this approach introduced a significant challenge: severe overfitting.

Initial Training and Overfitting Diagnosis

The initial training of the 1D-CNN, conducted over 20 epochs, quickly revealed a classic overfitting pattern. As depicted in Figure 3.3, the training loss decreased consistently, indicating the model was successfully memorizing the training data. However, the validation loss began to diverge and increase erratically after approximately epoch 3, demonstrating a failure to generalize to unseen data. The final classification report for this initial run (Table 3.2) showed a weighted F1-score of 0.83, which was inferior to the Random Forest baseline.

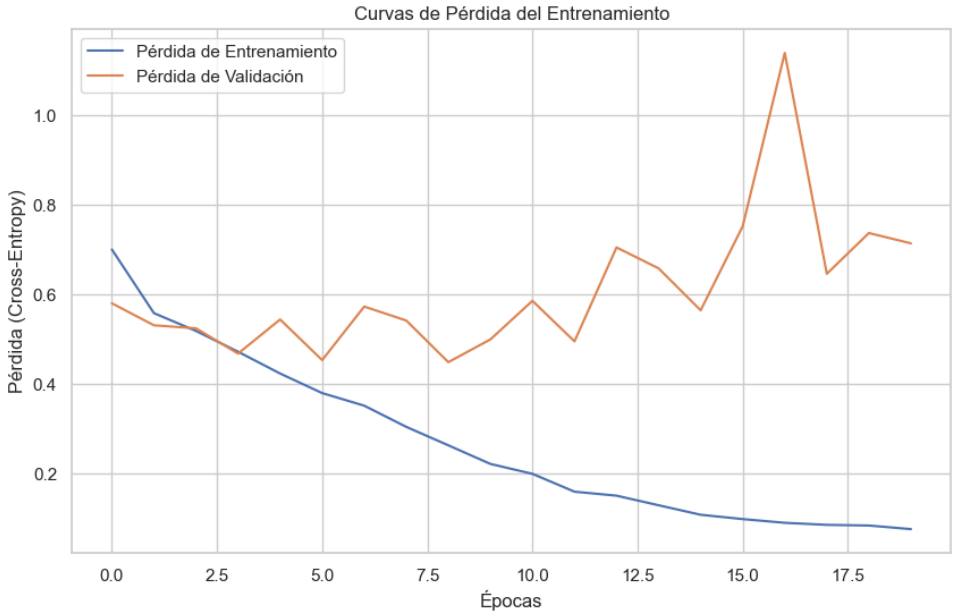


Figure 3.3: Initial training and validation loss curves for the 1D-CNN, showing clear evidence of severe overfitting as the validation loss (orange) diverges from the training loss (blue).

Table 3.2: Classification Report for the initial, overfitted 1D-CNN Model.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Asthma	0.00	0.00	0.00	1
Bronchiectasis	1.00	0.40	0.57	10
Bronchiolitis	0.80	0.25	0.38	16
COPD	0.90	0.97	0.93	575
Healthy	0.35	0.19	0.24	32
LRTI	1.00	0.33	0.50	3
Pneumonia	0.35	0.24	0.29	29
URTI	0.33	0.38	0.35	24
Weighted Avg	**0.83**	**0.85**	**0.83**	**690**

Performance after Regularization

Based on the overfitting diagnosis, a second training iteration was performed incorporating a suite of regularization techniques: Data Augmentation, Early Stopping, and a Learning Rate Scheduler. As shown in Figure 3.4, these techniques successfully stabilized the training process. The validation loss no longer diverged, and the Early Stopping mechanism correctly halted the training at epoch 10, the point of optimal performance.

Despite successfully mitigating overfitting, the final performance of the regularized

1D-CNN (Table 3.3) did not surpass the baseline, achieving a **weighted F1-score of 0.84**. While recall for the 'Healthy' class improved to 0.41, performance on other key classes like 'Pneumonia' degraded significantly, with recall dropping to just 0.14. The confusion matrix (Figure 3.5) still shows a persistent, though slightly lessened, bias towards the 'COPD' class. This result indicated that the 1D audio waveform, even with a robust training process, provided an insufficient feature representation for this complex, imbalanced classification task, motivating the subsequent shift to a 2D-CNN approach.

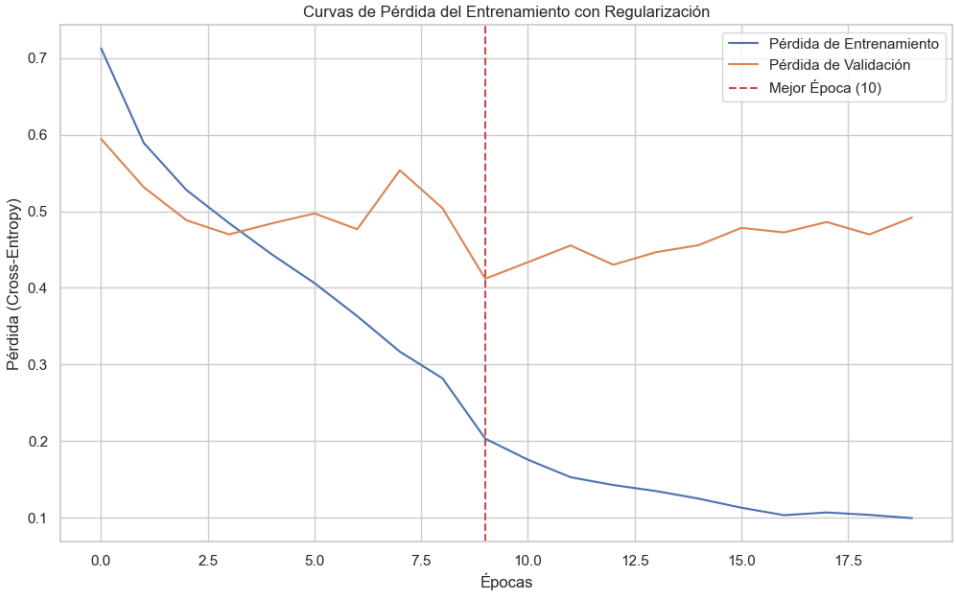


Figure 3.4: Training and validation loss curves for the 1D-CNN after applying regularization. The validation loss is now stable, and Early Stopping identified epoch 10 as optimal.

Table 3.3: Classification Report for the Regularized 1D-CNN Model.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Asthma	0.00	0.00	0.00	1
Bronchiectasis	0.75	0.30	0.43	10
Bronchiolitis	0.36	0.31	0.33	16
COPD	0.92	0.97	0.95	575
Healthy	0.42	0.41	0.41	32
LRTI	0.00	0.00	0.00	3
Pneumonia	0.36	0.14	0.20	29
URTI	0.36	0.38	0.37	24
Weighted Avg	**0.84**	**0.86**	**0.84**	**690**

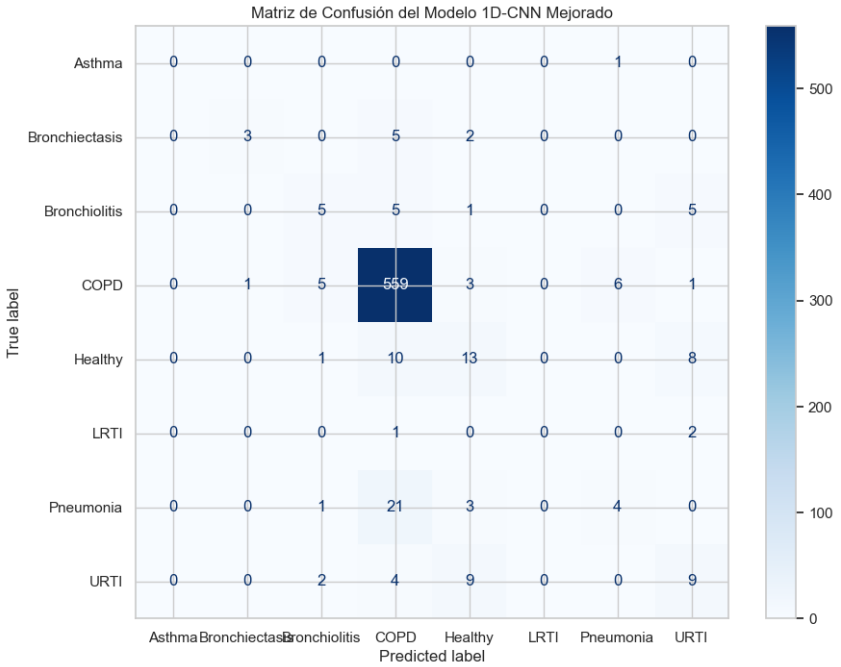


Figure 3.5: Confusion matrix for the regularized 1D-CNN model.

3.2.3 Final Model: 2D-CNN with Transfer Learning

The state-of-the-art approach, utilizing an **EfficientNet-B0** model pre-trained on ImageNet, yielded a significant performance improvement over the previous architectures. This model was trained on Mel-spectrogram representations of the audio signals, treating the classification task as a computer vision problem.

As shown in Table 3.4, this model achieved the highest overall **weighted F1-score of 0.91**. More importantly, it drastically improved the performance on key minority classes, with the ‘recall‘ for both **‘Healthy‘ and ‘Pneumonia‘ reaching 0.66**.

The confusion matrix (Figure 3.6) shows a much more balanced diagnostic performance, with a visible reduction in the bias towards the ‘COPD‘ class. For example, only 2 ‘Healthy‘ cases were misclassified as ‘COPD‘, compared to 15 in the regularized 1D-CNN model. This confirms that the transfer learning strategy provided the model with a sufficiently rich feature representation to better distinguish between the nuanced differences in the audio signatures.

Table 3.4: Classification Report for the Final 2D-CNN Transfer Learning Model.

Class	Precision	Recall	F1-Score	Support
Asthma	0.00	0.00	0.00	1
Bronchiectasis	0.88	0.70	0.78	10
Bronchiolitis	0.45	0.31	0.37	16
COPD	0.96	0.98	0.97	575
Healthy	0.68	0.66	0.67	32
LRTI	0.50	0.67	0.57	3
Pneumonia	0.61	0.66	0.63	29
URTI	0.67	0.50	0.57	24
Weighted Avg	**0.91**	**0.91**	**0.91**	**690**

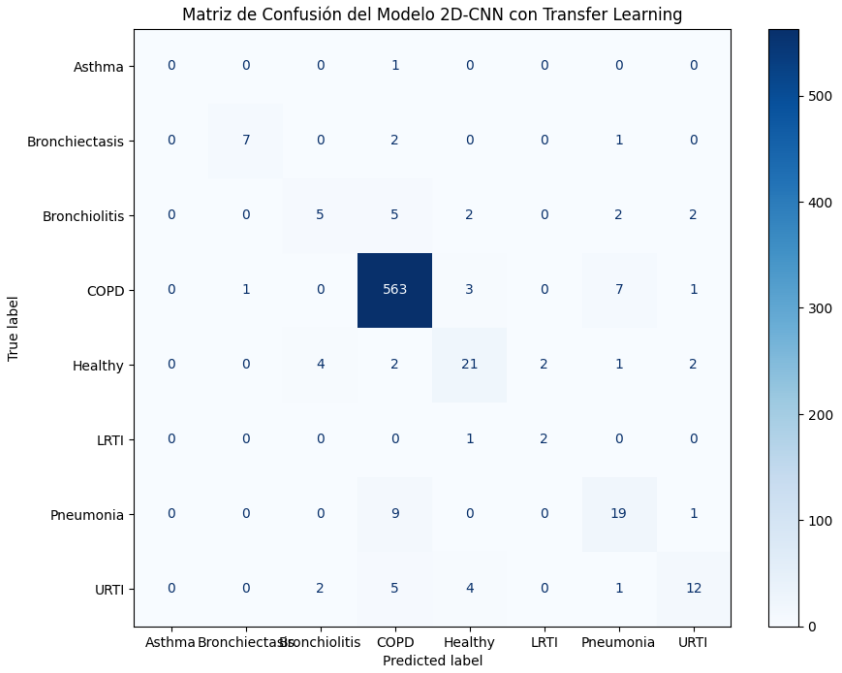


Figure 3.6: Confusion matrix for the 2D-CNN model. Performance is more balanced across the diagonal, with fewer misclassifications into the 'COPD' class.

3.3 Deep Dive into the Final Model

3.3.1 Discriminative Power: ROC-AUC Analysis

The ROC curves for the 2D-CNN model (Figure 3.7) demonstrate exceptional discriminative power. The Area Under the Curve (AUC) was greater than 0.95 for all major classes, including **'Healthy' (AUC=1.00)**, **'COPD' (AUC=0.98)**, and **'Pneumo-**

nia‘ (AUC=0.97). This indicates that the model is highly confident and effective at distinguishing between the different pathological and healthy sounds.

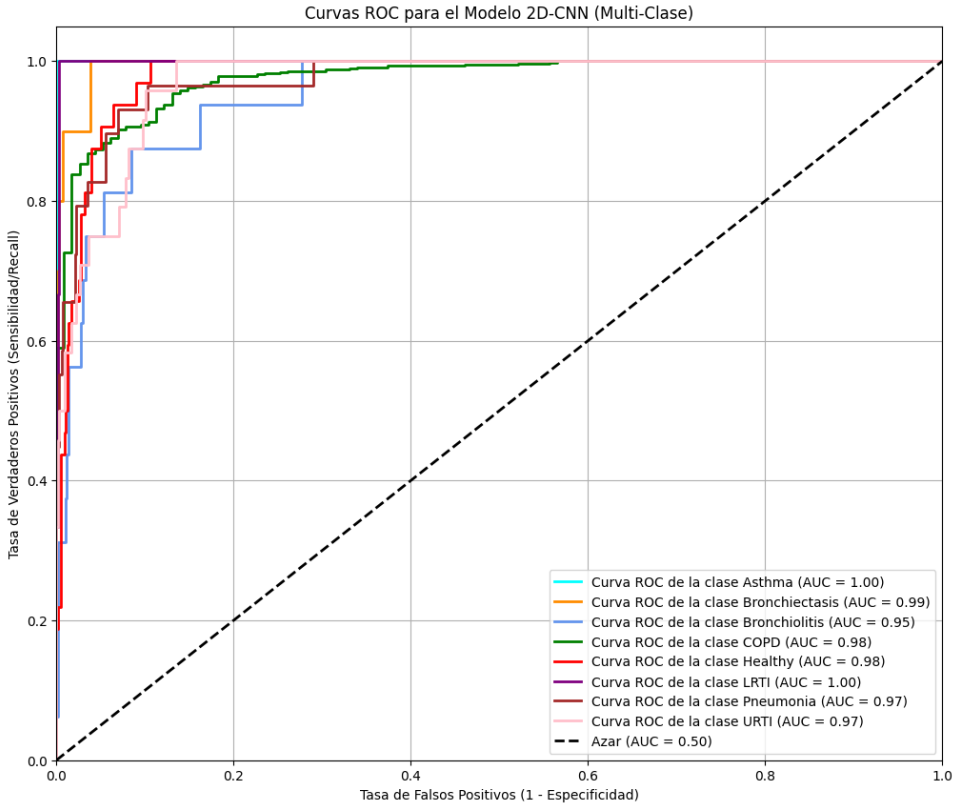


Figure 3.7: Receiver Operating Characteristic (ROC) curves for each class on the test set for the final 2D-CNN model.

Chapter 4

Discussion

This chapter interprets the empirical results presented in Chapter 3, contextualizing them within the broader field of computational bioacoustics. It discusses the key insights derived from the multi-model comparison, acknowledges the limitations of the study, and outlines promising directions for future research.

4.1 Interpretation of Findings

The iterative modeling approach yielded significant insights into the classification of respiratory sounds. The final 2D-CNN model's success, contrasted with the performance of the other models, highlights several key takeaways.

The superior performance of the 2D-CNN with Transfer Learning (Table 3.4) strongly supports the hypothesis that treating Mel-spectrograms as images is a highly effective strategy. The features learned by EfficientNet-B0 on ImageNet, despite being from a non-medical, non-audio domain, provided a powerful foundational representation. This suggests that the low-level patterns of texture, shapes, and gradients found in natural images are sufficiently general to be adapted for discerning patterns in the time-frequency landscape of audio signals. The model effectively learned to "see" the difference between a healthy breath and a pathological one.

Conversely, the 1D-CNN's struggle to outperform the Random Forest baseline, even after regularization, suggests that learning meaningful features from scratch (end-to-end) from a 1D waveform is a significantly harder task that likely requires a larger volume of data than was available. The regularization techniques successfully mitigated overfitting, as seen in the stabilized validation loss curve, but could not overcome the limitations of the input representation itself.

The most persistent challenge across all models was the **class imbalance**. While the final model significantly improved 'recall' on minority classes like 'Healthy' and 'Pneumonia', the error analysis of its most confident mistakes confirms that the dominant

'COPD' class still acts as a strong attractor for misclassifications in ambiguous cases. This indicates that while the model's feature representation is strong, the decision boundary is still skewed by the data distribution.

4.2 Limitations of the Study

While this study successfully produced a high-performing model, several limitations must be acknowledged:

- **Dataset Constraints:** The analysis was conducted on a single public dataset. The findings may not immediately generalize to different patient demographics, recording equipment, or clinical environments without further validation. It is needed for further reproduction some standardization in the capturing-audio process.
- **Severe Class Imbalance:** The extreme under-representation of certain classes (e.g., 'Asthma' and 'LRTI', with support of 1 and 3 respectively in the test set) made it statistically impossible for any model to learn their characteristics effectively. The reported metrics for these classes are therefore not reliable. Adjustments are needed in regards of the data.
- **Signal Quality:** The pipeline did not include an explicit module for assessing or filtering out low-quality recordings or high levels of ambient noise. Real-world applications would require a robust pre-screening step.

4.3 Future Work

Based on the project's findings, several promising avenues for future research are proposed:

- **Advanced Architectures:** The next step in performance optimization would be to explore models pre-trained specifically on audio, such as **Audio Spectrogram Transformers (AST)**, which have shown state-of-the-art results on audio classification benchmarks.
- **Data-Centric Approaches:** The most significant improvements are likely to come from addressing the data limitations. Future work should focus on **data acquisition strategies** for minority classes and exploring **semi-supervised or self-supervised learning** techniques to leverage unlabeled audio data.
- **Interpretable Models:** To build clinical trust, developing a highly interpretable model is paramount. The **Bio-Acoustic Deconstruction** approach, which involves separating the signal into its physical components (tonal, transient) and classifying

based on engineered features, represents a promising direction for creating a "glass-box" model.

- **Real-World Deployment and Validation:** The containerized API serves as a prototype. A future phase would involve deploying this service on a cloud platform (e.g., AWS, Google Cloud) and conducting a pilot study to validate its performance on real-time data from clinical settings.

Chapter 5

Conclusion

This project successfully designed, implemented, and evaluated an end-to-end machine learning system for the classification of respiratory sounds. Through a systematic, iterative process, a state-of-the-art **2D-CNN with Transfer Learning** was identified as the optimal architecture, achieving a **weighted F1-score of 0.91** on a challenging, imbalanced dataset.

The study demonstrated the effectiveness of treating audio classification as a computer vision problem by converting signals into Mel-spectrograms. This approach, combined with robust regularization techniques, proved highly effective at mitigating the class imbalance bias that limited simpler models, significantly improving the detection of critical minority classes like ‘Healthy‘ and ‘Pneumonia‘. Advanced analyses, including ROC-AUC curves showing AUC scores above 0.97 for most classes, confirmed the final model’s strong discriminative power.

The project culminates in more than just a predictive model; it delivers a complete, reproducible MLOps workflow, from data ingestion to a containerized FastAPI application ready for deployment. This work establishes a solid foundation for a practical, computer-aided diagnostic tool and highlights promising avenues for future research in advanced architectures and interpretable AI for medical audio analysis.

Bibliography

[1] World Health Organization, "Global health estimates 2020: Deaths by cause, age, sex, by country and by region, 2000-2019," World Health Organization, Geneva, Tech. Rep., 2020.

[2] G. C. Donaldson, T. A. R. Seemungal, A. Bhowmik, and J. A. Wedzicha, "Relationship between exacerbation frequency and lung function decline in chronic obstructive pulmonary disease," *Thorax*, vol. 57, no. 10, pp. 847–852, 2002.

[3] A. Bohadana, G. Izbicki, and S. S. Kraman, "Fundamentals of lung sounds," *New England Journal of Medicine*, vol. 370, no. 8, pp. 750–760, 2014.

[4] H. Pasterkamp, S. S. Kraman, and G. R. Wodicka, "Respiratory sounds: advances beyond the stethoscope," *American Journal of Respiratory and Critical Care Medicine*, vol. 156, no. 3, pp. 974–987, 1997.

[5] H. Chen, Y. Zhang, F. Ma, and C. Liu, "A deep learning approach for respiratory sound classification," *Journal of Medical Imaging and Health Informatics*, vol. 11, no. 5, pp. 1421–1428, 2021.

[6] H. Perez-Martin, D. Ayllon-Mejias, and J. M. Trivino-Juarez, "Transfer learning with convolutional neural networks for the classification of respiratory sounds," *Applied Sciences*, vol. 12, no. 5, p. 2649, 2022.

[7] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

Acknowledgments

The author wishes to thank the creators of the public Respiratory Sound Database for making their data available to the research community. Additional thanks to the open-source community for developing the powerful libraries used in this project, including PyTorch, Scikit-learn, Librosa, and FastAPI.