UNIVERSIDAD DE JAÉN

CENTRO
DEPARTAMENTO

TESIS DOCTORAL

•

Research and development of signal processing and artificial intelligence techniques applied to the retrieval of biomedical information through the analysis of respiratory sound signals

-

Investigación y desarrollo de técnicas de procesado de señal e inteligencia artificial aplicadas a la recuperación de información biomédica a partir del análisis de señales sonoras respiratorias

PRESENTADA POR:
Loredana Daria Mang

DIRIGIDA POR:
Francisco Jesús Cañadas Quesada
Julio José Carabias Orti

Linares, Enero-2024

ISBN

This doctoral thesis has been carried out under the supervision of

Dr. Francisco Jesús Cañadas Quesada

Departamento de Ingeniería de Telecomunicación

Escuela Politécnica Superior de Linares

Universidad de Jaén

Linares, Jaén, España

Dr. Julio José Carabias Orti

Departamento de Ingeniería de Telecomunicación

Escuela Politécnica Superior de Linares

Universidad de Jaén

Linares, Jaén, España

ACKNOWLEDGEMENT

This doctoral thesis, carried out in the Department of Telecommunication Engineering at the University of Jaén and focused on the design of algorithms for the processing of wheezing signals to enhance the diagnosis of the human respiratory system, has been made possible thanks to the grant 1257914 funded by Programa Operativo FEDER Andalucia 2014–2020, grant P18-RT-1994 funded by the Ministry of Economy, Knowledge and University, Junta de Andalucía, Spain.

This thesis has been a challenge and a great motivation for me in my professional field. Especially because of the opportunity to make contributions to the field of pulmonology by providing scientific insights that offer an additional source of information to pulmonologists, aiming to enhance the understanding of possible obstructive pulmonary diseases and, consequently, increase the reliability of the initial diagnosis. Furthermore, this thesis marks the beginning of a research line focused on biomedical signal processing with a long path ahead to explore.

The development of this thesis has been made possible thanks to the effort, collaboration, and support of several individuals who have been with me throughout the entire process. Firstly, I would like to express my deepest gratitude to my thesis advisors, **Dr. Francisco Jesús Cañadas Quesada** and **Dr. Julio José Carabias Orti**, and my mentor, Dr. Pedro Vera Candeas. Thanks to them, I was able to embark on my journey in the world of research. They began as my professors in various courses during my undergraduate and master's studies at the School of Telecommunications Engineering in Linares. I want to thank them for becoming my mentors during this academic and professional phase of my life and for the enormous ammount of time that they invested in this work and especialy in me. Furthermore, I would like to express my gratitude to you for spearheading the research initiative that marked the inception of this doctoral thesis. In essence, it is thanks to your unwavering dedication and the countless hours of work we have put in to bring the outcomes of this thesis to fruition. Thank you for everything, and above all, for the professional and personal support that I have consistently received from you. You have been a source of inspiration for me, and you can always count on my support.

Similarly, I would like to express my gratitude to Dr. Nicolás Ruiz Reyes, the

director of the research group 'Signal Processing in Telecommunication Systems (TIC-188),' for the assistance provided in signal processing matters and scientific discussions.

Place: Linares, Jaén (Spain)

Date: 10/January/2024 **Loredana Daria Mang**

RESUMEN

A nivel global, las enfermedades pulmonares obstructivas plantean un desafío creciente para la salud pública debido a su alta prevalencia, morbimortalidad significativa y costos socioeconómicos sustanciales. Actualmente, el proceso de auscultación sigue siendo el procedimiento clínico principal que emplean los neumólogos para evaluar el estado del sistema respiratorio. Esto se debe a que es un método no invasivo, económico, fácil de usar y especialmente seguro para los pacientes. Sin embargo, el diagnóstico derivado de la auscultación sigue siendo altamente subjetivo y depende de la habilidad, experiencia y formación de cada médico en la escucha e interpretación de señales de audio respiratorias. Como resultado, se producen un alto número de diagnósticos erróneos, lo que pone en riesgo la salud de los pacientes y aumenta los costos asociados con las instituciones de atención médica. Una de las principales tareas que realiza un médico durante la auscultación es la detección y análisis de sonidos adventicios que pueden manifestarse durante la respiración. Estos sonidos adventicios son producidos por la obstrucción de las vías respiratorias y son uno de los síntomas más comunes en las enfermedades pulmonares obstructivas. Específicamente, los sonidos sibilantes y los crepitantes se consideran dos de los sonidos adventicios más relevantes, ya que alertan sobre la posible presencia de importantes enfermedades pulmonares obstructivas como el asma, la bronquiolitis, la bronquiectasia o la enfermedad pulmonar obstructiva crónica (EPOC).

En esta investigación, abordamos la necesidad crítica de la detección temprana de enfermedades respiratorias, que sigue siendo una prioridad de salud mundial. La auscultación, una técnica no invasiva, económica y ampliamente utilizada, depende de la experiencia de los médicos para identificar sonidos respiratorios anormales como los crepitantes. Los sonidos escuchados por el profesional médico, mediante un estetoscopio, se pueden grabar mediante el uso de un estetoscopio electronico, proporcionando de esta manera audios viables para su analisis mediante tecnicas de procesado de señal.

Para mejorar este proceso, proponemos un enfoque en tres fases. Primero, un novedoso método que combina características espectrales basadas en autoregresión (AR) y un clasificador de máquina de soporte vectorial (SVM) para detectar eventos de crepitación en señales de sonido respiratorio. Empleamos una etapa de preproce-

samiento para centrarnos en los componentes de señal relevantes y realizar un análisis de señal a corto plazo. Nuestro enfoque es robusto y logra resultados competitivos con una precisión que varía del 80% al 100% en varias relaciones señal-ruido. Al combinar el modelo AR con SVM, ofrecemos una solución efectiva para detectar estos eventos, mejorando la precisión en la detección sobre señales simuladas a partir de su formula matematica. En segundo lugar, a pesar de los avances en el aprendizaje profundo, a menudo se pasa por alto la elección de las representaciones tiempo-frecuencia para los modelos de Redes Neuronales Convolucionales (CNN). En nuestro estudio, introducimos el cocleograma, que modela la selectividad de frecuencia de la cóclea humana, como una representación tiempo-frecuencia superior para la clasificación de sonidos adventicios respiratorios. El cocleograma supera a otros métodos, logrando una precisión promedio del 85.1% en sibilancias y del 73.8% en crepitancias. Esta investigación resalta la precisión y robustez del cocleograma, ofreciendo mejoras significativas en la clasificación basada en CNN. Y finalmente, exploramos el uso de la arquitectura Vision Transformer (ViT) para la clasificación de sonidos respiratorios, centrándonos en la representación de datos de entrada. ViT ha mostrado prometedores resultados en la clasificación de audio al aplicar autoatención a fragmentos de espectrogramas. Nuestra innovación radica en la combinación del cocleograma, que captura características temporales y espectrales únicas de los sonidos adventicios respiratorios con una arquitectura novedosa, el Vision Transfromer (ViT). Uno de los principales desafíos en las líneas de investigación relacionadas con el procesamiento de señales de audio biomédicas es la falta de bases de datos estandarizadas. Por esta razón evaluamos nuestra metodología en el conjunto de datos público de la Conferencia Internacional de Informática Biomédica y Salud (ICBHI), comparando ViT con otros enfoques de CNN utilizando diversas representaciones de datos de entrada. Nuestros resultados demuestran la eficacia de la representación del cocleograma y el potencial de ViT para la clasificación fiable de sonidos respiratorios. Esta investigación contribuye a los esfuerzos en curso para desarrollar técnicas avanzadas de procesamiento de señales y de inteligencia artificial con el objetivo de mejorar significativamente la velocidad y efectividad de la detección de enfermedades respiratorias, abordando así una necesidad crítica en el campo médico.

ABSTRACT

At a global level, obstructive lung diseases pose a growing challenge to public health due to their high prevalence, significant morbidity and mortality, and substantial socioeconomic costs. Currently, the auscultation process remains the primary clinical procedure employed by pulmonologists to assess the respiratory system's condition. This is because it is a non-invasive, cost-effective, easy-to-use method, and especially safe for patients. However, the diagnosis derived from auscultation remains highly subjective and depends on the skill, experience, and training of each physician in listening and interpreting respiratory audio signals. As a result, a high number of misdiagnoses occur, jeopardizing patients' health and increasing costs associated with healthcare institutions. One of the main tasks a physician performs during auscultation is the detection and analysis of adventitious sounds that may manifest during breathing. These adventitious sounds are produced by the obstruction of the airways and are one of the most common symptoms in obstructive lung diseases. Specifically, wheezing and crackles are considered two of the most relevant adventitious sounds, as they alert to the possible presence of significant obstructive lung diseases such as asthma, bronchiolitis, bronchiectasis, or chronic obstructive pulmonary disease (COPD).

In this research, we address the critical need for early detection of respiratory diseases, which remains a global health priority. Auscultation, a non-invasive, cost-effective, and widely used technique, relies on physicians' expertise to identify abnormal respiratory sounds such as crackles. Sounds heard by the medical professional, through a stethoscope, can be recorded using an electronic stethoscope, thus providing viable audio for analysis using signal processing techniques.

To improve this process, we propose a three-phase approach. Firstly, a novel method that combines autoregression (AR)-based spectral features and a Support Vector Machine (SVM) classifier to detect crackle events in respiratory sound signals. We employ a preprocessing stage to focus on relevant signal components and perform short-term signal analysis. Our approach is robust and achieves competitive results with accuracy ranging from 80% to 100% at various signal-to-noise ratios. By combining the AR model with SVM, we offer an effective solution for detecting these events, im-

proving the precision of detection on simulated signals based on their mathematical formula. Secondly, despite advances in deep learning, the choice of time-frequency representations for Convolutional Neural Network (CNN) models is often overlooked. In our study, we introduce the cochleogram, which models the frequency selectivity of the human cochlea, as a superior time-frequency representation for classifying respiratory adventitious sounds. The cochleogram outperforms other methods, achieving an average accuracy of 85.1% in wheezes and 73.8% in crackles. This research highlights the accuracy and robustness of the cochleogram, offering significant improvements in CNN-based classification. Finally, we explore the use of the Vision Transformer (ViT) architecture for respiratory sound classification, focusing on input data representation. ViT has shown promising results in audio classification by applying self-attention to spectrogram patches. Our innovation lies in the combination of the cochleogram, capturing unique temporal and spectral features of respiratory adventitious sounds, with a novel architecture, the Vision Transformer (ViT). One of the main challenges in research related to biomedical audio signal processing is the lack of standardized databases. For this reason, we evaluate our methodology on the International Conference on Biomedical and Health Informatics (ICBHI) public dataset, comparing ViT with other CNN approaches using various input data representations. Our results demonstrate the effectiveness of the cochleogram representation and the potential of ViT for reliable classification of respiratory sounds. This research contributes to ongoing efforts to develop advanced signal processing and artificial intelligence techniques with the aim of significantly improving the speed and effectiveness of respiratory disease detection, addressing a critical need in the medical field.

Keywords: *Adventitious respiratory sounds*, *wheezes*, *crackles*, *auto-regressive (AR) models*, *convolutional neural networks (CNN)*, *cochleogram*, *vision transformer*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Research Proposal and Knowledge Review

The aim of this chapter is to underscore the motivation behind this research line. To achieve this, we will begin by elucidating the current issues surrounding obstructive pulmonary diseases and how the analysis of adventitious sounds (anomalous and indicative of a pulmonary disorder), particularly wheezing and crackles sounds, can aid in enhancing the initial diagnosis resulting from auscultation. Secondly, we will outline and justify the research objectives, taking into consideration the most pertinent tasks and challenges faced by pulmonologists in the analysis of adventitious sounds. Finally, we will provide an overview of the publications included in this doctoral Thesis and introduce the various chapters into which the book is divided.

1.1 Context and motivation of the research

The most important challenge facing researchers in any scientific field is to ensure that the advancements in their research and their scientific contributions enhance the quality of life for citizens. The concept of electronic Health (eHealth) represents this intention in the healthcare sector. The World Health Organization (WHO) [13] defines this concept as the use of Information and Communication Technologies (ICT) for healthcare. On the other hand, the Technology and Health Foundation [14] defines it as the set of ICT used in the healthcare environment for prevention, diagnosis, treatment, monitoring, and health management, serving as a catalyst for change in healthcare systems that enables cost savings and improved efficiency. In this regard, the challenge pursued with this Thesis is to make scientific contributions that, in terms of diagnosis, enhance the efficiency of diagnoses to prevent jeopardizing patients' health, and in terms of healthcare management, reduce the costs associated with healthcare centers resulting from erroneous diagnoses.

Respiratory illnesses are on the rise and currently rank third among the leading causes of mortality, following cardiovascular diseases and cancer. The World Health Organization (WHO) highlights that the primary cause of lung-related deaths is tobacco-related harm [15], and urban air pollution has significantly worsened the situation, increasing the risk of acute respiratory conditions. According to a study [16] presented

by the European Union's Statistical Office (Eurostat) on respiratory system ailments, in 2016, 339,000 deaths occurred in the EU due to respiratory diseases, accounting for 7.5% of all European Union (EU) fatalities. Furthermore, it reveals that in 2017, respiratory diseases were responsible for at least one in every ten deaths in countries like Spain, Portugal, Belgium, Greece, Malta, and others. Examining data from the National Institute of Statistics (INE) [17], there has been a significant increase in deaths caused by respiratory diseases in Spain over the past four decades. As illustrated in Figure 1.1, in 2021, there were approximately 35543 deaths in Spain. Among the most common respiratory system diseases are asthma, chronic obstructive pulmonary disease (COPD), pulmonary emphysema, lung cancer, pneumonia, and allergies. When considering some of the most common lung diseases, such as asthma and COPD (Chronic Obstructive Pulmonary Disease), the WHO projects that there are presently more than 339 million individuals worldwide living with asthma. This makes it the most prevalent chronic condition among children, affecting 14% of the global pediatric population. Specifically, in 2016, the WHO recorded approximately 420,000 asthma-related fatalities, with the majority of these occurring in the elderly population, as both the elderly and children are the most vulnerable groups to this ailment [18]. Conversely, the WHO estimates that around 64 million people are currently afflicted by COPD, and roughly 3 million people succumb to the disease annually, ranking it as the third leading cause of death worldwide [19].

Analyzing respiratory diseases specifically, one of the primary challenges is that a significant number of these diseases often go undiagnosed correctly in the initial diagnosis by the physician. This error results in the patient's return with a more severe form of the disease that wasn't initially detected. This is where the concept of eHealth (electronic health) becomes relevant. It aims to apply new tools derived from Information and Communication Technologies (ICTs) to achieve, among other goals, an improvement in the physician's initial diagnosis, thereby avoiding missed early diagnoses. This, in turn, enhances the quality of life for patients and minimizes the costs of the healthcare system by increasing its efficiency. Improvement occurs not only by avoiding a more serious illness but also by applying treatment at the precise moment. In this regard, the Spanish Society of Primary Care Physicians reports that misdiagnoses cost the Spanish healthcare system 2,000 euros per patient per year, compared to the 400 euros for a correctly diagnosed patient [20]. Today, globally, the first diagnosis made by a physician in a consultation to evaluate a patient's respiratory health is conducted by analyzing sound signals through auscultation. This process is safe, non-invasive, cost-effective, easy to apply to patients of all ages (patient-friendly), and quick to perform with simple medical equipment, such as a stethoscope. However, the process of auscultation has certain limitations and disadvantages that jeopardize the effectiveness of the diagnosis. Firstly, the diagnosis derived from auscultation remains highly subjec-

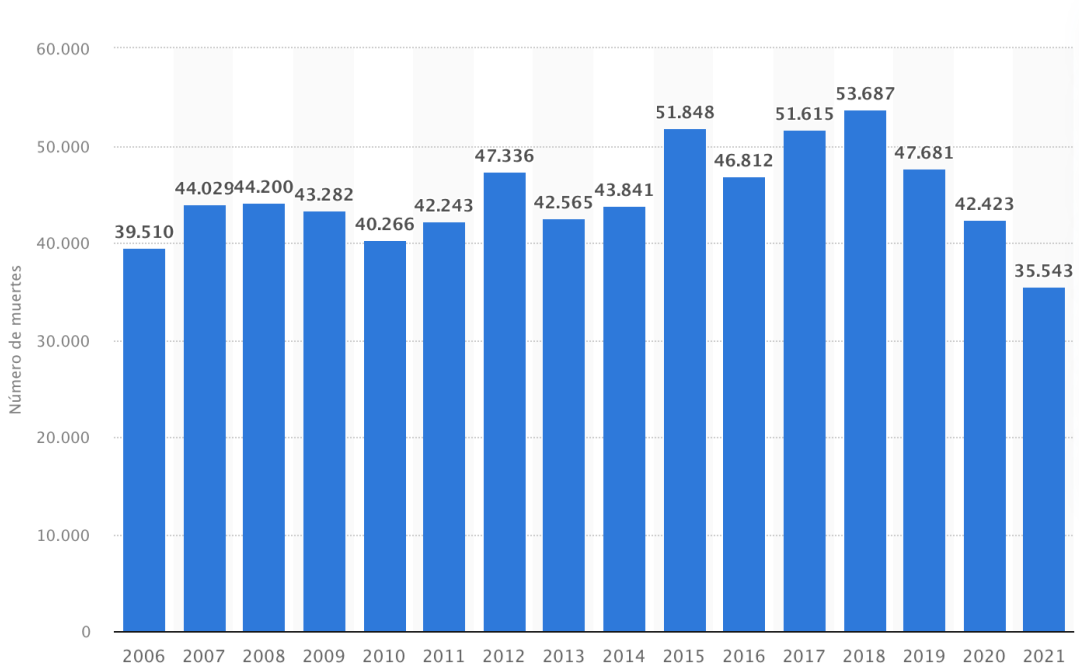


Fig. 1.1 The annual evolution of the number of deaths recorded due to respiratory system diseases in Spain from 2006 to 2021. Figure origin: `https://www.statista.com/statistics/753679/number-of-deaths-from-respiratory-diseases-in-spain/`

tive and depends on the skill, experience, and training of each physician in interpreting these sound signals. This subjective ability for sound signal analysis means that, on numerous occasions, physicians do not correctly identify the source of the patient's potential illness. Furthermore, the fact that normal respiratory sounds and adventitious sounds occur simultaneously in time and frequency makes their separation impossible using traditional filtering techniques [21]. This leads to the normal respiratory sounds, combined with the ambient noise surrounding the patient, interfering with the detection of the relevant adventitious sounds. This interference hampers the cognitive ability of the physician, as they are distracted by these interfering sounds, likely resulting in an increase in false negatives in the diagnoses made. It is also common for the cognitive capacity of the physician to diminish throughout the day, as the number of hours dedicated to analyzing respiratory sounds increases, a situation exacerbated by the stress experienced by the physician in certain medical cases [22, 23].

The adventitious or abnormal sounds that occur when the respiratory system is disordered are diverse. Among them, you can find wheezing, crackles, stridor, rhonchi, and pleural rubs, among others [24, 25, 26]. Specifically, wheezing and crackling sounds are considered a reliable acoustic biomarker of the degree of bronchial obstruction associated with various lung diseases such as asthma, acute bronchitis, bronchiolitis, bronchiectasis, COPD, pneumonia, interstitial lung disease (ILD), pulmonary edema,

bronchiectasis, interstitial pneumonia and alveolar proteinosis [27, 28, 29, 30, 31, 32, 33, 34]. Consequently, this fact has led to the detection, analysis, characterization, and improvement of wheezes and crackles being defined as a challenging research area in the field of biomedical signal processing and artificial intelligence (AI). In the last two decades, a series of contributions have emerged aimed at improving the diagnosis derived from adventitious sounds. Specifically, there are works focused on eliminating the ambient noise surrounding the patient during auscultation [35, 36, 37], detecting and temporally locating adventitious sounds [30, 38, 39], and classifying the type of adventitous sounds [40, 41, 42]. However, this line of research can be considered open due, on the one hand, to the need to enhance the performance of the proposed algorithms to increase their reliability and establish ICT as an aid to improve the efficiency of diagnoses, and on the other hand, to the emergence of new tasks that could assist the physician in diagnosis derived from auscultation. A recent study has shown that doctors fail to detect some adventitious sounds due to the overlap of respiratory sounds during inspiration and expiration [43].

Taking advantage of digital stethoscopes and specialized sensors for capturing respiratory signals during the auscultation process, and leveraging the signal processing capabilities in the field of biomedical acoustic signals, this Thesis addresses the development of methods and algorithms to enhance the reliability of the initial diagnosis of the respiratory system's condition and aid in identifying adventitious sounds related to pathologies. Specifically, it proposes strategies to explore different research directions offered by the Auto-Regresive (AR) modelling method in combination with the Support Vector Machine (SVM) classifier for the detection of the crackle sounds.

We also propose a novel time frequency representation as input for different state of the art neural networks classifiers for the detection and classification of different types of adventitious sounds at the same time that we develop an extensive research on different neural networks approaches in order to determine which alternative is best suited for the research field. Considering that, currently, the performance of most methods and algorithms in this research area depends on a training database (machine learning, neural networks, etc.). The contributions made in this Thesis propose algorithms that do not rely on any external database. Instead, they focus on modeling the time-frequency characteristics that differentiate between the various sounds present during auscultation.

1.2 Justification and research objectives

The importance of this Thesis lies in the fact that the diagnosis derived from the auscultation process remains a subjective diagnosis that is dependent on the skill, experience, and training of each doctor to recognize and interpret the sound signals heard through the stethoscope. The main objective of this Thesis is therefore to advance research on

adventitious sound detection and classification methods by exploiting the information contained in the audio signals from auscultation using signal processing and deep learning algorithms. Specifically, the objective of this Thesis is the development of novel methods and algorithms, time frequency representations and deep learning algorithms applied to the processing of the audio signal from the respiratory system to provide an additional information source to the doctor that helps identify the presence of pathologies resulting from adventitious sounds and increases the reliability of the diagnosis by analyzing the sound signals captured through the auscultation process. The fulfillment of this general objective aims to reduce the rate of false negatives in the detection of potential lung diseases resulting from wheezes and crackles, avoiding putting patients' health at risk and reducing the associated cost to healthcare centers.

Taking into account the tasks and significant challenges for pulmonologists in the analysis and characterization of adventitious sounds, it is necessary to establish a series of specific objectives whose achievement will allow the attainment of the general objective:

- Creation and compilation of databases of pulmonary sound recordings from healthy patients (only respiratory sounds) and patients with illnesses (respiratory sounds mixed with adventitious sounds).

- Development and implementation of signal processing algorithms for the detection of adventitious respiratory sounds (crackles) in simulated respiratory sound signals. For this objective two main tasks were performed; The creation of a database of simulated events using an algorithm based on the mathematical description of crackling events and a system based on an auto-regressive method in combination with an SVM and deep learning classifiers for the two-class classification, crackles and healthy patient.

- Development and implementation of different time-frequency representation as an input for different state of the art deep learning classifiers for 4 class the detection of adventitious respiratory sounds (wheezes, crackles, both, and none (healthy patient)) in respiratory sound signals. As a results of this research, the cochleogram was determined to be the most suitable time frequency representation for the task.

- Development and implementation of signal processing algorithms for the classification of adventitious respiratory sounds in respiratory sound signals. For this objective, several deep learning-based systems were implementd, for the classification of adventitious events into 4 classes (crackles, wheezes, both, none). As a results of this research, the Vision Transformer was determined to be the most suitable deep learning classififcation method for the task.

1.3 Scientific contributions

This Thesis dissertation include two published works and a third work that is currently under review. The natural order of this works is as follow.

[P1] Automatic Robust Crackle Detection and Localization Approach Using AR-Based Spectral Estimation and Support Vector Machine (Mang, L. D., Carabias-Orti, J. J., Canadas-Quesada, F. J., de la Torre-Cruz, J., Muñoz-Montoro, A., Revuelta-Sanz, P., Combarro, E. F. (2023). Applied Sciences, 13(19), 10683. `https://doi.org/10.3390/app131910683`)

Auscultation primarily relies upon the acoustic expertise of individual doctors in identifying, through the use of a stethoscope, the presence of abnormal sounds such as crackles because the recognition of these sound patterns has critical importance in the context of early detection and diagnosis of respiratory pathologies. In this paper, we propose a novel method combining autoregressive (AR)-based spectral features and a support vector machine (SVM) classifier to detect the presence of crackle events and their temporal location within the input signal. A preprocessing stage is performed to discard information out of the band of interest and define the segments for short-time signal analysis. The AR parameters are estimated for each segment to be classified by means of support vector machine (SVM) classifier into crackles and normal lung sounds using a set of synthetic crackle waveforms that have been modeled to train the classifier. A dataset composed of simulated and real coarse and fine crackles sound signals was created with several signal-to-noise (SNR) ratios to evaluate the robustness of the proposed method. Each simulated and real signal was mixed with noise that shows the same spectral energy distribution as typically found in breath noise from a healthy subject. This study makes a significant contribution by achieving competitive results. The proposed method yields values ranging from 80% in the lowest signal-to-noise ratio scenario to a perfect 100% in the highest signal-to-noise ratio scenario. Notably, these results surpass those of other methods presented by a margin of at least 15%. The combination of an autoregressive (AR) model with a support vector machine (SVM) classifier offers an effective solution for detecting the presented events. Even though this approach exhibits enhanced robustness against variations in the signal-to-noise ratio that the input signals may encounter, the main limitation of this work was the lack of real world sound samples. Future work would focus on combining recurrent and convolutional neural networks approaches using different time-frequency representations in order to develop novel criteria to determine the most reliable and discriminant feature map in terms of the real world abnormal respiratory sound to be detected.

[P2] Cochleogram-based adventitious sounds classification using convolutional neural networks (Mang, L. D., Canadas-Quesada, F. J., Carabias-Orti, J. J., Combarro, E. F., Ranilla, J. (2023). Biomedical Signal Processing and Control, 82,

104555. https://doi.org/10.1016/j.bspc.2022.104555)

The World Health Organization (WHO) establishes as a top priority the early detection of respiratory diseases. This detection could be performed by means of recognizing the presence of acoustic bio-markers (adventitious sounds) from auscultation because it is still the main technique applied in any health center to assess the status of the respiratory system due to its non-invasive, low-cost, easy to apply, fast to diagnose and safe nature. Despite the novel deep learning approaches applied in this biomedical field, there is a notable lack of research that rigorously focuses on different time–frequency representations to determine the most suitable transformation to feed data into Convolutional Neural Network (CNN) architectures. In this paper, we propose the use of the cochleogram, based on modeling the frequency selectivity of the human cochlea, as an improved time–frequency representation to optimize the learning process of a CNN model in the classification of respiratory adventitious sounds. Our proposal is evaluated using the largest and most challenging public database of respiratory sounds. The cochleogram obtains the best binary classification results among the compared methods with an average accuracy of 85.1% in wheezes and 73.8% in crackles, and a competitive performance evaluating a multiclass classification scenario in comparison with other well-known state-of-the-art deep learning models. The cochleogram provides a suitable time–frequency representation since it is able to model respiratory adventitious content more accurately by means of non-uniform spectral resolution and due to its increased robustness to noise and acoustic changes. This fact implies a significant improvement in the learning process of CNN models applied in the classification of respiratory adventitious sounds.

[P3 - Under Review] Classification of Adventitious Sounds combining Cochleogram and Vision Transformers (L.D. Mang, F.D Gonzalez-Martinez, D. Martínez-Muñoz, S. García-Galán, R. Cortina). Biomedical Sensors, Advanced Machine Intelligence for Biomedical Signal Processing.

Early detection of respiratory diseases is crucial for improving lung health and reducing morbidity and mortality worldwide. The analysis of respiratory sounds plays a significant role in characterizing the respiratory system's condition and identifying abnormalities. In this paper, we investigate the performance of the Vision Transformer (ViT) architecture for respiratory sound classification, focusing on the input data representation. ViT has shown promising results in audio classification tasks by applying self- attention to spectrogram patches. We extend this approach by exploring the use of the cochleogram time frequency representation, which captures unique temporal and spectral features of adventitious respiratory sounds. The proposed methodology is evaluated on the publicly available dataset from the International Conference on Biomedical and Health Informatics (ICBHI). We compare the classification performance of ViT with other state-of-the-art convolutional neural network (CNN) approaches using spec-

trogram, Mel frequency cepstral coefficients, constant-Q transform, and cochleogram as input data. Our results demonstrate the effectiveness of the cochleogram representation and highlight the potential of ViT for reliable respiratory sound classification. This study contributes to the ongoing efforts in developing advanced signal processing and artificial intelligence techniques with the aim to significantly augment the speed and effectiveness of respiratory disease detection, thereby addressing a critical need in the medical field.

1.4 Thesis structure

In this section the structure of the research work of the Thesis is presented. It is developed into seven chapters as described above:

Chapter 1: The first chapter, which is where we are currently, primarily focuses on enhancing the motivation that has sparked this line of research, as well as presenting the objectives and organization of the doctoral Thesis. Finally, it presents the scientific contributions of this Thesis.

Chapter 2: The second chapter is dedicated to presenting the fundamental concepts related to the analysis of respiratory sounds. First, it describes the respiratory system responsible for generating biomedical respiratory sounds from an anatomical, physiological, and pathological perspective. Second, an introduction is provided to the auscultation process, emphasizing its basic principles, advantages, and limitations, as well as several options commercial available for recording auscultated sounds. To conclude the chapter, a comprehensive classification of the types of respiratory sounds produced by the respiratory system is introduced, with a focus on adventitious sounds, specifically wheezes and crackles in detail.

Chapter 3: The third chapter is dedicated to the literature review for the detection and classification of adventitious sounds. Starting with the fundamentals and works related to the preprocesing, feature extraction and different classifiers used in the analysis of adventitious sounds signals. Furthermore, we give an extensive description of the databases and repositories, public or privates, that are widely used in the literature review for this work, including real life adventitious sounds signals and artificially created adventitious sounds signals. Finally, we explain the most relevant metrics used in the publications included in this work.

Chapter 4: The fourth chapter is dedicated to the first publication we achieved during this Thesis. In this research, we address the critical issue of identifying abnormal respiratory sounds, specifically crackles, during auscultation, which is heavily reliant on the skills of individual physicians. We propose a novel approach that combines autoregressive-based spectral features with a support vector machine (SVM) classifier to detect and locate crackle events from monaural respiratory sounds. The method

includes a preprocessing step to focus on the relevant frequency band and segment the signals. Using SVM, we classify these segments as either crackles or normal lung sounds based on synthetic crackle waveforms used for training. Our dataset comprises both synthetic and real crackle sound signals with varying signal-to-noise ratios.

Chapter 5: The fifth chapter is dedicated to the second publication of this Thesis. In this work we argue than while deep learning has made strides in this field, there is a lack of research on optimal time-frequency representations for Convolutional Neural Network (CNN) models. This study introduces the cochleogram, which models human cochlear selectivity, as an improved representation for CNNs in classifying respiratory adventitious sounds. The cochleogram is non-uniform spectral resolution and robustness to noise enhance CNN learning in respiratory sound classification.

Chapter 6: The sixth chapter is dedicated to our last work of this Thesis. This research explores the use of the Vision Transformer (ViT) architecture for respiratory sound classification, focusing on input data representation. ViT, applying self-attention to spectrogram patches, has shown promise in audio classification. This study extends ViT's capabilities by using the cochleogram representation, which captures unique features of adventitious respiratory sounds. Evaluation on a publicly available dataset demonstrates the cochleogram's effectiveness, highlighting ViT's potential for reliable respiratory sound multiclass classification. This research contributes to advancing signal processing and AI techniques, addressing the critical need for faster and more effective respiratory disease detection.

Chapter 7: This chapter serves as the culminating point in this comprehensive Thesis, encapsulating the essence of the research journey, and offering vital insights into the future prospects and potential enhancements that can be undertaken. This final chapter consolidates the findings and contributions made throughout the Thesis, underlining their significance in the broader context of the field of study. The future work section discusses how the research findings can be further refined, expanded, or applied in real-world scenarios. It explores the gaps and limitations of the current study and suggests strategies for addressing them. The research community, policymakers, and practitioners can take inspiration from these recommendations to advance the field, develop new methodologies, and promote innovation.

CHAPTER 2

Fundamentals of biomedical respiratory sounds

Biomedical respiratory sounds (acoustic biomarkers) serves as the input signal for the various algorithms or methods developed in this Thesis. Therefore, it is essential to understand the nature of these sounds, how they are generated, how they are recorded during auscultation, and how they can be classified. In this regard, this chapter begins with a description of the human respiratory system responsible for generating these sounds from an anatomical, physiological, and pathological perspective. Secondly, it outlines the fundamental principles related to the auscultation process, which allows for the capture of respiratory system sounds. Lastly, a comprehensive classification of the various types of biomedical sounds produced by the respiratory system is provided.

2.1 Human respiratory system

The respiratory system is of vital importance to humans. A person can live several weeks without food and several days without water, but only a few minutes without oxygen. In fact, most people wouldn't be able to survive without breathing for more than 3 minutes, and even if they tried to hold their breath longer, their autonomic nervous system would take control. This is because every cell in the body requires a continuous supply of oxygen to produce energy, grow, repair, or rebuild, and to maintain the vital functions of the organism. However, while oxygen is a critical need for cells, it's actually the accumulation of carbon dioxide that primarily drives the need to breathe.

Below, the basic principles of the human respiratory system are presented in terms of anatomy, physiology, and obstructive pathologies it can suffer from. The goal is not to conduct an exhaustive study on a topic as vast and complex as the one at hand but to provide foundational knowledge about the apparatus that generates the biomedical respiratory sounds under study in this Thesis. For a more in-depth exploration of this topic, the bibliography used to develop the information in this section is referenced [44, 45, 46, 47, 48, 49, 50].

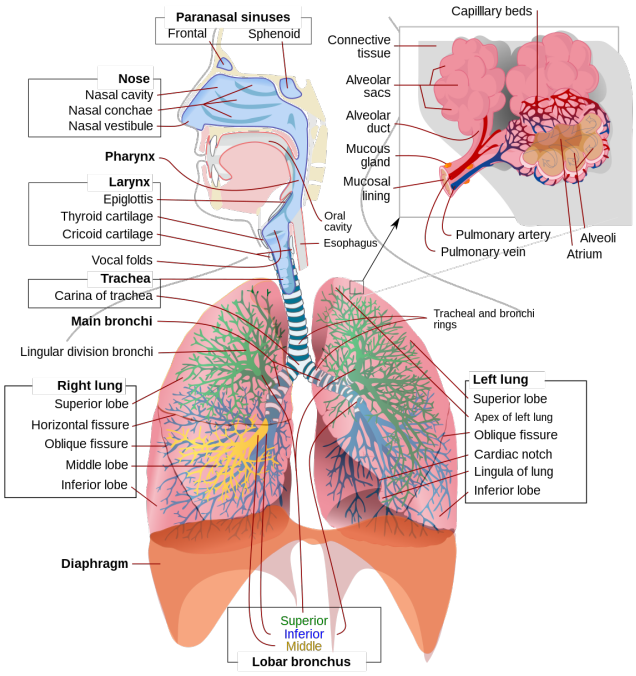


Fig. 2.1 Components of the human respiratory system. Figure origin: `https://www.britannica.com/summary/human-respiratory-system`

2.1.1 Anatomy of the human respiratory system

The human respiratory system is defined as the network of organs and tissues involved in respiration [46, 51]. In general terms, it includes the respiratory pathways, the lungs, blood vessels, and the muscles that drive the lungs. These parts work together with the primary goal of moving oxygen throughout all the cells in the body and eliminating harmful gases from the organism, such as carbon dioxide. Specifically, the human respiratory system begins its operation in the nasal cavity, continues in the pharynx and larynx, which lead to the trachea branching to create bronchi, and ultimately descending through bronchioles to reach the alveolar sacs. This tree terminates in swollen structures called alveoli, which are composed of a single layer of squamous cells surrounded by a network of capillaries. Within the alveoli, the main function of the respiratory system takes place, the exchange of gases (oxygen and carbon dioxide). In Figure 2.1, you can see the different parts that make up the respiratory system, starting from the nasal cavity and ending in the alveoli located inside the lungs [52, 53, 54].

The organs of the respiratory system form a continuous system of cavities or passages called the respiratory tract, through which air enters and exits the body. The respiratory pathways can be divided into two generic groups: the upper respiratory pathways (upper respiratory tract) and the lower respiratory pathways (lower respiratory tract). The main organs that make up each group are shown in Figure 2.1.

Upper respiratory tract: The upper respiratory tract consists of organs located

outside the thoracic cavity: nasal cavity, pharynx, and larynx. All the organs and structures that make it up participate in the conduction or movement of air in and out of the body. Specifically, the organs of the upper respiratory tract provide a pathway for air to move between the external atmosphere and the lungs. They also clean, moisten, and warm incoming air. However, no gas exchange occurs in these organs. The different parts that make up the upper respiratory tract are described below [46, 47]:

Nasal Cavity: it is an air-filled space located behind the nose. Specifically, it is situated above the bone that forms the palate and curves downward and backward to join with the throat. It is divided into two sections called nasal passages. As inhaled air flows through the nasal cavity, it is warmed and moistened. The nose hairs help trap larger particles in the air before they enter further into the respiratory tract. In addition to its respiratory functions, the nasal cavity also contains chemoreceptors necessary for the sense of smell and makes a significant contribution to the sense of taste.

Pharynx: it is a tubular structure that connects the nasal cavity and the back of the mouth with other lower structures in the throat, including the larynx. The pharynx serves a dual function: it allows both air and food (or other ingested substances) to pass through, making it a common passage that is part of both the respiratory and digestive systems. Air moves from the nasal cavity through the pharynx to the larynx (as well as in the opposite direction). On the other hand, food travels from the mouth through the pharynx to the esophagus. This approximately 12.5 cm long passage is divided into three regions: nasopharynx, oropharynx, and laryngopharynx.

Larynx: it is a mobile structure that connects the pharynx and the trachea, assisting in the passage of air through the respiratory pathways, usually acting as a valve that prevents swallowed items and foreign bodies from entering the lower respiratory tract, thus preventing choking. The larynx is also known as the vocal organ because it contains the vocal cords, which vibrate when air flows over them, producing sound. Specifically, it is formed by 9 cartilages (epiglottis, thyroid, cricoid, and three pairs of smaller cartilages). Some of these cartilages allow the separation of the vocal cords to facilitate the breathing process, while others enable the movement of the vocal cords to produce the various types of sounds emitted by humans.

Lower respiratory tract The lower respiratory tract consists of a series of organs located in the thoracic cavity. It includes the trachea and other passages of the lower respiratory tract that carry air from the upper respiratory tract to the lungs. These passages form an inverted tree known as the bronchial tree, with repeated branching as they extend into the lungs (bronchi, bronchioles, and alveolar ducts), ultimately connecting with alveolar sacs containing the alveoli. In total, there are approximately an astonishing 2000 km of airways that carry air through the human respiratory tract. However, it is only in the lungs, specifically in the alveoli, that gas exchange between the air and the bloodstream occurs. The different parts that make up the lower respiratory tract are

described below [45, 50, 44]:

Trachea: it is the widest conduit of the respiratory pathways. Its dimensions typically range from about 10 to 15 cm in length and 2.5 cm in width. It is a cartilaginous tube that connects the larynx to the primary bronchi of the lungs, allowing the passage of air. The trachea extends from the larynx and bifurcates into the two main left and right bronchi. It is composed of a series of hyaline cartilages (usually 16 to 20) shaped like the letter 'C'. The open part of each ring (the open side of the 'C') is made up of muscle and connective tissue. A moist and smooth tissue called mucosa lines the interior of the trachea. The trachea slightly widens and lengthens with each inhalation, returning to its resting size with each exhalation, playing a vital role in the breathing process. The trachea's starting point connects to the larynx through a cartilage called the cricoid, and its endpoint branches into the primary bronchi that connect to each lung.

Bronchi and Terminal Airways: The first bronchi branching off from the trachea are the left and right main bronchi, also known as primary bronchi. Their structure is similar to that of the trachea, although instead of rings, they have overlapping cartilaginous plates. The right main bronchus is short and wide, while the left one is long and narrow. The main bronchi connect the trachea to the lungs and enter the lungs through a region called the HILUM. Upon reaching the respective lungs, the main bronchi subdivide into secondary or lobular bronchi, and these further branch into narrower tertiary bronchi or segmental bronchi. Subsequently, there are further divisions of the segmental bronchi that group together to form the so-called subsegmental bronchi. When these passageways are too narrow to be supported by cartilaginous tissue, they are called bronchioles, which have walls consisting of smooth muscle and lack cartilage. The bronchioles continue to branch, giving rise to alveolar ducts, which lead to alveolar sacs or clusters of microscopic alveoli. Alveoli are primarily composed of simple squamous epithelium, enabling rapid oxygen and carbon dioxide diffusion. These tiny air sacs are the functional units of the lungs where gas exchange occurs between the air in the lungs and the blood in the capillaries. The two lungs can contain up to 700 million alveoli, providing an enormous total surface area for gas exchange.

Lungs: they contain all the components of the bronchial tree beyond the primary bronchi and are the largest organs in the respiratory system, occupying most of the space within the thoracic cavity. These are two lightweight, elastic, and spongy organs suspended within the pleural cavity of the chest, separated by an area known as the mediastinum, where the heart is located. The right lung is shorter than the left due to the upward push from the liver against the diaphragm, though the left lung has a smaller volume because of the position of the heart. Specifically, the right lung is divided into three lobes, while the left lung is divided into two, with each lobe being supplied by one of the secondary bronchi. On the mediastinal surface of each lung, you will find the pul-

monary HILUM, the only point of attachment through which blood vessels, lymphatics, bronchi, and nerve fibers pass to each lung. Each lung is enclosed by a double-layered membrane called the pleura. The inner layer (visceral pleura) adheres to the lung's surface. The outer layer (parietal pleura) attaches to the chest wall, mediastinum, and diaphragm. The small space between the visceral and parietal pleura is known as the pleural cavity. This cavity contains a thin film of fluid that serves as a lubricant to reduce friction as the two layers slide against each other and helps to keep the two layers together as the lungs expand and contract during breathing.

Respiratory Muscles: In addition to the extensive network of passages that make up the bronchial tree, the muscles involved in breathing are typically considered part of the lower respiratory tract. The most important of these is a large, dome-shaped muscle called the diaphragm, which curves toward the lungs and separates the chest from the abdomen. When it contracts, it flattens, increasing the volume of the thoracic cavity. Similarly, the contraction of the external intercostal muscles moves the ribs upward and outward. This increase in volume leads to a drop in pressure within the lungs, allowing air to flow passively into the airways. While the diaphragm and intercostal muscles play a key role in the breathing process, their respiratory action is assisted and augmented by a complex set of other muscle groups (abdominals, scalenes, sternocleidomastoid, etc.). Additionally, the muscles of the larynx, pharynx, and nasal cavity adjust the resistance to the movement of gases through the upper respiratory pathways during inhalation and exhalation.

2.1.2 Physiology of the human respiratory system

Although the human respiratory system serves a wide range of functions, its primary role is to facilitate the exchange of gases with the atmospheric air. On one hand, it ensures a constant concentration of oxygen in the blood, which is necessary for metabolic reactions, and on the other hand, it serves as a means to eliminate the body's waste gases resulting from these reactions. This is achieved through the performance of more specific functions such as the regulation and control of respiration, the mechanics of breathing (inspiration and expiration), and gas exchange [55, 53].

Regulation and Control of Respiration

Breathing is an automatic and rhythmic act regulated by the nervous system. The centers that control respiration are located in the brainstem. Specifically, there are two centers, one in the medulla (pneumotaxic center) and another in the pons (apneustic center). The pneumotaxic center is responsible for initiating inhalation, while the apneustic center is responsible for initiating exhalation. Neural networks direct the muscles that form the chest and abdominal walls to create a pressure gradient that moves air in and out of the lungs. The respiratory rhythm is regulated through the recipro-

cal interconnection of stimuli and inhibitions among brainstem neurons. Additionally, sensors distributed throughout the body send signals to the centers that control respiration. On one hand, chemoreceptors detect changes in blood oxygen levels, and on the other hand, mechanoreceptors monitor lung expansion, airway size, and the force of respiratory muscle contraction.

The human respiratory system has the ability to adjust the frequency and depth of respiration based on changes in the internal or external environment. Specifically, within the context of a resting state among adults, the breathing typically ranges from 12 to 16 respirations per minute (rpm) [56, 57]. These variations primarily occur when carbon dioxide levels are high or oxygen levels are low. Considering the modifications in depth and frequency of respiration that can occur, three types of regulation are defined: hyperpnea, hypopnea, and apnea. Hyperpnea involves an increase in the depth and frequency of respiration, hypopnea involves a decrease in respiratory frequency, and apnea is defined as a complete suspension of breathing. In essence, respiratory frequency adapts to the oxygen demand of cells and the necessary elimination of carbon dioxide. Additionally, the respiratory system can self-regulate when there is a disturbance in the airways, such as an asthma attack. Lastly, the process of respiration is regulated when the mechanics of respiratory muscles are altered, for instance, during physical exercise.

Although the respiratory muscles mentioned earlier greatly enhance the flexibility of the act of breathing, they also complicate respiratory regulation. In reality, respiratory muscles serve other functions (such as maintaining posture or even speaking), which is why the process of respiration can be influenced by higher brain centers, leading to voluntary control over respiratory mechanics. Consequently, humans can voluntarily hold their breath or breathe at different frequencies [48, 49].

Mechanics of Respiration (Inspiration and Expiration)

The exchange of air between the lungs and the environment is a cyclical process that involves two movements: inspiration (inhalation) and expiration (exhalation). These processes are essential for providing oxygen to cells and removing carbon dioxide from the body. Inspiration is an active process that involves the expansion of the chest cavity and the inhalation of air. On the other hand, expiration is a passive process in which the reduction of the chest cavity leads to the expulsion of air from the lungs to the outside. Figure 2.2 illustrates the mechanics of respiration during the inspiration and expiration phases.

In the mechanics of respiration, the contraction and relaxation of respiratory muscles serve to change the volume of the chest cavity. As the chest cavity and lungs contract or expand, there is a corresponding change in lung volume, which in turn alters the pressure inside the lungs. From a physical standpoint, Boyle's law states that in a closed space, the volume of gas is inversely proportional to the gas pressure (when

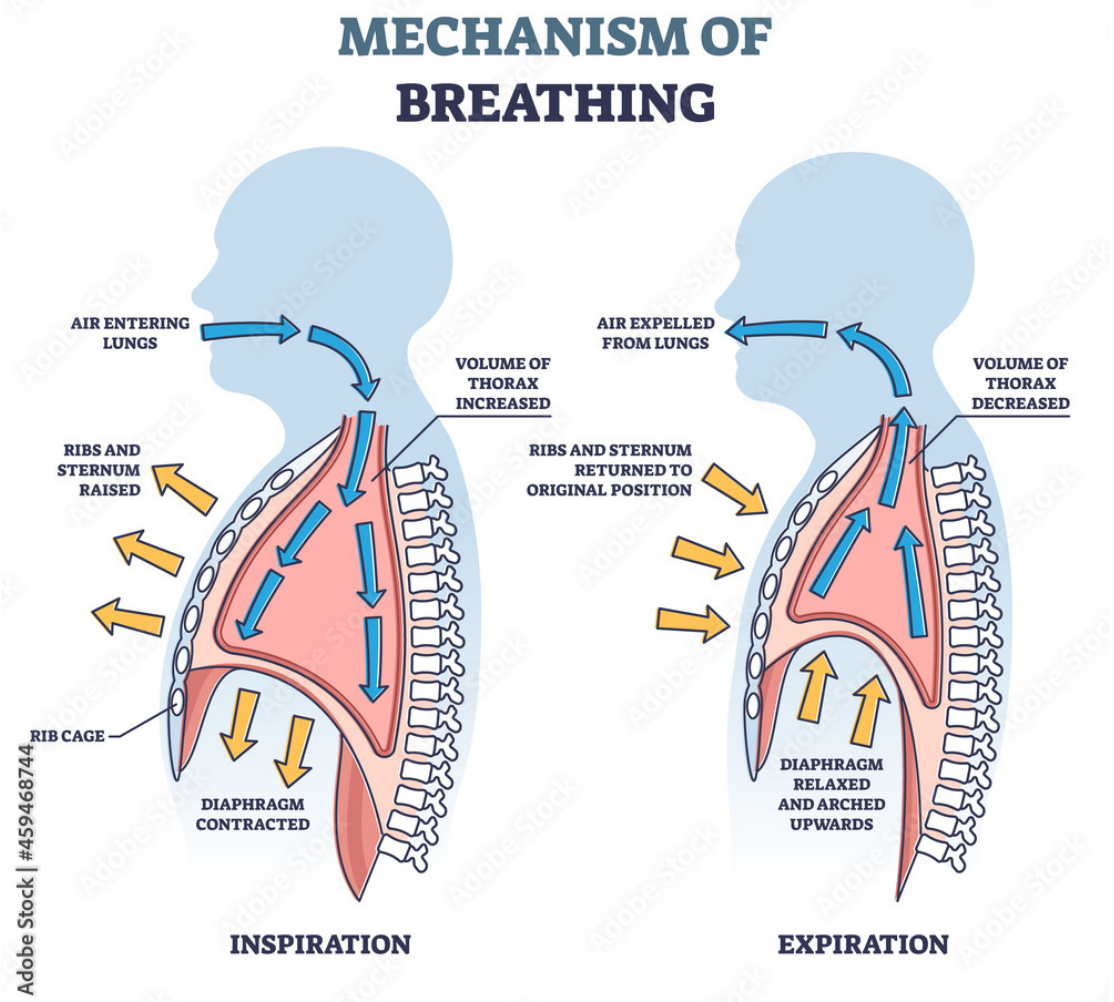


Fig. 2.2 Inspiration and Expiration of the human respiratory system. Figure origin: `https://stock.adobe.com/es/images/mechanism-of-breathing-as-anatomical-process-explanation-outline-dia 459468744`

temperature is constant). In this sense, when the volume of the chest cavity increases, the volume of gas (air) in the lungs also increases, and the air pressure inside the lungs decreases. Conversely, when the volume of the chest cavity decreases, the volume of air in the lungs also decreases, and the air pressure inside the lungs increases. The pressure gradient between the external environment and the inside of the lungs causes air to flow from the area of higher pressure to the area of lower pressure [50, 48].

Inspiration Process: Inspiration is the phase of the respiratory cycle in which air enters the lungs. During this process, the volume of the chest cavity increases in three directions:

- Increase in vertical diameter: caused by the contraction of the diaphragm, which descends towards the abdomen.

- Increase in anteroposterior diameter: caused by the elevation of the chest cavity, brought about by the contraction of various pairs of muscles, including the external intercostal, sternocleidomastoid, pectoralis minor, scalene, and serratus anterior muscles.

- Increase in left-right diameter: caused by the lateral elevation of the ribs when the intercostal muscles contract.

The expansion of the chest cavity involves an expansion of the parietal pleura, which is transmitted to the visceral pleura and the lungs. In this way, the pressure of the gas contained in the lungs decreases, resulting in a suction of air until it reaches atmospheric pressure. As mentioned earlier, an increase in lung volume leads to a decrease in pressure inside the lungs. The pressure of the external environment to the lungs is now greater than the pressure of the air inside the lungs, which means that air moves into the lungs due to the pressure gradient.

Exhalation Process: Exhalation is the phase of ventilation in which air is expelled from the lungs. The relaxation of the respiratory muscles involves a decrease in the volume of the chest cavity, sufficient to allow the air to exit to the outside. The contraction of the chest cavity is transmitted to the lungs. At that moment, the pressure of the gas within the lungs increases, causing the air to be expelled until it equals atmospheric pressure. According to Boyle's law, a decrease in lung volume results in an increase in pressure within the lungs. The pressure inside the lungs is now greater than in the external environment, which means that air exits the lungs due to the pressure gradient.

Gas exchange and gas transport in the blood:

Gas exchange in the lungs occurs between the air entering the alveoli and the blood flowing through the capillaries. The pulmonary gas exchange process removes carbon dioxide from the blood and replenishes the oxygen supply in the blood. The circulatory system is responsible for transporting gases from the lungs to the tissues throughout the body and vice versa.

Gas exchange is facilitated by a pressure gradient existing between the alveoli and the capillaries, through a process known as diffusion. When the respiratory mechanism takes place, the process of inhalation introduces a mixture of gases, including oxygen and carbon dioxide, into the lungs. Each of these gases exerts a different pressure, which is related to its concentration within the gas mixture. These individual pressures are called partial pressures. The difference in partial pressures between the gases in the alveoli and the capillaries creates a pressure gradient across the respiratory membrane (the membrane that separates the alveoli and the blood capillaries). If the pressure on each side of the membrane were the same, there would be no gas exchange. It is the variation in partial pressures of oxygen and carbon dioxide that drives this process.

Gases move in both directions (towards the blood capillaries and towards the lung

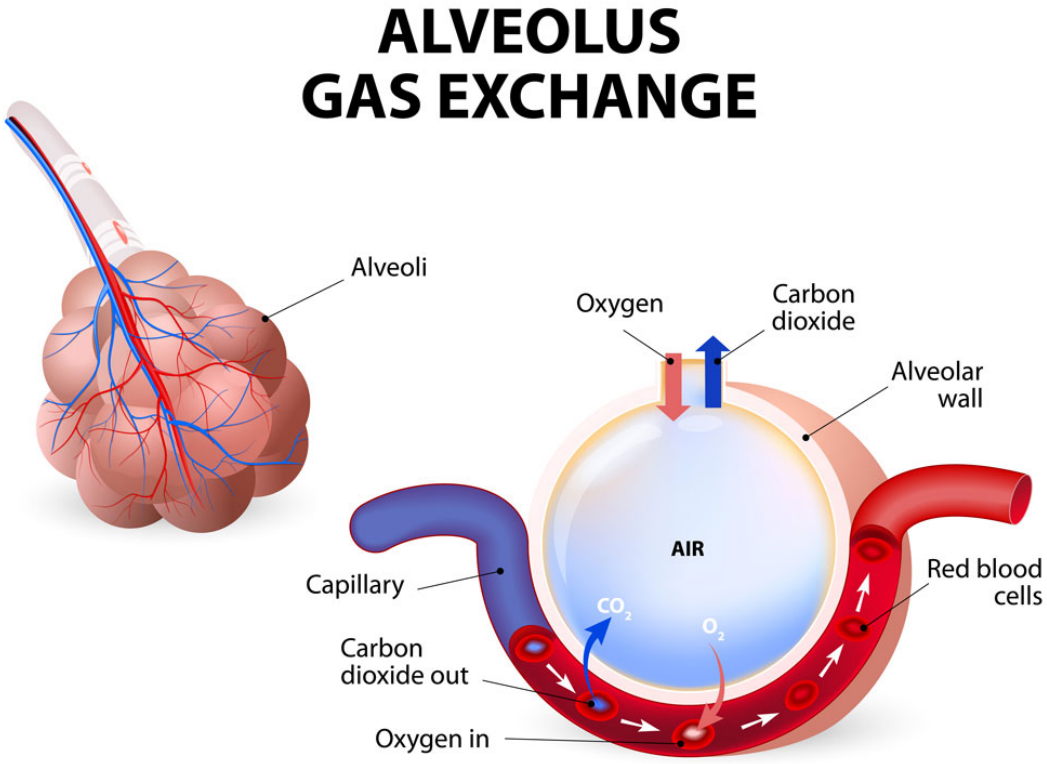


Fig. 2.3 Alveolus gas exchange - Pulmonary alvelous. Figure origin: `https://www.pedilung.com/pediatric-lung-diseases-disorders/anatomy-of-a-childs-lung/alveolus-gas-exchange-pulmonary-alveolus/`

alveoli). Specifically, gases move from areas of high pressure (high concentration) to areas of low pressure (low concentration). During the inspiration phase, the concentration of oxygen in the alveoli is high, and in the blood capillaries, it is low; thus, oxygen travels from the alveoli to the capillaries due to the oxygen gradient. Conversely, there is an inverse gradient for carbon dioxide, which diffuses from the blood to the alveoli to be expelled during the expiration process. Figure 2.3 provides a representative illustration of the gas exchange process occurring in the alveoli.

Once oxygen has crossed the respiratory membrane and reaches the pulmonary blood, it is transported to the tissue capillaries for diffusion into the cells. On the other hand, carbon dioxide diffuses from the body's cells into the blood. The transport of oxygen and carbon dioxide in the blood is primarily facilitated by hemoglobin, a protein located inside red blood cells. Blood with a high level of oxygen traveling from the lungs to all cells in the body is typically referred to as oxygenated blood and appears bright red due to the binding of hemoglobin and oxygen. In contrast, blood with a high level of carbon dioxide traveling from the body's cells to the lungs is usually called deoxygenated blood and is much darker red due to the lack of available oxygen to bind with hemoglobin [49, 50].

2.1.3 Pathologies of the human respiratory system

There is a wide range of pathologies related to the respiratory system, which are typically classified based on the organs, muscles, and tissues they affect. The objective of this section is to provide an introduction to the pathologies that affect the airways of the bronchial tree, which are responsible for producing adventitious sounds. The presence of these sounds, generated by the obstruction of the airways, indicates an abnormality in the proper functioning of the respiratory mechanics, providing relevant information to the physician for determining the patient's diagnosis. Below is a description of the most common and relevant pathologies that affect the airways and are primarily associated with the presence of wheezes DNA crackles sounds. These pathologies include asthma, chronic obstructive pulmonary disease (COPD), bronchiolitis, bronchiectasis, pneumonia, pulmonary edema, interstitial lung disease (ILD).

Asthma:

Asthma is a common respiratory disease that affects over 339 million people worldwide [58], making it the most prevalent chronic disease in children globally. Asthma is primarily characterized by symptoms such as wheezing, shortness of breath, chest tightness, and/or cough, along with variable limitation of the respiratory airflow. These symptoms can vary in their characteristics over time and intensity. These variations are mainly triggered by factors such as exercise, exposure to allergens or irritants, air pollution, changes in weather, or viral respiratory infections. Symptoms and airflow limitation can resolve spontaneously or in response to medication and may sometimes be absent for weeks or months. On the other hand, patients may experience episodic asthma attacks in which the lining of the bronchial tubes swells, causing narrowing of the airways and reducing the airflow in and out of the lungs.

Asthma is often associated with hypersensitivity of the airways to direct or indirect stimuli and chronic inflammation of the airways. These characteristics may persist even when there are no symptoms or when lung function is normal but can be normalized with treatment. Recurrent asthma symptoms can often lead to insomnia, daytime fatigue, reduced activity levels, and school or work absenteeism. Asthma is a disease with a relatively low mortality rate, as appropriate medication can help avoid asthma triggers, and its severity can be controlled. However, the World Health Organization (WHO) estimated that around 420000 asthma-related deaths occurred worldwide in 2016 [58], with countries having weaker healthcare systems being the most affected. Therefore, early detection of the disease is essential to apply appropriate treatment, avoiding potential delays and worsening the quality of life for patients [59].

COPD:

Chronic Obstructive Pulmonary Disease (COPD) is a generic term that encompasses several respiratory diseases causing breathing difficulties and worsening over time. The

World Health Organization (WHO) [24] highlighted that in 2016, the global prevalence of COPD exceeded 250 million cases, and in 2015, approximately 3 million people died from COPD, accounting for 5% of all recorded deaths that year. In a healthy individual, the airways and air sacs in the lungs (alveoli) are elastic. When the mechanics of respiration begin, during the inhalation phase, the airways carry air into the alveoli to initiate the gas exchange process and provide oxygen to all parts of the body. Then, during the exhalation phase, carbon dioxide extracted from the blood is expelled to the outside. However, in the case of individuals with COPD, the process described above is compromised mainly for the following reasons: the airways and the walls of the alveoli become less elastic and rigid, the walls of many alveoli are destroyed, the walls of the airways become thicker and inflamed, and the airways produce more mucus than usual, sometimes leading to obstruction. The most common symptoms experienced by people with COPD are chronic cough, shortness of breath during daily activities (dyspnea), respiratory infections, increased fatigue, increased mucus production (phlegm or sputum), chest tightness, and the presence of wheezing sounds during the mechanics of respiration. The main causes that lead to the development of COPD include exposure to tobacco smoke (both active and passive smokers), indoor/outdoor air pollution, exposure to dust and chemicals, and recurrent lower respiratory infections during childhood. COPD affecting most people translates into the development of emphysema and chronic bronchitis depending on the damaged structures or tissues [60].

- Emphysema: affects the alveoli of the lungs as well as the walls between them. They lose elasticity and can become damaged [61].
- Chronic bronchitis: affects the lining of the airways, which become constantly irritated and inflamed. This causes the airways to swell, making it difficult for air to enter and exit and leading to increased mucus production [19].

Bronchiolitis:

Bronchiolitis is a globally significant lung infection. According to the WHO bulletin [270], it is estimated that 150 million new cases occur each year, of which 11-20 million (7-13%) are severe enough to require hospitalization. Worldwide, 95% of cases occur in developing countries. Specifically, bronchiolitis is a respiratory disease primarily caused by a virus that affects the smaller airways (bronchioles). The function of bronchioles is to control the airflow in the lungs. When these airways become infected or damaged, they can become inflamed or congested. This reduces or blocks the flow of oxygen to the alveoli, where the gas exchange process should occur normally. While this disease is generally common in young children or infants, bronchiolitis can also affect adults. The most common virus associated with bronchiolitis is the respiratory syncytial virus (RSV). However, over the years, it has been discovered that many other

viruses can cause the same infection, including human rhinovirus, coronavirus, human metapneumovirus, adenovirus, human parainfluenza viruses, and human bocavirus. The main symptoms associated with this condition include shortness of breath and fatigue, audible wheezing sounds during breathing, crackling sounds, rapid breathing, labored breathing, cough, nasal congestion, and mucus. There are two main types of bronchiolitis [62], which are:

- Viral bronchiolitis: typically occurs primarily in babies and is due to the presence of a virus in the bronchioles [63].
- Obliterative bronchiolitis or constrictive bronchiolitis: This is a rare and dangerous condition observed in adults. This disease causes scarring in the bronchioles. This blocks the airways, creating an irreversible obstruction of the respiratory passages [64].

Bronchiectasis:

Bronchiectasis is on the rise within respiratory diseases. Scientists currently define it as an emerging ”global epidemic” and label it as an evolving clinical issue due to the lack of therapy and a lack of understanding of its inherent heterogeneity [75]. Recent data suggest that this condition has a prevalence of up to 566 per 100000 individuals in the population, with an approximately 40% increase over the last decade [65]. Specifically, bronchiectasis is a condition in which the bronchial tubes that make up the bronchial tree undergo abnormal widening. In a healthy patient, small glands in the lining of the airways produce a small amount of mucus to keep these tubes moist and trap dust and dirt from inhaled air. However, due to the widening caused by this condition, mucus and bacteria tend to accumulate in these widened areas. This can lead to frequent infections and blockage of the airways. Bronchiectasis symptoms may take months or even years to develop. Some of the typical symptoms that may be present include daily chronic cough, coughing up blood, abnormal chest sounds or wheezing when breathing, shortness of breath, chest pain, coughing up large amounts of thick mucus daily, daily fatigue or tiredness, and frequent respiratory infections. Although there is no cure for this disease, it is manageable. With effective treatment, a normal life can be maintained. However, exacerbations must be treated promptly to maintain the oxygen flow to the rest of the body and prevent further airway damage. Two main categories can be defined for this condition [66], which are:

- Bronchiectasis due to cystic fibrosis. Cystic fibrosis (CF) is a genetic condition that causes abnormal mucus production [67].
- Non-cystic fibrosis bronchiectasis. The most common known conditions that can lead to this condition are: an abnormal immune system, inflammatory bowel

disease, autoimmune diseases, COPD, human immunodeficiency virus (HIV), aspergillosis (an allergic lung reaction to fungi), and lung infections (such as whooping cough or tuberculosis) [66].

Pneumonia:

Pneumonia is a common and potentially serious lung infection that can affect one or both lungs. It is characterized by inflammation of the air sacs in the lungs, leading to the accumulation of pus or other fluids [68]. Pneumonia can be caused by various microorganisms, including bacteria, viruses, fungi, and even chemical irritants. It can result in a range of symptoms, from mild to severe, and it may lead to significant health complications if not properly treated. Pneumonia can originate from various sources, with the most common causes being infectious microorganisms. Some of the primary causes and origins of pneumonia are:

- Bacterial Pneumonia: This is the most common type of pneumonia and is often caused by bacteria, such as Streptococcus pneumoniae (pneumococcus). Other bacterial pathogens that can cause pneumonia include Haemophilus influenzae, Mycoplasma pneumoniae, and Legionella pneumophila.

- Viral Pneumonia: Various respiratory viruses can cause pneumonia. Influenza viruses, respiratory syncytial virus (RSV), and rhinoviruses are examples of viruses that can lead to pneumonia. Viral pneumonia tends to be more common in children and is often less severe than bacterial pneumonia.

- Fungal Pneumonia: Fungal pneumonia is usually less common and is more likely to affect individuals with weakened immune systems. Fungal pathogens like Candida, Aspergillus, and Pneumocystis jirovecii can lead to fungal pneumonia.

- Aspiration Pneumonia: This type of pneumonia occurs when foreign materials, such as food, drink, saliva, or stomach contents, are inhaled into the lungs. Aspiration pneumonia is more common in individuals with conditions that affect swallowing or those who have impaired consciousness, such as due to alcohol intoxication or certain medications.

- Chemical Pneumonia: Exposure to toxic chemicals, gases, or fumes can irritate and damage the lungs, causing chemical pneumonia. Industrial accidents or accidental inhalation of harmful substances can lead to this type of pneumonia.

The symptoms of pneumonia can range from mild to severe and may include fever, cough, difficulty breathing, chest pain, and fatigue. Treatment depends on the specific type of pneumonia and its underlying cause, but it typically involves antibiotics for bacterial pneumonia and antiviral medications for viral pneumonia. It is essential to

seek prompt medical attention if you suspect you have pneumonia, as early diagnosis and treatment can prevent complications and aid in a full recovery.

Interstitial Lung Disease (ILD):

Interstitial Lung Disease (ILD) is a group of lung disorders characterized by inflammation and scarring (fibrosis) of the interstitium, which is the tissue that surrounds and supports the air sacs (alveoli) in the lungs. The interstitium consists of a network of tiny blood vessels, connective tissue, and air spaces, and its primary function is to support the alveoli and facilitate the exchange of oxygen and carbon dioxide between the lungs and the bloodstream. ILD affects the interstitium, making it difficult for the lungs to function effectively [69]. In many cases, the specific cause of ILD is unknown, and it is referred to as idiopathic interstitial lung disease (IILD). In other cases ILD can also be caused by various factors, such as exposure to environmental toxins (e.g., asbestos, dust, or mold), infections (e.g., pneumonia or tuberculosis), medications, autoimmune diseases (e.g., rheumatoid arthritis or systemic sclerosis), and hypersensitivity pneumonitis (an allergic reaction to inhaled substances). There are over 200 different types of ILD, with some of the most common forms being:

- Idiopathic Pulmonary Fibrosis (IPF): A progressive and often fatal form of ILD with no known cause.
- Non-specific Interstitial Pneumonia (NSIP): A less severe form of ILD with inflammation and scarring.
- Cryptogenic Organizing Pneumonia (COP): A condition where small airways and alveoli become blocked.
- Sarcoidosis: A rare inflammatory condition that can affect multiple organs, including the lungs.

ILD is a complex group of diseases, and each subtype may have unique characteristics and treatment approaches. It's essential to work closely with a healthcare team, including pulmonologists and specialists, to receive an accurate diagnosis and an appropriate treatment plan tailored to the specific condition. Early diagnosis and management are crucial in improving the quality of life and long-term outcomes for individuals with ILD.

Pulmonary Edema:

Pulmonary edema is a medical condition characterized by the accumulation of excessive fluid in the air sacs (alveoli) and the interstitium of the lungs [70]. This accumulation of fluid impairs the ability of the lungs to exchange oxygen and carbon dioxide efficiently. Pulmonary edema can be a potentially life-threatening condition, and it may be caused by various underlying factors. It can manifest suddenly (acute

pulmonary edema) or develop gradually over time (chronic pulmonary edema). The most common type and is often associated with heart-related issues. It can result from conditions such as congestive heart failure, heart attacks, heart valve problems, or high blood pressure. The second type of pulmonary edema is non-cardiogenic pulmonary edema. This type is not primarily related to heart problems. It can be caused by factors like acute respiratory distress syndrome (ARDS), pneumonia, lung infections, exposure to toxic gases or chemicals, high-altitude pulmonary edema (HAPE), or severe kidney disease. The symptoms of pulmonary edema can vary depending on its severity and the underlying cause, but they often include: Severe shortness of breath, especially when lying down, a persistent cough, often producing pink or frothy sputum, rapid, shallow breathing, anxiety or restlessness, wheezing or crackling sounds in the lungs, cyanosis (bluish discoloration of the skin and lips) due to a lack of oxygen.

Treating the root cause, such as managing heart failure, addressing infections, or treating toxic exposures. Pulmonary edema is a medical emergency when it occurs acutely, especially if it results from heart-related issues. Prompt medical intervention is essential to manage symptoms and prevent life-threatening complications. Chronic pulmonary edema may be a sign of an underlying health problem that needs appropriate management and treatment. It is crucial to consult a healthcare professional for a thorough evaluation and diagnosis if you suspect pulmonary edema.

2.2 Auscultation process

Doctors have been listening to patients' bodies to make their diagnoses probably since the practice of healing began. Specifically, it was Hippocrates who initiated the concept of auscultation by applying the ear to the patient's chest to hear the respiratory sounds transmitted inside, calling this procedure "immediate or direct auscultation." However, the concept of the stethoscope did not emerge until 1816. The French physician René Laënnec needed to listen to the sounds that occurred in a patient's chest, so he rolled up a long piece of paper into a tube and observed that with that device, he could hear much better than by placing his ear directly on the patient's chest. Laënnec published his masterpiece in 1819 [71], and it was at that moment that the art of auscultation began, quickly becoming popular worldwide. Laënnec coined the name "stethoscope" from two Greek words: "stethos" (chest) and "skopein" (to observe). He also called the process auscultation, from the Latin word "auscultare" (to listen). Twenty-five years after Laënnec invented the stethoscope, George P. Camman developed a design that included a headset for each ear [72]. Medical professionals continued to use this design with few changes for almost a century. It wasn't until the early 1960s when Dr. David Littmann patented a new design that significantly improved the acoustic performance of the stethoscope. A few years later, 3M acquired Dr. Littmann's stethoscope business.

Eventually, Dr. Littmann's designs became the new standard for stethoscopes, and now 3M Littmann is the most trusted brand in the business [72].

Although the stethoscope emerged over 200 years ago, the process of auscultation remains the first clinical examination that a physician uses to diagnose possible obstructive pathologies in the human respiratory system. This is because respiratory system auscultation is a low-cost, non-invasive, safe, and easy-to-perform diagnostic technique. Furthermore, the ability to differentiate between normal and abnormal sounds (adventitious sounds) remains essential in clinical practice for an efficient diagnosis. Thanks to the invention of the electronic stethoscope and specialized sensors and microphones, in the past two decades, significant contributions have appeared that address the analysis of respiratory system sounds to aid in improving diagnosis, as well as in health teaching and pedagogy. The purpose of this section is to present the most relevant information for understanding the principles of respiratory auscultation, its advantages and limitations, the parts and types of stethoscopes, the most current important commercial electronic stethoscopes (used by doctors), the microphones or sensors used as an alternative to capture the sounds from inside the patient, and the alternative tools and techniques available to analyze and diagnose the human respiratory system. To do this, some of the most relevant articles and bibliography in the field of auscultation have been used [73, 74, 75, 76].

2.2.1 Principles of respiratory auscultation

To carry out the clinical examination of the respiratory system, a widely known and used assessment approach in medicine is typically employed, in which auscultation plays a fundamental role. This assessment approach is known by its acronym IPPA (International Positive Psychology Association) [73, 77], as it can be divided into the following four stages:

- Inspection: the use of the senses of sight, smell, and hearing to observe the normal condition, or any deviation from normal, of the person as a whole, as well as of a specific area.

- Palpation: touching and feeling parts of the body with the hands to assess temperature, texture, moisture, movement, and the consistency of structures. Regarding respiratory assessment, you can detect the symmetry of the lungs.

- Percussion: the transmission of sound that occurs when the doctor taps with a finger in short, sharp strokes against another finger placed firmly on a specific organ. This allows determining the density of structures within a cavity, such as the thoracic cavity. The lungs should sound hollow upon percussion because they are filled with air.

- Auscultation: listening to the movement of air that occurs in the airways comprising the upper and lower respiratory tract to determine the presence of various normal and abnormal sounds that occur during the mechanics of respiration.

In general terms, auscultation involves listening to the internal sounds of the body, typically using a stethoscope. Auscultation is performed to examine the circulatory system and the respiratory system (cardiac and respiratory sounds), as well as the gastrointestinal system (intestinal sounds). This process is part of a patient's clinical examination and is commonly used to provide strong evidence either including or excluding different pathological conditions that manifest clinically in the patient. In the case of the circulatory system, the doctor examines the four main areas where the sounds of heart valves are loudest (aortic, pulmonary, mitral, and tricuspid). During the listening, the doctor pays special attention to how the heart sounds, the frequency of each sound, and the intensity with which it sounds. Cardiac sounds are traditionally rhythmic in the short term, between 60-100 beats/min in a healthy adult, so any variation may indicate that some areas may not be receiving sufficient blood or that some valves may be leaking. In the case of the gastrointestinal system, the doctor examines different regions of the abdomen to listen to the various sounds present. In normal conditions, sounds should be heard in any region of the abdomen. The absence of sound in any region may indicate that digested material may be stuck or the intestine may be twisted. Finally, in the case of the respiratory system, the doctor examines the main areas of the bronchial tree. The airflow sounds different when the airways are obstructed, narrowed, or filled with liquid and mucus. Detecting these abnormal sounds is crucial for identifying respiratory pathology. Focusing on the subject of this Thesis, this section describes the basic principles of auscultation, considering the biomedical sounds produced by the respiratory system, which are considered rhythmic long-term sounds, between 12-20 breaths/min in a healthy adult.

Traditional auscultation of respiratory sounds is usually characterized by a series of considerations that allow for an efficient interpretation of respiratory sounds [78, 79], which are summarized in the following list:

- To begin with, the optimal environment [73] during respiratory auscultation should be characterized by conditions: (i) quiet, as background noise can interfere with the listening of the sounds of interest, so the patient should remain silent; (ii) warm, to ensure the patient's comfort and to prevent shivering that can be added as interfering noise; and (iii) well-lit, to accurately detect the positions that need to be auscultated.

- During the examination, the patient should breathe slightly deeper than usual through the mouth. This allows for an increase in the intensity of the respiratory sounds of interest, facilitating their detection.

- The auscultation process should be performed around the thoracic surface, which is divided into three areas: the posterior thoracic surface, the anterior thoracic surface, and the lateral thoracic surface.

- Auscultation should be performed systematically and comparatively between the left and right sides to detect any asymmetry between the sounds in both lungs. To do this, a ”stepladder” approach is commonly used, which involves auscultating different regions following a zig-zag pattern (see Figure 2.4).

- The procedure starts by auscultating the 7 focal points on the anterior surface and lateral surfaces of the thorax, as shown in Figure 2.4 (points 6 and 7 correspond to the lateral surfaces). A zig-zag pattern is used, starting in the right supraclavicular region and moving downward while comparing the left and right sides alternately. Following the same approach, the 6 focal points on the posterior surface and lateral surfaces of the thorax are auscultated, as shown in Figure 2.4 (points 5 and 6 correspond to the lateral surfaces). It's important to avoid the scapula, as lung sounds cannot be heard through bone.

- During auscultation, at least one complete respiratory cycle should be listened to in each auscultated position to appreciate the sounds generated during both the inspiration and expiration phases.

- To differentiate between adventitious respiratory sounds and normal respiratory sounds, it is essential to identify the properties or attributes that characterize the auscultated respiratory sounds: frequency, pitch, intensity, sonority, timbre, duration, etc. The attributes mentioned earlier are detailed in Section 2.3.

2.2.1.1 Advantages and limitations

The process of auscultation has several advantages compared to other techniques for diagnosing the respiratory system, making it the initial clinical examination performed in any healthcare center for diagnosing the respiratory system [78]. The main advantages are as follows:

- It is an easy-to-use technique that any physician can perform. In contrast, other techniques, such as radiography, require a specialist (radiologist) to conduct the clinical examination, and the required instrumentation is more expensive.

- It is a quick technique, which helps avoid overcrowding at healthcare centers.

- It is a cost-effective technique, allowing it to be used worldwide regardless of the economic level. Any healthcare facility can have a stethoscope to perform

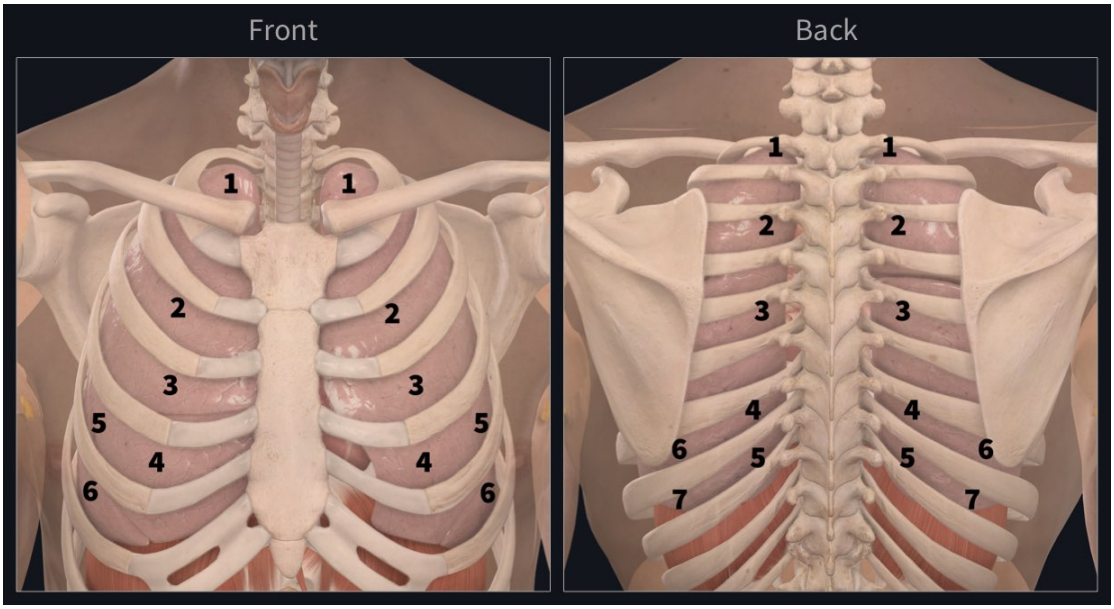


Fig. 2.4 Auscultation point, front and back of human body. Figure origin: `https://3d4medical.com/blog/auscultation-of-the-lungs`

the examination, whereas not all healthcare centers or countries can afford X-ray machines.

- It is a non-invasive and safe technique for both the patient and the specialist conducting the examination. In contrast, X-rays have various potentially harmful effects, including the risk of cancer, skin burns, the development of cataracts, and more.

- It is a technique that allows for the detection of adventitious sounds, which indicate the presence of obstructive respiratory pathologies, thus enabling a comprehensive diagnosis of the respiratory system. In this regard, the accuracy of adventitious sounds in identifying respiratory pathologies and assessing their severity has been widely confirmed [80, 73, 81, 82].

On the contrary, the auscultation process has several limitations, which constitute the main motivation of this Thesis. The drawbacks of this technique revolve around the cognitive capacity of physicians to accurately interpret the various biomedical sounds heard. The main limitations or disadvantages of auscultation are as follows:

- First, the diagnosis derived from auscultation remains highly subjective, conditioned by each physician's ability, experience, and training in listening to these sound signals. This subjective skill for analyzing the auscultated sound signals means that, on many occasions, the physician does not accurately determine the origin of the patient's potential illness. Therefore, the reliability of auscultation has been legitimately questioned due to the high variability in observations among specialists [73, 81]

- Furthermore, normal respiratory sounds and adventitious sounds (anomalous and indicative of a pulmonary disorder) are simultaneously mixed in time and frequency. This causes normal respiratory sounds to interfere with the listening of the adventitious sounds of interest, hindering the physician's cognitive ability by being distracted by these interfering sounds, which will likely result in an increase in false negatives in the diagnoses made. A recent study has shown that physicians fail to detect some adventitious sounds due to the overlap of respiratory sounds during inspiration and expiration [43].

- Likewise, the ambient noise surrounding the subject during auscultation causes interference with the listening of the adventitious sounds of interest, limiting the physician's cognitive ability [83].

- The capacity of the human auditory system is limited. The specialist physician must be able to analyze the properties that define the sound (frequency, amplitude, timbre, etc.) to determine the type of adventitious sound heard. However, discriminating between adventitious sounds with similar characteristics, such as monophonic and polyphonic wheezing sounds or fine and coarse crackles, is a fairly complex task to perform through auscultation [40].

- Lastly, it is common for the physician's cognitive capacity to decrease throughout the day, as the number of hours dedicated to analyzing respiratory sounds increases, a fact exacerbated by the stress that the physician experiences with certain medical cases [22, 23].

Taking advantage of the best of auscultation and considering the limitations it entails, the objectives of this Thesis have been defined. Due to these limitations, there is a need to design algorithms and methods that can address the drawbacks of auscultation to increase the efficiency of diagnoses, thereby preventing misdiagnoses that could jeopardize patients' health. The need for a complementary system to clinical decision-making for diagnosing potential pulmonary disorders in patients, applying eHealth, has become crucial in recent years to avoid compromising patient health, with the additional need to make healthcare systems sustainable and efficient [84, 85].

2.2.2 Types of stethoscopes

A stethoscope is composed of a variety of important components that allow it to transfer internal sounds from a patient's body to the ears of medical professionals so they can diagnose and treat the patient's medical condition. Regardless of the type of stethoscope, they are all characterized by the same basic design and the components that make it up. Using the most prominent company in stethoscope manufacturing as a reference

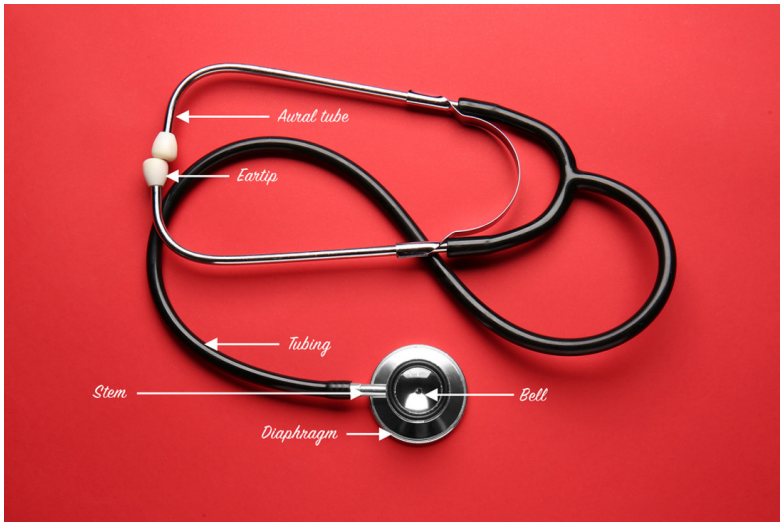


Fig. 2.5 Littman Sthetoscope. Figure origin: `https://www.stethoscope.com/blog/anatomy-of-a-stethoscope-everything-you-need-to-know-/`

(Littmann [72]), Figure 2.5 illustrates the main components of a stethoscope, which are:

- Metallic arch (Headset): it is the top part of the stethoscope that comes into contact with the specialist doctor's head and is composed of:
  - Eartips: these are the parts that come into contact with the doctor's ears, where they receive the auscultated biomedical sounds. Eartips are generally made of rubber or silicone and are designed to create a snug fit within the ears to isolate unwanted external sounds. These cushions should be comfortable to prevent doctors, who spend a lot of time auscultating various patients, from suffering added stress that could impair the diagnosis.
  - Eartubes: these are the metallic/steel parts of the stethoscope that connect to the eartips and the tubing. Eartubes are designed to isolate and transfer sound to the doctor's ears with minimal loss in the quality of the auscultated biomedical sounds. These tubes help separate sounds in the left and right channels to provide a better auditory experience, making it easier for the user to diagnose their patients' medical conditions.
- Hose or Tubing: these are flexible tubes generally made of PVC or neoprene. The purpose of the tubing is to transfer and relay the sounds that are captured by the diaphragm or bell and then send them to the eartubes. The walls of the tube are designed to prevent noise from mixing with other sounds. Additionally, some models have a dual lumen tubing that allows sounds to propagate more effectively and reach the specialist's ears more clearly.

- Base or Stem: it is the part of the stethoscope that connects the tubing to the auscultation receiver. In some models, it also allows the user to switch between the diaphragm and bell that make up the auscultation receiver.

- Chest-piece or Auscultation Receiver: it is the bottom part of the stethoscope that comes into contact with the patient's skin and is responsible for capturing internal body sounds. In a standard design, it consists of:

  - Diaphragm or Membrane: is a flat, circular, and often larger surface located on one side of the chest piece. It is usually the larger of the two sides and is designed to be placed directly on the patient's skin or clothing. The diaphragm is responsible for transmitting high-frequency sounds, such as normal heart sounds (S1 and S2) and breath sounds. Here's how the diaphragm works:

    * Surface Contact: To listen to sounds, the healthcare professional places the diaphragm against the patient's body, usually over the area of interest (e.g., chest, heart, or abdomen).
    * Sound Transmission: When positioned correctly, the diaphragm picks up vibrations caused by the body's internal sounds. These vibrations travel through the chest piece and the tubing to reach the healthcare professional's ears.
    * Amplification of High-Frequency Sounds: The diaphragm is particularly adept at capturing high-frequency sounds. As a result, it is especially useful for detecting sounds like heart murmurs, breath sounds, and certain bowel sounds.

  - Bell: is the smaller, usually concave, and bell-shaped side of the chest piece, which is the part that comes into direct contact with the patient's body. The stethoscope typically has two sides: the bell and the diaphragm. Each side is designed to pick up different frequencies of sounds, allowing healthcare professionals to assess a variety of bodily sounds during physical examinations. Here's how the bell of the stethoscope works:

    * Low-Frequency Sounds: The bell is specifically designed to capture low-frequency sounds. These include certain heart murmurs, certain bowel sounds, and some vascular sounds.
    * Proper Placement: To use the bell, the healthcare professional places it lightly on the patient's skin or clothing, using minimal pressure. It is essential to ensure a good seal between the bell and the patient's body to prevent the escape of sound.

Fig. 2.6 Double Head Sthetoscope. Figure origin: `https://www.mdfinstruments.es/products/dual-head-blackout-black-stethoscope`

* Sound Transmission: When positioned correctly, the bell captures the low-frequency vibrations generated by internal body processes. These vibrations travel through the chest piece and the tubing to reach the healthcare professional's ears.

Switching between Bell and Diaphragm: Some stethoscopes have a dual-sided chest piece, allowing the user to switch between the diaphragm and the bell. To listen to low-frequency sounds, the healthcare professional will use the bell by lightly placing it on the patient.

After presenting the components of a stethoscope, the rest of the section is devoted to describing the most relevant aspects of the types of stethoscopes. Currently, a wide range of stethoscopes is available on the market, and although they follow the standard structure described earlier, they can be classified into various types based on certain factors. For example, some stethoscopes are tailored to specific patient types or medical specialties, including fetal, neonatal, pediatric, nursing, pulmonology, cardiology, veterinary, etc. In other cases, they can be classified based on the components included in the chest-piece (auscultation receiver): single head (diaphragm or bell, see Figure 2.5), double head (the standard composed of both diaphragm and bell, see Figure 2.6), and triple head (the least common, composed of three pieces that cover a wider frequency range, see Figure 2.7). However, the classification shown below is more generic, consisting of two types of stethoscopes: acoustic stethoscopes and digital stethoscopes.

Traditional or conventional acoustic stethoscopes:

Within this type (see Figure 2.5), all stethoscopes that are not composed of a digital system are included. These are the traditional stethoscopes that have existed since the invention of the first stethoscope and have the most functional limitations in the quality of the auscultated biomedical sound. An acoustic stethoscope works by channeling and

Fig. 2.7 Triple Head Sthetoscope. Figure origin: `https://www.frafito.net/stethoscopes/670-stethoscope-harvey-deluxe-triple-pavillon.html`

directing the majority of the relevant sound waves toward the ears. For us to hear a sound, sound waves must cause vibrations in air molecules, resulting in changes in air pressure that cause our eardrums to vibrate in turn. Internal sounds of the body, such as heartbeats or airflow in the airways, generate sound waves that strike the metal piece (diaphragm or bell) of the stethoscope when it is placed on a patient. Subsequently, the rubber hose channels these sound waves in a specific direction until they strike the earpieces of the stethoscope and, finally, reach the ears through the ear tips. Because sound waves are transmitted through the hose, they reach the ears with greater intensity. This is why when listening to a patient's heart with a stethoscope, it sounds louder than if you were to place your ear directly next to their chest. The main limitations of this type of stethoscope revolve around the quality of the captured sound and the limited amplification of sounds. It has even been shown that both the bell and diaphragm have significant attenuation above 200 Hz, which limits the ability to discern sounds in this frequency range [38]. Although the human ear's sensitivity ranges from 20 to 20000 Hz, the ear follows a logarithmic sensitivity to frequency, requiring greater changes in higher frequencies to distinguish them as different [86]. Therefore, the limitation of acoustic stethoscopes in amplifying the captured sounds is the main disadvantage. On the other hand, compared to digital stethoscopes, another disadvantage is the inability to capture the sound signal in digital format for further analysis. Lastly, and briefly, the main advantages of this type of stethoscope are: i) lower cost compared to digital ones; and ii) they do not require batteries to operate. The most common traditional analogic sthetoscope brands are: Littmann, Spirit, MDF, ADC, Premium y Rappaport.

Electronic or digital stethoscopes:

The advancement of technology led to the emergence of electronic or digital stethoscopes. Unlike acoustic stethoscopes, electronic stethoscopes capture the physical vibrations of sound, translate them into an electrical signal, and optimize it to improve the audio quality of the auscultated sounds and, as a result, the diagnosis. Electronic stethoscopes use a variety of sensors, including condenser microphones and piezoelectric sensors, to convert acoustic waves into electrical signals for filtering and processing [87].

Some of the advantages supported by electronic stethoscopes include [88]:

- Amplification of relevant sounds.
- Adjustable frequency range.
- Noise reduction.
- Recording and playback.
- Connectivity with external devices (smartphone, tablet, or computer) via Wi-Fi or Bluetooth.

- Some models include a screen for visualizing the spectrogram of what is being captured.

  On the other hand, the main disadvantages that can be highlighted are:

- Greater weight.
- External power source (batteries).
- Higher cost than acoustic stethoscopes.
- Electronic components may be damaged if not properly maintained.
- Potential interference from other electronic devices.

It's worth noting that most electronic stethoscopes are designed like traditional stethoscopes, but there are some models (such as the Thinklabs One digital stethoscope [89]) that consist of just a chest-piece connected directly to headphones. Considering everything mentioned earlier and focusing on the objectives of this Thesis, it's important to highlight that this type of stethoscope enables the development of algorithms and methods to assist in the diagnosis of respiratory pathologies, as they transform acoustic waves from inside the body into sound signals that can be digitally processed.

Next, a comprehensive analysis of a set of electronic stethoscopes, from the most reputable companies recommended by specialists in the field of auscultation, is described.

As mentioned earlier, electronic stethoscopes, unlike acoustic stethoscopes, allow for the capture of sound signals from inside the body for subsequent processing. This is why electronic stethoscopes receive special attention for the development of algorithms or methods to enhance respiratory diagnosis. The objective of this section is to identify the most relevant electronic stethoscopes available today. To achieve this, a comprehensive analysis has been conducted on the most reputable companies recommended by specialists in the field of auscultation. Among the most prestigious companies, the following stand out: 3M Littmann [90], Thinklabs [89], Eko [91], and Ekuore [92]. Below, for each company, the electronic stethoscope model that offers the best features and performance is described.

3M Littmann:

3M Littmann is considered the pioneer company in the field of auscultation, offering throughout its history the most potent and relevant stethoscopes. Currently, it is the company with the longest history and the most highly regarded by medical professionals. The latest model developed is called the 3M Littmann Electronic Stethoscope Model Core (see Figure 2.8), and its main features are:

- It connects to the Eko software to visualize and share heart sound waves. Gain of up to 40x (at maximum frequency, compared to analog mode).

Fig. 2.8 Examples of digital sthetoscopes: top-left - 3M Littmann Core, top-right - Thinkslab, bottom-left - Eko, bottom-right - Ekuore

- Active noise cancellation reduces unwanted background noise.
- Switch between analog listening mode and amplified mode.
- Soft-sealing ear tips provide excellent acoustic sealing and a comfortable fit.
- Adjustable double-sided stainless steel chest piece with open or closed bell.
- Suitable for use in adult and pediatric patients. FDA approved and HIPAA compliant.

Thinklabs:

However, nowadays, other companies have managed to position themselves with innovative proposals in the electronic auscultation market. Thinklabs was founded in 1991 by Clive Smith, a Caltech graduate in Electrical Engineering with a passion for medical electronics, sound, music, and signal processing. True to their motto, ”think

deeply about the problems that matter and develop imaginative solutions," their innovative designs and product efficiency have made Thinklabs a worthy competitor to the pioneer 3M Littmann. Their electronic stethoscope, considered the best in the market, is known as Thinklabs ONE (see Figure 2.8), and its main features include:

- Innovative design that breaks away from the familiar aesthetics of traditional stethoscopes. Specifically, it consists only of the chest piece, which connects directly to headphones.
- Considered the smallest and most powerful stethoscope in the market (fits in the palm of your hand).
- Amplification of auscultated sounds by a factor of over x100. It is the electronic stethoscope on the market that achieves the greatest sound amplification.
- Recording and storing of sounds for later analysis and processing using Thinklabs' proprietary application.
- It does not come with built-in wireless connectivity, but it has a mobile kit that allows it to be connected to any mobile device with iOS, Android, or Windows.
- It does not have an ambient noise reduction system. This results in both the auscultated sounds and the ambient noise surrounding the patient being amplified.

Eko:

On the other hand, it is worth mentioning the company Eko for the diversity of products it offers. Eko made a mark in the market with the design of a device called the CORE Digital Attachment, which can transform an acoustic stethoscope into an electronic one simply by placing it between the chest piece and the tubing. Additionally, they have several models of electronic stethoscopes. The most versatile and powerful one is called the CORE Digital Stethoscope (see Figure 2.8), and its main features include:

- Innovative technology for active background noise cancellation.
- Amplification of auscultated sounds by a factor of up to x40.
- Connectivity with external devices via Bluetooth. Compatible with any device running iOS, Android, or Windows.
- Incorporates artificial intelligence applied to heart sounds. It has become the first electronic stethoscope capable of detecting heart murmurs.
- Includes software that, among other functions, allows real-time display of the phonocardiogram.

- Recording and storage of auscultated sounds for later analysis and processing in tracks of up to 120 seconds in duration.
- Platform to facilitate the management of auscultated files for work from home utilities.

Ekuore:

Finally, undoubtedly the riskiest and most innovative proposition comes from the company Ekuore. Their innovation represents a significant evolution in the field of auscultation. Specifically, they have designed a device that they define as the first smart electronic stethoscope, allowing for remote auscultation. This device enables the remote listening of auscultated sounds, without the need for the physician to be in the same room as the patient. The device transmits the auscultated sounds to any device (smartphone, tablet, or computer) via WiFi for listening and viewing without the need for physical contact with the patient. This device is known as the Electronic Stethoscope eKuore Pro (see Figure 2.8), and its main features are described below:

- Comprised of a single device responsible for capturing and transmitting auscultated sounds.
- Considered the first stethoscope designed for telemedicine, introducing the era of remote auscultation. Through an app developed by eKuore, a physician can listen to a patient remotely. For example, this is advantageous in the monitoring or continuous monitoring of chronic patients, as the patient can use the device to auscultate themselves and then send the results to the physician for analysis, thus avoiding unnecessary hospital visits. Additionally, in cases involving patients with contagious diseases (COVID-19), direct contact can be avoided to reduce the risk of transmission.
- The component that comes into contact with the patient is disposable, reducing the risk of potential infections.
- Incorporates predefined filters to improve cardiac and pulmonary signals. Amplification of the auscultated sounds by a factor of up to x20.
- Connectivity with external devices via WiFi. It also allows for wired or Bluetooth headphones to be connected.
- Includes an app to control various storage and recording options. Furthermore, the app offers other capabilities, such as visualizing the phonocardiogram or even editing recorded sound signals (`https://apps.apple.com/es/app/ekuore/id672630692` for apple users and `https://play.google.com/store/apps/details?id=com.kukupia.app.ekuorepro&hl=es&gl=US` for android users).

2.2.3 Alternatives to the auscultation process for the diagnosis of respiratory pathologies

There is a wide range of alternative procedures to auscultation that help determine the condition of the human respiratory system and diagnose potential pathologies. Without diminishing the auscultation process, as it forms the basis of this Thesis, it is considered interesting to introduce alternative techniques for analyzing respiratory biomedical sounds used to diagnose respiratory pathologies. A classification of various alternative procedures for examining patients has been made. These procedures can be grouped into three categories: laboratory methods, respiratory function tests, and imaging techniques.

Laboratory methods [93]:

In addition to routine blood and urine laboratory analyses, there are several specific tests available to help determine potential specific respiratory pathologies (for example, asthma can be detected with a test that measures immunoglobulin E levels, denoted as IgE). This group includes:

- Microbiological tests: they play an essential role in the investigation of infectious respiratory diseases caused by viruses, bacteria, fungi, or parasites.
- Histological and cytological examinations: they play a fundamental role in the diagnosis of many malignant and benign respiratory diseases, including infections. Besides sputum, which can be examined cytologically, some samples are obtained through various biopsy techniques, which are examined later and sent for histological and/or cytological evaluation.

Respiratory function tests [93]:

These types of tests focus on evaluating the proper functioning of the respiratory system during breathing mechanics and gas exchange that occurs in the lungs. This group includes:

- Spirometry: involves instructing the patient to exhale all the air from their lungs after a maximal inhalation, for as long as needed, using a specific device (see Figure 2.9). Spirometry assesses the patient's lung capacity and is a key test for identifying the presence of asthmatic episodes. Specifically, these tests analyze the volume of exhaled air and are used to measure the effect of bronchodilator medications on the reversibility of obstruction, as well as to determine responsiveness to bronchial provocation tests.
- Lung capacity and airway resistance: total lung capacity can be determined using gas dilution techniques or body plethysmography. The latter method also allows for measuring airway resistance. Forced oscillation technique, which measures

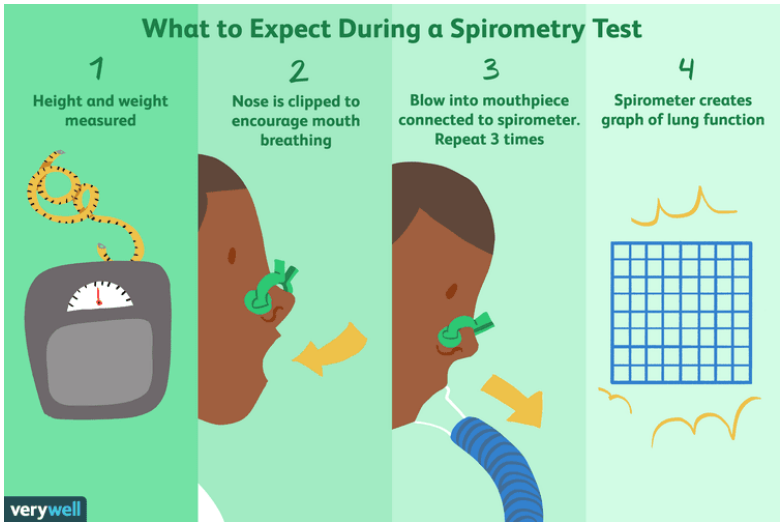


Fig. 2.9 Spirometry process. Figure origin: `https://www.verywellhealth.com/asthma-and-spirometry-200531`

the resistance of the entire respiratory system, has the advantage that the patient does not need to perform specific breathing maneuvers.

- Diffusing capacity: the capacity for carbon monoxide diffusion in the lung (also known as transfer factor) is usually performed with a test over a single breathing cycle and measures the overall gas exchange function of the lung.

- Arterial blood gas analysis: this is one of the most useful diagnostic tests and involves measuring the amount of oxygen and carbon dioxide in the blood. This examination also determines the blood's acidity (pH).

- Cardiopulmonary exercise tests (CPET): evaluate the function of the heart, circulation, respiration, and muscle metabolism at rest and under increasing physical exertion, up to the maximum possible load. Simultaneous measurement of oxygen and carbon dioxide concentrations in inhalation and exhalation allows the determination of how much oxygen is inspired (VO_2) and how much carbon dioxide (VCO_2) is exhaled.

- Measurement of respiratory muscle function: this is commonly evaluated by measuring the maximum pressures generated in the mouth during inhalation and exhalation while the airway is occluded.

- Diagnosis of sleep-related respiratory disorders: this is usually performed through polysomnography. It involves recording brain activity, breathing, heart rate, muscle activity, and blood oxygen levels while sleeping.

Imaging Techniques [93]:

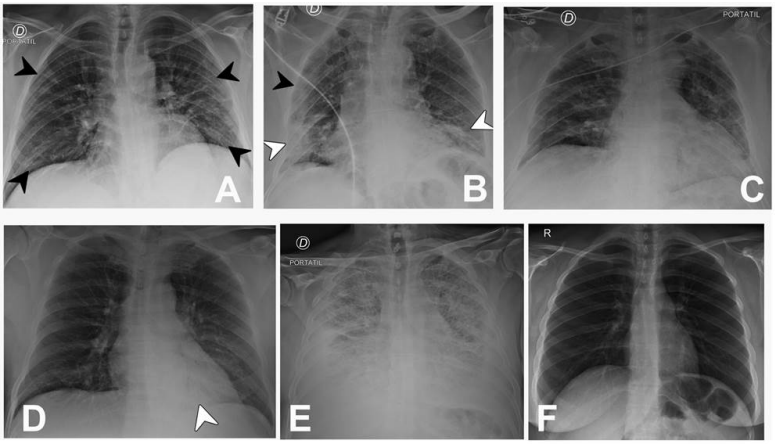


Fig. 2.10 Examples of chest X-rays for each pattern of the BSTI structured reporting system in COVID-19. Classic/probable pattern (A-B): Patient 1 (A) Bilateral peripherally and lower-distributed ground-glass opacities (black arrowheads). Patient 2 (B): Bilateral ground-glass opacities (black arrowheads) associated with multiple consolidation foci in the described distribution (white arrowheads). Indeterminate pattern (C): Diffuse ground-glass opacities with no lower or peripheral predominance. Non-COVID-19 pattern (D-E): Patient (D) with retrocardiac unifocal consolidation consistent with bacterial pneumonia. Patient (E) with signs of diffuse bilateral interstitial and alveolar edema associated with bilateral pleural effusion consistent with decompensated heart failure. Normal pattern (F): Examination without radiological findings suggestive of pneumonia in a patient with COVID-19 confirmed by PCR test. It is relevant to mention that this pattern does not rule out the presence of disease.. Figure origin: `https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0717-93082020000300088`

These are techniques that are based on the analysis of images representing the condition of the human respiratory system. This group includes:

- Chest X-ray (Radiography): it is an essential part of diagnosing and monitoring respiratory pathologies (see figure 2.10). It is the first step in the radiological evaluation of patients with suspected respiratory diseases. Modern digital radiography offers high image quality and the possibility to reduce radiation doses.
- Chest Computed Tomography (CT): is the second most important radiological modality in respiratory medicine, allowing for much more detailed visualization of thoracic structures compared to X-rays.
- Pulmonary and Bronchial Angiography: these are invasive techniques for obtaining images of blood vessels and are only used if less invasive techniques (CT/magnetic resonance with contrast) fail or need confirmation.
- Fluoroscopy: is an X-ray technique that directly visualizes respiratory movement. It is mainly used for guiding biopsies in peripheral lung lesions and for the differential diagnosis of a raised diaphragm.

- Magnetic Resonance Imaging: has the advantage of avoiding radiation. It is primarily useful when there is suspicion of tumor invasion in the mediastinum and the chest wall.

- Ultrasonography: has become an important imaging technique. Its advantages include no radiation, low cost, and mobility. It is mainly used in the investigation of pleural effusions but also in pleural thickening, chest wall abnormalities, diagnosing pneumothorax, and biopsies of lesions adjacent to the chest wall.

- Nuclear Medicine Techniques: These primarily include the Pulmonary Ventilation/Perfusion Scan to measure both breathing and circulation in all areas of the lungs. It is mainly used in diagnosing pulmonary embolism.

2.3 Classification of respiratory sounds

Respiratory sounds refer to the set of sounds generated by the airflow through the different airways or passages that compose the human respiratory system during the mechanics of respiration (the process of inhalation and exhalation) [94]. This section will provide a classification of the different types of respiratory sounds that can be produced by the respiratory system. As shown in Figure 2.11, the different types of respiratory sounds can be grouped into two sets differentiated by the patient's condition (a healthy patient or a patient with a specific respiratory condition). On one hand, normal respiratory sounds refer to the sounds produced by the respiratory system during the breathing cycle. These sounds are always present during a person's breathing, regardless of the presence or absence of a specific pathology. As we will see in Section 2.3.2, normal respiratory sounds can be classified based on where they are generated within the network of airways that make up the respiratory system. On the other hand, adventitious sounds, also known as abnormal or incidental sounds, appear in patients with a respiratory pathology. Therefore, identifying acoustic bio-markers is a vital task to detect abnormalities in the proper functioning of the respiratory mechanics caused by potential respiratory disorders.

In 1816, when Laënnec initiated the process of auscultation, popularizing the listening of respiratory sounds, there was no standardized classification of the different types of respiratory sounds. Although in recent decades, a more standardized classification has been established, clearly distinguishing the different respiratory sounds (as in Figure 2.11), there is still a need to define a single classification that encompasses all types of sounds present during a person's breathing and their respective characteristics. The main challenge lies in the time-frequency characterization of different types of respiratory sounds.

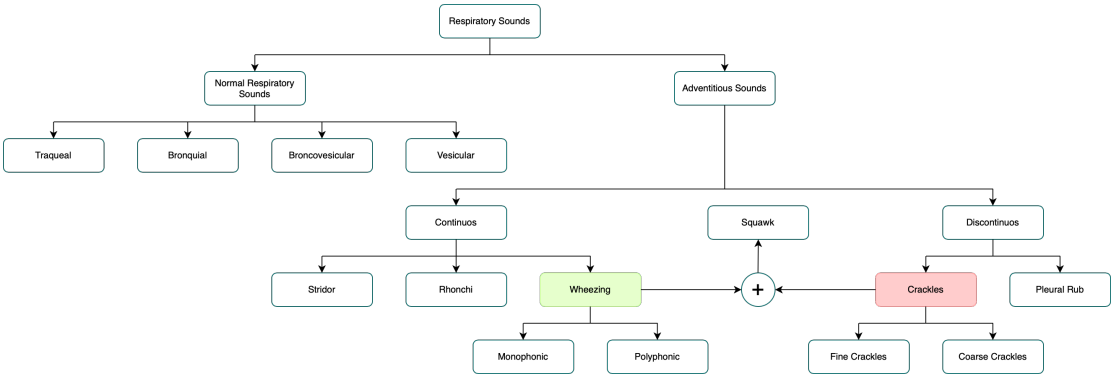


Fig. 2.11 Classification of respiratory sounds. Note that the adventitious sounds analysed in detail in this Thesis are remarked in pink and light green colour.

Although cardiac sounds are not the subject of study in this Thesis, it is important to consider the interference that respiratory sounds may experience due to these sounds. This interference occurs because the auscultation areas for respiratory and cardiac sounds are very close, resulting in spectral overlap between both types of sounds. Specifically, the dominant frequency of all respiratory sounds typically ranges from 60 to 1000 Hz [95], as we will see throughout this section. However, the dominant frequency of cardiac sounds is usually below 100 Hz [21, 96, 97, 98, 24]. The difference in frequency ranges makes it easier to filter cardiac sounds from respiratory sounds. In general, a high-pass filter is often applied above 100 Hz to eliminate significant heart noise as well as electrical interference [21, 96, 98, 99, 24]. Although this procedure is common for removing interfering cardiac sounds, other studies propose more robust methods that assume the dominant frequency of cardiac sounds is usually below 320 Hz, leading to greater spectral overlap between cardiac and respiratory sounds between 60 and 320 Hz [100, 101, 102, 103, 104, 105]. Specifically, in this section, we will first detail the nature of respiratory sounds and how they are generated, as well as the characteristics or sound properties necessary to model and distinguish the various types of respiratory sounds. Second, we will present the different types of normal respiratory sounds. Finally, we will provide a detailed explanation of the various types of adventitious sounds, with an emphasis on wheezes and crackles, the subject of study in this Thesis. To tackle this task, we will use some of the most relevant works in the field of respiratory sound classification [21, 96, 97, 24, 99, 78, 73, 76, 106, 75, 107, 108, 109].

2.3.1 Characteristics of respiratory sounds

Before discussing respiratory sounds, it is useful to elaborate on the nature of sound itself and how it is generated. The flow of air through the respiratory pathways of the bronchial tree causes turbulence, resulting in the vibrations we perceive as respiratory sounds. Turbulence occurs in areas where air velocity is high and in those with geomet-

ric conditions that hinder laminar flow. Specifically, turbulent airflow is disorganized and chaotic by nature and occurs when high-velocity airflow passes through a large-diameter respiratory passage, especially one with irregular walls. This primarily occurs in the trachea and at the bifurcations of the primary, lobular, and segmental bronchi. However, only turbulent airflow is responsible for the production of respiratory sounds [110]. In contrast to turbulent airflow, laminar flow occurs in low-flow situations and is silent. Specifically, laminar flow has a parabolic shape, as air in the central layers moves faster than air in the peripheral layers, with little to no transverse flow. In the airways closer to the alveoli, airflow is laminar, and thus there is no turbulence, and no respiratory sounds are produced. For a more detailed explanation, Sarkar [78] provides a helpful description of the production of respiratory sounds, including that the laminar flow pattern can be modeled with the Poiseuille equation.

On the other hand, variations or distortions that occur during the propagation of respiratory sounds are linked to several factors [21]. First, there is the acoustic response of the stethoscope or sensors used for auscultation. Second, the asymmetry of sounds present in auscultation, which may indicate the presence of a pathology, is mainly due to airway obstruction. Lastly, the heterogeneous composition of the body can act as a filter for these sounds. Specifically, the human thorax is composed of four different types of materials with significantly different acoustic properties: hard tissue (bone), soft tissue (muscle, fat, etc.), air in the major conductive airways of the bronchial tree, and parenchymal tissue, which is a heterogeneous mixture of soft tissue and air found in alveolar sacs and smaller bronchioles. The characteristics of these different components affect how sounds are transmitted through the chest during auscultation. Specifically, sound inside the airways experiences frequency-dependent absorption in the airway walls and the surrounding parenchymal tissue. It has been shown that high-frequency sounds propagate further within the branching structure of the airways, while low-frequency sounds tend to couple to the airway walls earlier. However, due to the attenuation of high-frequency sounds in the surrounding parenchymal tissue, most of the signal energy of recorded respiratory sounds on the torso's surface is concentrated in the lower frequencies. As a result, the analysis of sound transmission in the thoracic cavity suggests that the thorax, in general, acts as a low-pass filter, absorbing higher frequencies as sound travels through the bronchial tree. Consequently, depending on the auscultation points, the frequency range and intensity of sounds will vary. As will be seen in the next section, tracheal and bronchial respiratory sounds have higher frequency and intensity than bronchovesicular and vesicular respiratory sounds.

The physics of sound plays a crucial role in discriminating or identifying the different types of respiratory sounds that may appear during auscultation. Furthermore, the properties or attributes of sound can describe what a human experiences when listening to a particular sound. Specifically, the objective characteristics or attributes of a sound

are frequency, intensity, and duration. On the other hand, the main subjective properties of a sound are pitch, loudness, and timbre. Below, these characteristics are defined according to respiratory sounds [78], and a relationship is established between objective characteristics (which define sounds with objective values) and subjective characteristics (which describe how the sound is perceived by the physician).

Frequency and Pitch: On one hand, frequency is an objective characteristic that measures the number of oscillations or vibrations per unit of time, in cycles per second, expressed in hertz (Hz). Therefore, higher frequency means a greater number of vibrations, and vice versa. As mentioned earlier, in larger airways (trachea and main bronchi), there are more vibrations (due to higher turbulent airflow), than in narrower airways (secondary bronchi and bronchioles). Thus, as will be seen in the next section, tracheal and bronchial respiratory sounds span a wider frequency range than bronchovesicular and vesicular respiratory sounds. On the other hand, pitch is a subjective characteristic that represents how "high (sharp)" or "low (deep)" a sound is perceived. It is directly related to frequency; the higher the frequency, the higher the pitch, and vice versa. This is why tracheal or bronchial respiratory sounds are perceived as higher-pitched than the rest. This characteristic is considered one of the most important for distinguishing between all types of respiratory sounds, both normal and adventitious. For example, each type of adventitious sound has a characteristic frequency range. In the case of wheezes, they can be characterized by a pitch situated between 100 and 1000 Hz and in the case of crackles it is between 380 and 1800Hz. this can make it easier to understand the relationship between frequency and pitch. Note that although not all respiratory sounds can be modeled as a pitch, it is common to use this subjective metric to understand how high or low the human ear perceives the sound based on its frequency. Specifically, the human ear can perceive sound waves in a wide frequency range, ranging from 20 to 20000 Hz.

Intensity and Loudness: Intensity is an objective characteristic related to the energy of sound waves and measures the average flow of energy per unit area perpendicular to the direction of propagation. On the other hand, loudness is a subjective characteristic that allows ordering sounds on a scale from softest to loudest in terms of intensity, meaning from sounds that are perceived as quieter to sounds that are perceived as louder. Loudness depends not only on the power of a sound but also on its duration and its spectral-temporal structure. Loudness is actually a more complex characteristic to standardize because it is determined by the initial source producing the sound, the amplitude of the vibrations, the distance vibrations travel, and the material they pass through afterward. This explains why some lung sounds are perceived as strong, as in a consolidated lung, or soft, as in a lung filled with emphysematous bullae. In general terms, sounds with higher intensity are perceived as louder than sounds with lower intensity. However, two sounds with the same intensity will not sound equally loud if

their frequencies are different. In 1933, Fletcher and Munson [111] determined that the human ear has its highest sensitivity between 2000 and 5000 Hz. The farther a frequency is from the previous spectral range, the higher the intensity needed to hear respiratory sounds.

Duration: Duration is an objective characteristic and refers to the time interval during which a specific event is active in the audio signal. This parameter is associated with the terms onset and offset, indicating the moment in time when the event begins and ceases to be active, respectively. The duration of vibrations determines whether the physician's ear perceives the sound as being of long or short duration. This metric, for example, allows distinguishing prolonged expiratory wheezes in a patient with COPD and at the same time it make is more complicated to determine whether a crackle is present or not due to its short duration. Duration is also considered one of the most relevant characteristics for distinguishing between different types of adventitious sounds that can occur. For instance, wheezing sounds are characterized by having a longer duration than crackling sounds.

Timbre: Timbre is a subjective characteristic that refers to the quality of the sound heard by the physician and depends mainly on the frequency components that characterize a particular sound. This property allows distinguishing two sounds with the same pitch and intensity. Specifically, wheezing sounds are often composed of multiple frequency components. The fundamental frequency or primary frequency is the lowest of the sound wave and determines the pitch of the sound. Frequencies higher than the fundamental are called secondary frequencies or harmonics. Depending on the location of these secondary frequencies, wheezing can be classified as monophonic or polyphonic. If these frequencies are harmonically related to each other (integer multiples of the fundamental frequency), wheezing is monophonic, and otherwise, it is polyphonic. Therefore, depending on the location of these secondary frequencies (also called partials), the timbre will vary, allowing differentiation of a sound with the same pitch and intensity. Thus, timbre can be thought of as the "musical" characteristics of breath sounds. In general, breath sounds are considered very complex mixtures of sounds of different frequencies, giving them a characteristic timbre that can be altered by various pathological conditions. For example, this allows the physician to distinguish between different sounds produced in the chest, such as vesicular breath sounds generated in normal lung tissue, and bronchial breath sounds transmitted through a consolidated lung.

In the processing of biomedical respiratory sounds, there are several methods of analysis for extracting the time-frequency features of respiratory sounds recorded during auscultation. These analysis techniques are fundamental for the classification and detection of different adventitious sounds (wheezing, crackles, etc.). Specifically, the magnitude spectrogram of respiratory sounds can be obtained using the magnitude of a variant of the Fourier transform called the "Short-Time Fourier Transform (STFT),"

applied with a specific window function (Hamming, Hann, etc.). This type of time-frequency representation has been widely used in the field of biomedical respiratory sounds [100, 96, 112, 113, 114] because it allows extracting the time, frequency, and intensity of different input sounds. It's worth noting that this type of representation has been used throughout this section to display the different types of respiratory sounds. Specifically, the parameters of the STFT and the sampling frequency have been adjusted in each representation to improve the visualization of each sound type.

In addition to the sound characteristics mentioned above, the development of an experimental and empirical study of different types of respiratory sounds has led to the definition of a series of more specific properties aimed at facilitating the distinction between the various sounds present.

Temporal Continuity/Discontinuity: This property is directly related to the duration of a particular event. Temporal continuity allows modeling sounds whose energy varies slowly over time (this concept is also called temporal smoothness), as is the case with continuous adventitious sounds (for example, wheezing). In Figure 2.12, you can observe how wheezing sounds are continuous over time. On the other hand, temporal discontinuity refers to sounds that occur intermittently or have an short duration (this concept is also known as temporal dispersion). This property models the behavior of discontinuous adventitious sounds (such as crackles). Therefore, distinguishing between these properties allows for a different modeling of wheezing sounds compared to crackles.

Spectral Smoothness/Dispersion: This property is directly related to the distribution of energy across the spectral range (frequency range). Spectral smoothness characterizes sounds that are smooth in frequency, meaning their energy varies slowly across the spectral range. This behavior generally occurs in normal respiratory sounds, which can be modeled as a wideband spectrum, as shown in Figure 2.12. On the other hand, spectral dispersion characterizes sounds that are spread out in frequency, meaning they can be modeled as narrowband spectral peaks. Therefore, distinguishing between these properties allows for differentiation between wheezing sounds compared to crackles.

Temporal Repetition: This property allows differentiation between events that repeat over time. For example, normal respiratory sounds can be considered repetitive patterns over time. During the mechanics of respiration, the processes of inhalation and exhalation occur repeatedly over time. However, adventitious sounds in general and wheezing and crackling sounds, in particular, cannot be modeled based on this property. Specifically, crackles may be present in certain stages of the respiratory signal and absent in others due to the unpredictable nature of pulmonary disorders. Figure 2.12 shows the behavior of normal respiratory sounds and wheezing sounds considering this property.

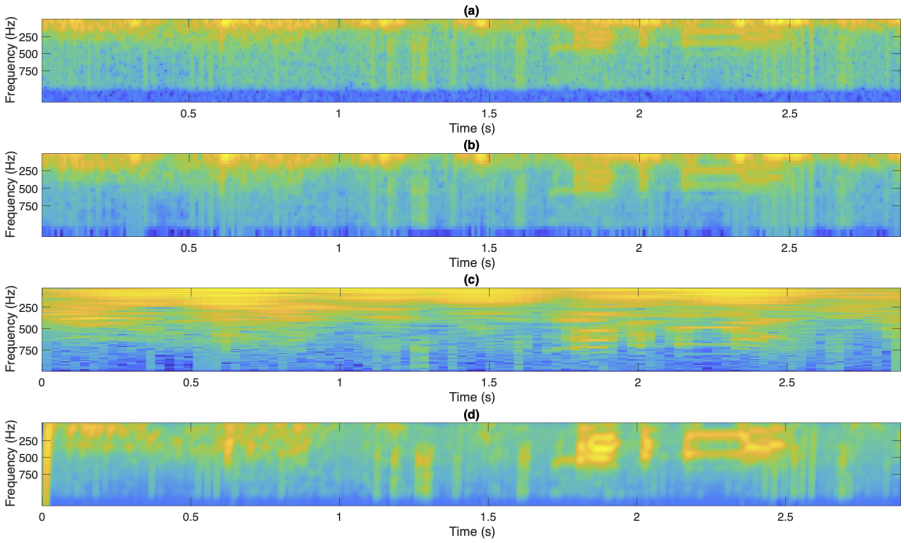
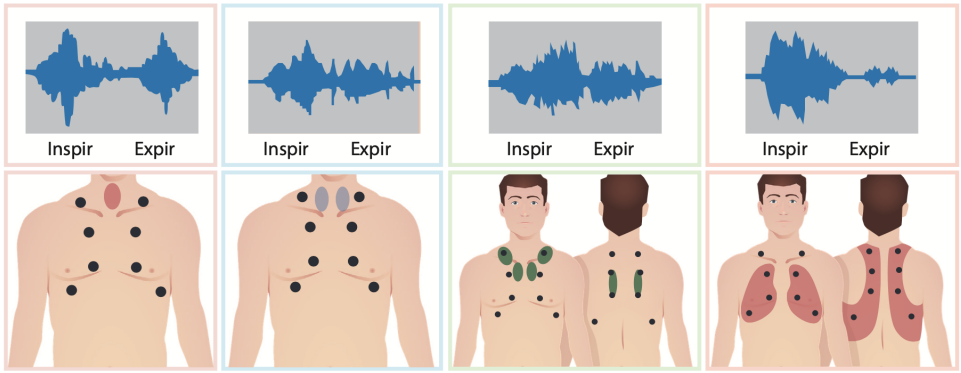


Fig. 2.12 Magnitude, in logarithmic scale, of the time frecuency representations analyzing a respiratory cycle with a time duration of 2.9 seconds associated to the patient number 103 from the Respiratory Sound Database of the International Conference on Biomedical Health Informatics (ICBHI). The respiratory cycle is composed by one wheeze sound located in the temporal range [2.1-2.6] seconds. STFT spectrogram (a), Mel-scaled spectrogram (b), Constant-Q (c) and cochleogram (d). Wheezing sound detected in range [2.2-2.4] seconds.

2.3.2 Normal respiratory sounds

As mentioned earlier, respiratory sounds are produced as a result of air flowing through the lungs and are classified as normal or abnormal (adventitious). Normal respiratory sounds are defined as those produced in healthy airways during unforced physiological breathing. In general, normal respiratory sounds can extend in frequency up to 5000 Hz [115], but most of the energy is between 60 and 1000 Hz [95]. Specifically, their frequency behavior can be characterized as a broadband spectrum, where the energy change is smooth. Normal respiratory sounds can be divided into respiratory cycles, and each respiratory cycle consists of the inspiration and expiration phases. The duration of the phases varies depending on the auscultated airways. Additionally, the intensity and frequency range of normal respiratory sounds vary throughout the auscultation of the bronchial tree. The sounds with higher intensity and a wider frequency range occur in the trachea. However, these parameters decrease along the bronchial tree until, ultimately, very close to the alveoli, there is laminar flow that does not produce sound. In Figure 2.12 a respiratory cycle with a time duration of 2.9 seconds is shown.

Normal respiratory sounds are typically classified into four types based on the auscultation area [116, 75, 107, 73, 24, 21, 78, 77]: tracheal, bronchial, bronchovesicular, and vesicular. As shown below, the frequency, intensity, and duration characteristics of



Lung sound	Tracheal	Bronchial	Bronchovesicular	Vesicular
Location best heard	Trachea	Large airways in the anterior chest near the second and third intercostal spaces	In the posterior chest between the scapulae and in the center of the anterior chest	Most lung spaces
Sounds like	Harsh, like air in a pipe	Loud and high in pitch	Softer than bronchial sounds with a tubular quality	Soft, blowing, or rustling sounds
Lung sound caused by	Normal breathing through the trachea	Normal breathing	Normal breathing	Normal breathing

Fig. 2.13 The different types of lung sounds can be heard best in the following locations: Bronchial lung sounds, Tracheal lung sounds, Bronchovesicular lung sounds, Vesicular lung sounds. Figure origin: `https://www.lecturio.com/nursing/free-cheat-sheet/charting-lung-sounds-normal-findings/`

each stage (Inspiratory-Expiratory ratio) vary across the four types of normal respiratory sounds. As an illustrative representation, Figure 2.13 provides a visual classification of the different types of normal respiratory sounds, taking into account the auscultation areas as well as the intensity and duration of each respiratory phase.

Tracheal breath sounds are rough, very loud, and high-pitched sounds heard over the trachea (above the subclavicular notch). The expiratory phase is longer than the inspiratory phase. Specifically, the Inspiratory-Expiratory (I:E) ratio is typically (1:2) or (1:3). There is also a pause between the two phases. The typical frequency range of tracheal breath sounds varies from 100 to 1,500 Hz, with a sharp drop in power above a cutoff frequency of around 800 Hz. The frequency range of tracheal sounds is much broader than that of vesicular or pulmonary sounds, extending up to 5000 Hz.

Auscultation over the trachea is not routine but can be useful in specific conditions. It has a hollow and tubular timbre, making it a good model for studying bronchial breath sounds. Additionally, some adventitious sounds, such as stridor, may be mixed with normal tracheal respiration. Lastly, the analysis of tracheal sounds is also useful for monitoring patients with sleep apnea-hypopnea syndrome (AHS) [117].

Bronchial breath sounds are strong and high-pitched, similar to tracheal sounds but with slight variations. These sounds are heard over the sternum (where the trachea divides into the primary bronchi). Like tracheal sounds, there is a pause between the inspiratory and expiratory phases, and the I:E ratio is typically (1:2) or (1:3). They are higher-pitched and louder than breath sounds heard over other parts of the lungs (bronchovesicular or vesicular), but quieter and more hollow (tubular) in sound compared to tracheal breath sounds. Tracheal and bronchial breath sounds are commonly grouped together since they have very similar characteristics, both being generated by the larger airways. Therefore, bronchial breath sounds, like tracheal sounds, usually occur between 100 and 1,500 Hz. However, the intensity of these sounds undergoes a more significant drop when the frequency exceeds 800 Hz. Various lung disorders such as pneumonia, lung tumors, atelectasis (lung collapse), or pneumothorax may cause bronchial breath sounds to be heard in other lung regions, indicating the aforementioned pathologies. These sounds are referred to as abnormal bronchial breath sounds when heard outside their auscultation area.

Bronchovesicular breath sounds are of intermediate loudness, intensity, and pitch. They can be heard over the first and second intercostal spaces along the sternum and between the scapulas. The inspiratory and expiratory phases of these sounds are of equal length, following an I:E ratio of (1:1). Bronchovesicular sounds are intermediate between bronchial and vesicular sounds, mainly generated by the secondary bronchi in both lungs. Therefore, the frequency range and intensity of these sounds have intermediate values between bronchial and vesicular (pulmonary) breath sounds.

Vesicular (pulmonary) breath sounds are soft and low-pitched. They are heard in most of the periphery of the lung, where the network of increasingly narrow bronchioles is located. The inspiratory phase is longer than the expiratory phase, following an I:E ratio of (2:1), (3:1), or (4:1). Similar to bronchovesicular sounds, there is no pause between inspiration and expiration. Specifically, pulmonary breath sounds are transmitted through lung tissue and the chest wall. These sounds are quieter than tracheal and bronchial sounds. Inspiration is more sonorous than expiration as it fades rapidly over time. This is because turbulent airflow during expiration is quickly distributed to larger airways (trachea and primary bronchi). In sound analysis, the frequency range of normal pulmonary sounds appears to be narrower than that of tracheal sounds, extending from less than 100 Hz to 1000 Hz.

However, most of the energy is usually distributed in a spectral range between 200

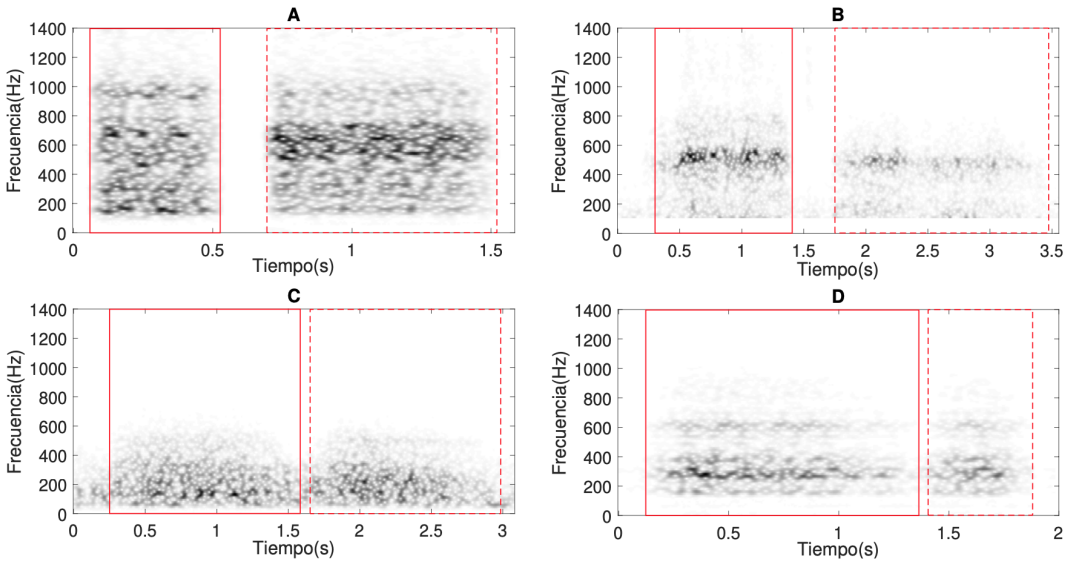


Fig. 2.14 Time-frequency representation (spectrogram) of a complete respiratory cycle (Inspiration and Expiration) for the four types of normal breath sounds: A) Tracheal breath sound; B) Bronchial breath sound; C) Bronchovesicular breath sound; and D) Vesicular or pulmonary breath sound. The solid rectangles indicate the inspiration phase, and the dashed ones represent the expiration phase. Figure origin: [83]

and 600 Hz. Nevertheless, it is usually considered that normal breath sounds contain the majority of their energy in the spectral range between 60-1,000 Hz [78]. Figure 2.14 provides an example of the spectrogram obtained for the different normal breath sounds described earlier. The clear pause between inspiration and expiration in the case of tracheal and bronchial breath sounds is due to the absence of the alveolar phase in the larger airways (trachea and primary bronchi).

Finally, it should be noted that in some cases, the normal breath sounds heard in the mouth are also included [78, 75]. However, in most studies, these sounds are not classified as normal breath sounds [116, 75, 107, 73, 24, 21, 78, 77], as they are not examined during the auscultation process used to assess the condition of the respiratory system, or simply because they are not considered auscultated sounds. Mouth-breath sounds have a frequency range distributed between 200 and 2000 Hz, similar to normal white noise [118]. Their intensity is as strong and rough as the breath sounds from the trachea and main bronchi and has a moderately high pitch [78]. In a healthy person, breathing is quiet in the mouth, but it is easily audible even at a distance in patients with chronic bronchitis and asthma. However, this sign is used less frequently today. One reason might be that stridor and wheezing are often confused with noisy breathing [118], but the simple method of listening to noisy breathing in the mouth without the need for auscultation equipment can be an important clinical sign. Generally, the mouth-breath sound is due to increased turbulence caused by irregularities in the airway surface, abrupt changes in flow direction, or narrowing of the airways that generate

faster flow [118].

2.3.3 Adventitous respiratory sounds

Adventitious or abnormal respiratory sounds breath sounds are defined as additional respiratory sounds that overlap with normal breath sounds [119, 97]. The presence of these sounds during the mechanics of breathing (inspiration and expiration) usually indicates the presence of a pulmonary disorder. Various lung pathologies and injuries cause alterations in the airways that transmit sound, leading to adventitious breath sounds that, if properly analyzed, can provide additional information about the severity and location of the disease. This section examines the different types of adventitious breath sounds and their spectral characteristics, with a particular focus on wheezes and crackles, the subject of study in this Thesis. The classification of adventitious sounds has been developed based on the definitions and spectral characteristics established by the European Respiratory Society (ERS) [120]. Adventitious sounds can be classified into two groups: continuous adventitious sounds (such as wheezes) and discontinuous adventitious sounds (such as crackles) [78, 43, 121, 122, 74, 123]. Furthermore, there are adventitious sounds that share common features between continuous and discontinuous adventitious sounds, known as squawks.

Continuous adventitious sounds (musical):

Continuous adventitious sounds (CAS), are primarily defined by their temporal continuity, with a duration exceeding 100 ms. Moreover, these sounds are often described as ”musical” because they exhibit spectral patterns similar to musical instrument notes (this behavior can be observed in the corresponding spectrogram). Depending on their spectral location, duration, and pitch of these spectral patterns, continuous adventitious sounds can be classified as wheezes, rhonchi, and stridor.

Wheezes are produced by localized obstruction of the airways, caused by a foreign body, mucous plug, or a possible tumor [78]. Wheezes are a non-specific finding and can even be detected in a healthy individual towards the end of expiration during forced exhalation. However, it has been demonstrated that pathological wheezes, those resulting from a respiratory disorder, can occur even when the mechanics of breathing are smooth [116]. Various studies have attempted to delve into the mechanism of wheeze production. Initially, in 1967, Forgacs proposed that wheezes are generated by oscillations of the bronchial walls initiated by the airflow passing through them, and the pitch of wheezes depends on the mechanical properties of the bronchial walls [124]. In his study, Forgacs defines wheezes as a musical sound and compares them to the sound produced by a toy trumpet, where the sound is generated by reed vibration. Later, Grotberg and Gavriely proposed a mathematical model based on fluid dynamics to try to explain the mechanics of wheeze production [125]. The oscillations that generate wheezing be-

gin when the airflow velocity reaches a critical value. This model shows that wheezing is always accompanied by airflow limitation, but airflow limitation is not necessarily accompanied by wheezing.

In general terms, wheezes are defined as continuous adventitious sounds generated by the obstruction of the trachea or bronchi that make up the bronchial tree. They can be located during inspiration, expiration, or in both phases of the respiratory cycle, as shown in Figure 2.15. Wheezes always overlap with normal breath sounds since both sounds are produced by the same airflow that traverses the bronchial tree. Thus, they can appear along with any of the four types of normal breath sounds (tracheal, bronchial, bronchovesicular, and vesicular or pulmonary). Furthermore, wheezes are characterized by having a high pitch, which makes them louder than normal breath sounds on certain occasions. This can lead to wheezes often being audible in the patient's mouth without the need for the auscultation process when the individual breathes, inhaling and exhaling the airflow through their mouth [119]. As mentioned earlier, there is no single time-frequency characterization for wheezes. On one hand, the American Thoracic Society (ATS) defines these sounds as having a pitch exceeding 400 Hz with a duration of over 250 ms [126]. However, according to the guidelines established by Computed Respiratory Sound Analysis (CORSA), wheezes are defined as having a pitch exceeding 100 Hz with a duration of over 100 ms [97]. Moreover, the waveform of a wheeze in the time domain resembles that of a sinusoidal sound. Therefore, wheezing sounds are characterized by narrow-band spectral trajectories (spectral peaks), as can be seen in Figure 2.15.

Wheezing is probably the most commonly used acoustic term in the field of human respiratory system medicine. This is because in hundreds of publications each year, wheezing is referenced as an acoustic biomarker of airway obstruction, as a parameter for measuring the severity of asthma, or as a classifier for epidemiological surveys, to name a few examples. Identifying these sounds during the normal respiratory cycle is of great importance in the diagnosis of obstructive airway diseases. In fact, [97] indicates that wheezing can show acoustic characteristics not only of the presence of abnormalities in the respiratory system but also of the severity and location of airway obstruction in some of the most common obstructive respiratory diseases. Although wheezing is often associated with asthma episodes, numerous studies demonstrate that these sounds are also indicators of other diseases such as COPD, bronchitis, or bronchiectasis, among the most common. The severity level of these diseases could be related to the duration, number, fundamental frequency, and temporal location of wheezing within the inspiration or expiration interval in a respiratory cycle. Specifically, wheezing is commonly defined as a clinical sign in patients with obstructive airway diseases, particularly during acute asthma episodes.

Wheezing can be classified as monophonic or polyphonic, depending on the har-

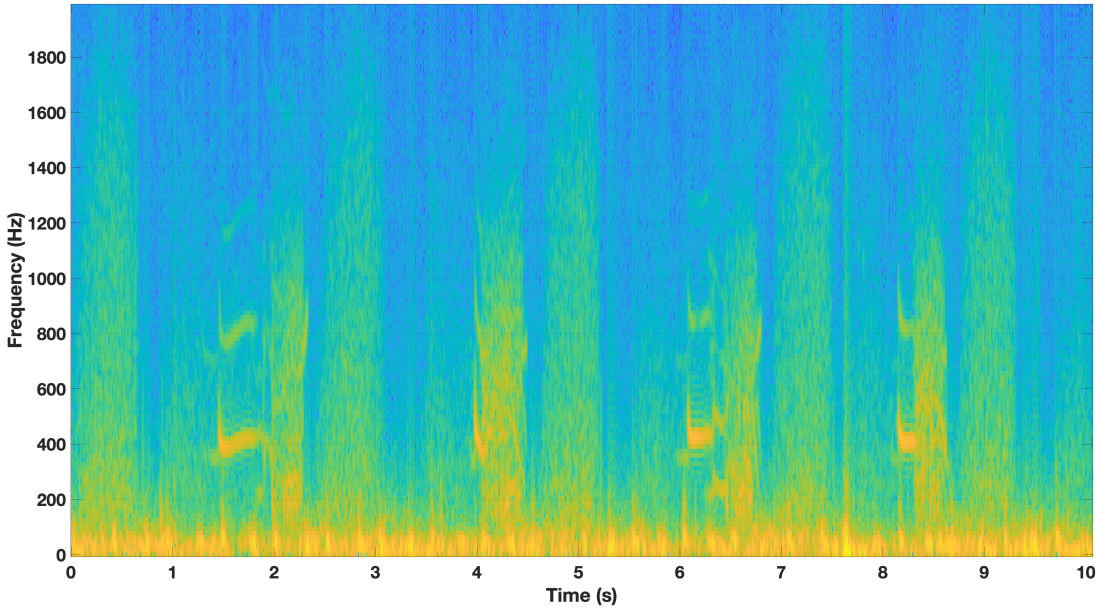


Fig. 2.15 Time-frequency representation (spectrogram) of respiratory signal with wheezing sounds present during the mechanics of breathing (inspiration and expiration). 4 wheezing sounds present here in intervals [1.5-2s], [4-4.5s], 6-6.5[s] and [8.1-8.5s]

monic structure that exists between different spectral paths or frequency components that make them up. Monophonic wheezing consists of a single narrow-band spectral peak (fundamental frequency) or the fundamental frequency along with its harmonics. Therefore, monophonic wheezing is characterized by defining a single spectral path over time or several harmonically related spectral paths. On the other hand, polyphonic wheezing is composed of a set of unrelated narrow-band spectral peaks (tones). Therefore, polyphonic wheezing is characterized by a set of non-harmonically related spectral paths over time. Monophonic wheezing originates from the obstruction of one of the larger airways (trachea, primary and secondary bronchi) and is associated with asthma. In contrast, polyphonic wheezing is caused by multiple obstructions in the smaller central airways in the lungs (bronchioles) and is commonly associated with COPD.

Stridor is a loud, high-pitched, continuous musical sound produced by the rapid flow of air through a blocked segment in the extrathoracic airways (trachea and larynx). Therefore, it is often overlapped with tracheal breath sounds. It is produced by a mechanism similar to the vibration of a reed in a musical instrument, much like wheezing. In signal analysis, it is characterized by a sinusoidal waveform with a fundamental frequency generally exceeding 500 Hz (can reach up to 1000 Hz), often accompanied by several harmonics, and a duration exceeding 250 ms. Considering the characteristics of stridor (see Figure 2.16), it is commonly defined as a type of monophonic wheezing sound. However, stridor can be distinguished from wheezing because it is more prominent in the neck than inside the chest wall, generally louder than wheezing, and typically

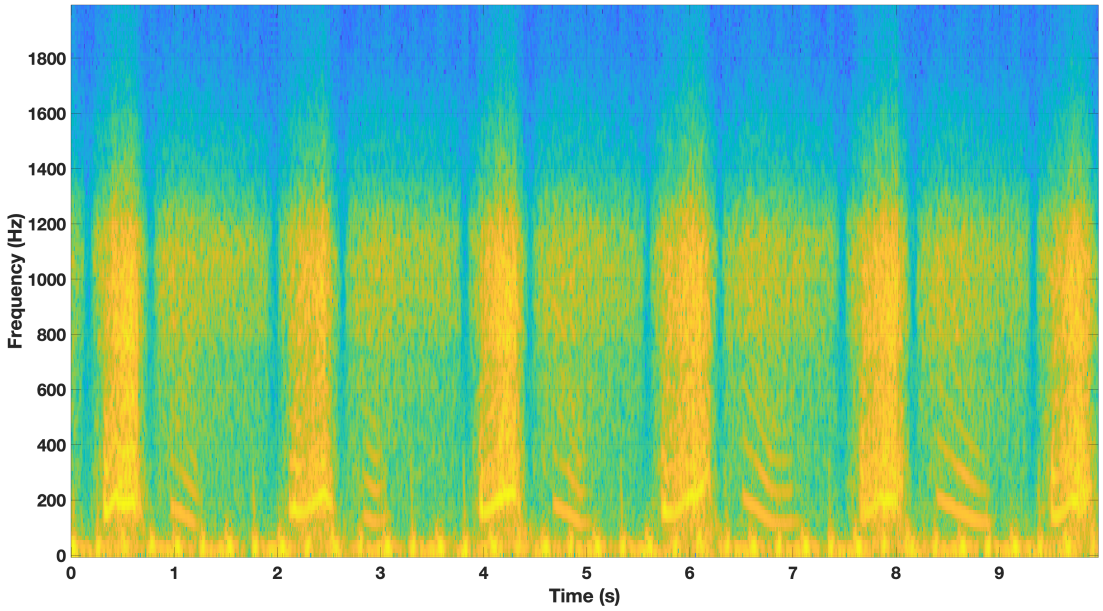


Fig. 2.16 Time-frequency representation (spectrogram) of respiratory signal with stridor sounds present during the mechanics of breathing (inspiration and expiration).

occurs primarily during inhalation. As mentioned, stridor is often associated with obstruction of the trachea or larynx, which can be caused by epiglottitis, laryngotracheitis, tracheomalacia, laryngomalacia, stenosis, anaphylaxis, tracheal carcinoma, vocal cord dysfunction, or inhalation of a foreign body into the trachea or larynx, among other possible pathologies or abnormalities. Furthermore, evaluating stridor is especially useful in intensive care unit patients who have been extubated, as its appearance can be a sign of extrathoracic airway obstruction requiring prompt intervention.

Roncus is considered a variant of wheezing, primarily distinguished by having a lower pitch. In signal analysis, it is characterized by a sinusoidal waveform with a fundamental frequency of less than 200 Hz, generally around 150 Hz (which is responsible for its resemblance to snoring), and a duration exceeding 100 ms. Roncus, being a low-pitched sound, is better heard over the chest wall, which is why it overlaps with bronchovesicular or pulmonary breath sounds (see Figure 2.17). It can occur during both inhalation and exhalation and is altered by coughing. It is a typical symptom in acute or chronic bronchitis (COPD) and is usually accompanied by excessive bronchial secretion. It normally disappears with coughing, except in cases known as fixed roncus, where coughing does not eliminate it, indicating airway obstruction by foreign bodies. Roncus and wheezing likely share the same generation mechanism; however, unlike wheezing, roncus can disappear after coughing, suggesting that secretions play a crucial role in generating these sounds. Therefore, this sound is often considered an indicator of airway wall constriction associated with mucosal thickening, edema, or bronchospasm.

Discontinuous Adventitious Sounds (non-musical):

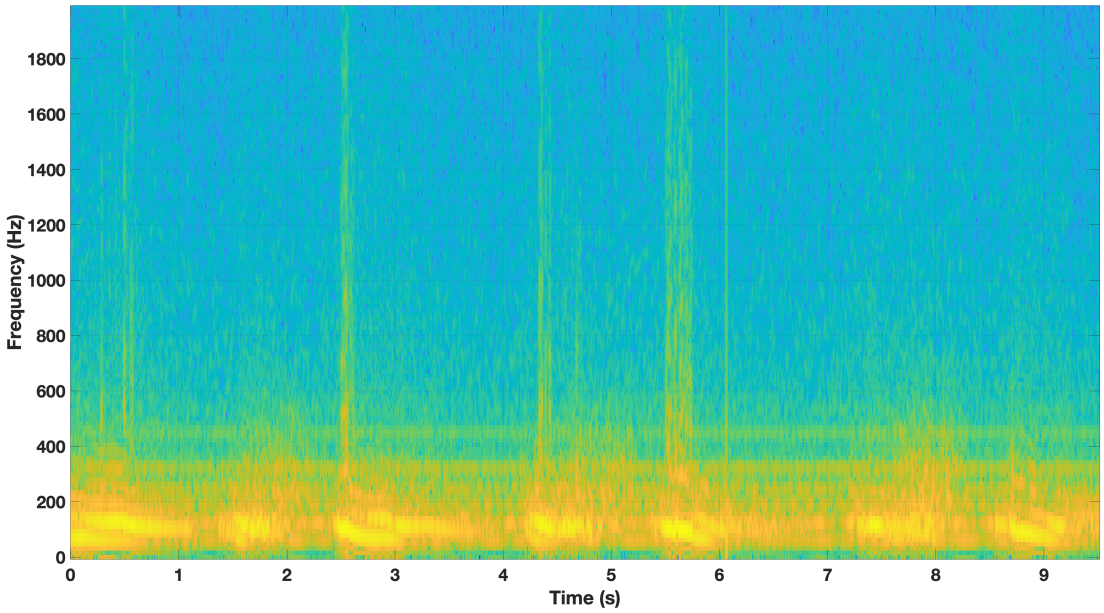


Fig. 2.17 Time-frequency representation (spectrogram) of several respiratory signals with roncus sounds present during the mechanics of breathing (inspiration and expiration).

Discontinuous Adventitious Sounds (DAS), are primarily defined by their temporal discontinuity. Unlike Continuous Adventitious Sounds (CAS), DAS lack musical properties since they cannot be characterized as notes describing a spectral trajectory over time. DAS are often defined as explosive or bubbling sounds with a duration of less than 25 ms and have a repetitive character. Instead of describing spectral trajectories over time, they can be represented by intermittent pulses or broad-spectrum patterns that appear periodically over time (this behavior can be seen in the spectrogram of these sounds). Among the main discontinuous adventitious sounds are crackles and pleural rub.

Crepitations (crackles) are discontinuous, explosive, and non-musical adventitious sounds typically heard during inspiration and sometimes during expiration. Crackles often indicate a pathological abnormality in lung tissue or the airways. In general, the frequency range of these sounds is [60-1000] Hz, with the majority of the power between [60-1,200] Hz [127]. Additionally, crackles, being discontinuous, have a duration of less than 20 ms, and they appear intermittently during the respiratory cycle in the form of broadband spectral patterns, as shown in Figure 2.18. Clinical conditions in which crackles can be present include pneumonia, pulmonary fibrosis, COPD, bronchiectasis, and heart failure, among others. In the detection of crackles, the number of crackles in a single respiratory cycle is important as it can indicate the severity of lung and airway disorders. However, more than the quantity of crackles, their position within the respiratory cycle and the variation in the waveform that generates them are the key characteristics that help determine the type of lung pathology.

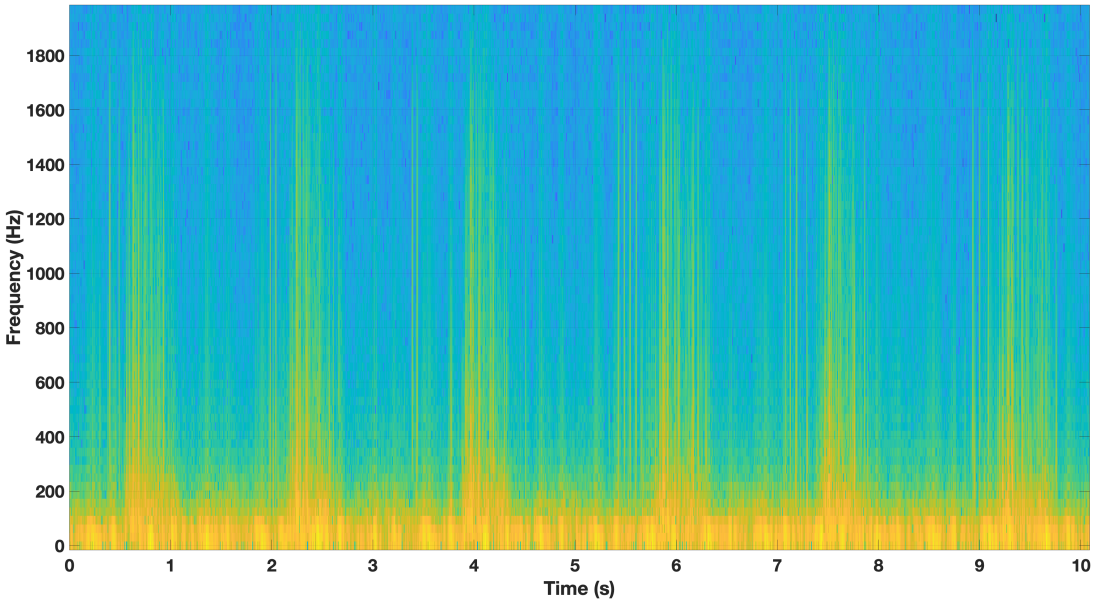


Fig. 2.18 Time-frequency representation (spectrogram) of respiratory signal with crackles sounds present during the mechanics of breathing (inspiration and expiration). 6 crackles events present here in intervals [0.5-1s], [2.3-2.7s], [4-4.4s], [5.9-6.3s], [7.5-7.9s] and [9.3-9.8s]

Crackles are often classified as fine crackles or coarse crackles based on several considerations [128]:

- The mechanism for generating fine crackles is the sudden inspiratory opening of small airways that were kept closed by surface forces during the previous expiration. On the other hand, coarse crackles are produced by air bubbles passing through long bronchi or bronchiectatic segments, intermittently opening and closing.

- During auscultation, fine crackles are typically heard during the middle or end of inspiration in dependent lung regions (where smaller airways are located) and are not transmitted to the mouth. Coarse crackles, on the other hand, can be heard at the beginning of inspiration and throughout expiration. They can be auscultated in any lung region and can be transmitted to the mouth.

- Fine crackles are not altered by coughing, although they may change or disappear with changes in body position (e.g., leaning forward). In contrast, coarse crackles may change or disappear with coughing and are not influenced by changes in body position.

- Coarse crackles are loud, low-pitched, and less numerous per breath, while fine crackles are soft, higher-pitched, and more numerous per breath. In terms of sound quality, coarse crackles are likened to salt being poured into a hot pan,

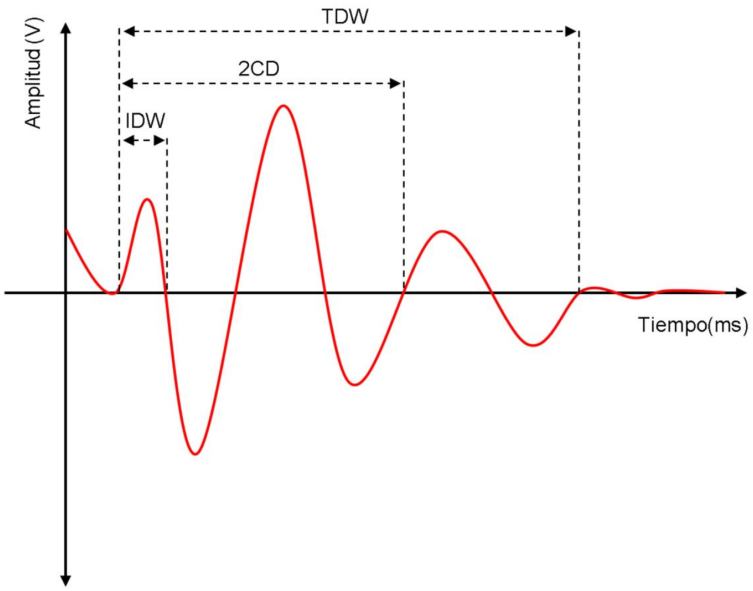


Fig. 2.19 Generic waveform of a crackling sound.

while fine crackles sound more like slowly separating strips of Velcro or a bottle of carbonated water being opened.

- In sound analysis, fine crackles have a shorter duration (around 5 ms) compared to coarse crackles (around 15 ms). Additionally, fine crackles are characterized by a higher frequency (around 650 Hz) compared to coarse crackles (around 350 Hz) [124].

- In a pathological context, fine crackles are associated with conditions like pneumonia, fibrosis, or heart failure, among others. Coarse crackles, on the other hand, are related to bronchiectasis, severe pulmonary edema, chronic bronchitis, and COPD, among other conditions.

On the other hand, Murphy [129] proposed an objective classification of the type of crackling sound by analyzing its characteristic waveform over time. As can be seen in Figure 2.19, the waveform of crackling sounds is generally represented as a long, damped sinusoidal wave [130, 131]. Specifically, the IDW parameter refers to the duration (in ms) between the start and the first intersection with the time line (either above or below); the 2CD parameter indicates the duration of the first two cycles of the crackling sound; and the TDW parameter corresponds to the total duration of the crackling sound. Considering these parameters, ATS classifies crackling sounds as coarse when the average durations of IDW and 2CD are around 1.5 and 10 ms, respectively, and as fine when they are around 0.7 and 5 ms [123]. On the other hand, CORSA states that coarse crackles occur when $2CD > 10$ ms, and fine crackles when $2CD < 10$ ms [97].

Pleural rub: It is a discontinuous adventitious sound, explosive, short in duration,

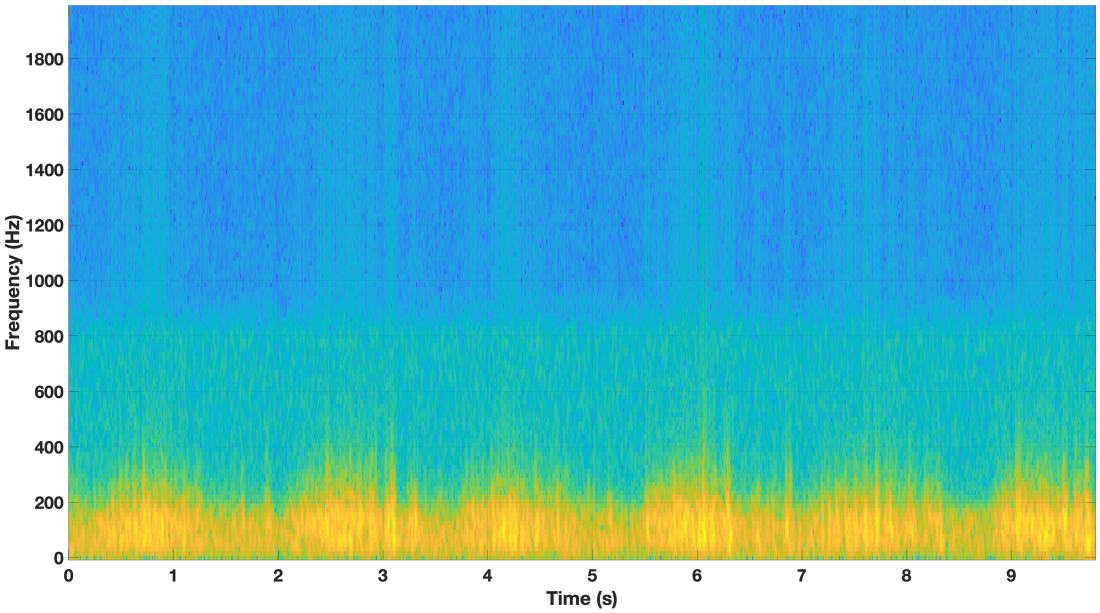


Fig. 2.20 Time-frequency representation (spectrogram) of a respiratory audio signal with the so-called adventitious sound pleural rub.

rhythmic, and squeaky in nature that can be heard during both inspiration and expiration (see Figure 2.20). Typically, the inspiratory component is shown in the expiratory component [82]. These sounds are caused by the rubbing of the pleural membranes during respiration. In healthy individuals, the parietal and visceral pleura glide over each other silently. However, in people with certain respiratory pathologies, the visceral pleura becomes inflamed and solidified. Therefore, the sliding between both pleural layers generates friction, producing the sound known as pleural rub [116]. In terms of loudness, it is often compared to the sound of walking on snow or the creaking of new leather [43]. In sound analysis, pleural rub is characterized by a duration of more than 15 ms and a frequency lower than 350 Hz [43, 26]. In pathological terms, these sounds are caused by pleurisy or a lung tumor and are often indicative of diseases like pleuritis or mesothelioma [116]. The differential diagnosis between pleural rub and coarse crackles is usually challenging because the waveforms of both sounds are very similar. However, pleural rub has a longer duration and a lower frequency. Additionally, pleural rub is usually biphasic (present in both phases of the respiratory cycle) and is not affected by coughing. In contrast, coarse crackles can occur independently in both phases and are altered by coughing [78].

Mixed adventitious sounds (continuous and discontinuous):

Within this category, we can find squawks, which contain both musical and non-musical components. Squawks are sounds that appear during late inspiration and are termed mixed because they are composed of short-duration wheezing and crackling sounds (see Figure 2.21). These sounds are generated by oscillations in the peripheral airways (the smaller ones) located in deflated lung regions when their walls remain in

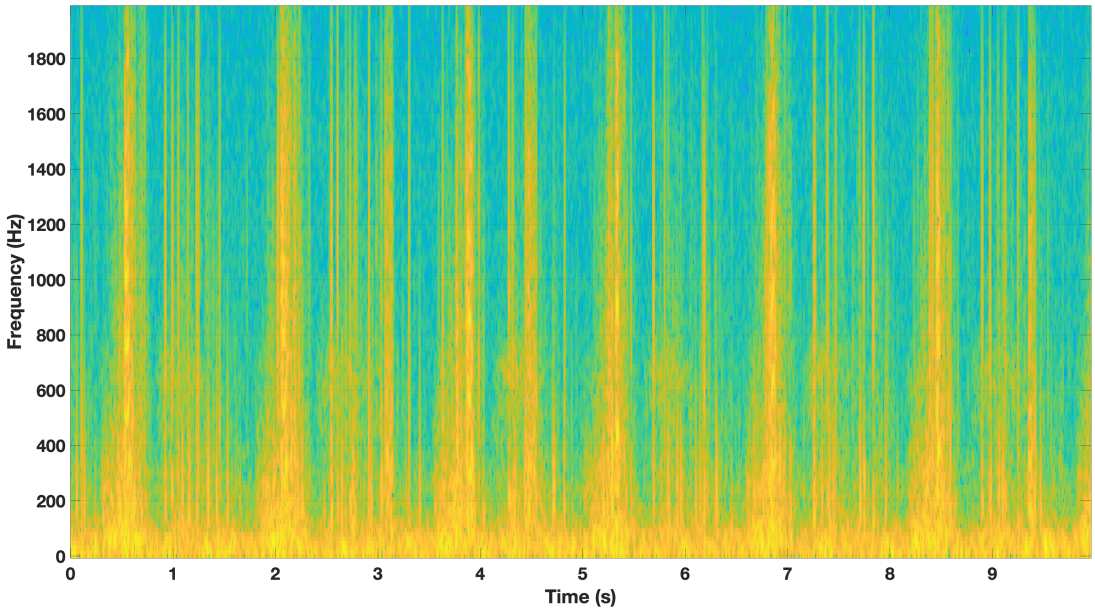


Fig. 2.21 Time-frequency representation (spectrogram) of a respiratory audio signal with the so-called squaks adventitious sound.

contact for an extended period and then open at the end of inspiration [116, 121]. In sound analysis, squawks are characterized by a short wheeze with a duration of less than 200 ms and a fundamental frequency ranging from 200 to 300 Hz, along with some crackling sounds. Additionally, the fundamental frequency of wheezing is often accompanied by a set of harmonics, as shown in Figure 2.21. In pathological terms, these sounds can be heard in patients with pulmonary fibrosis due to hypersensitivity pneumonitis, patients with interstitial lung disorders, and patients with pneumonia or obliterative bronchiolitis [132, 133, 134, 119, 97].

2.4 Conclusions

In this second chapter, basic concepts related to respiratory biomedical audio have been presented, which are necessary to understand how respiratory sounds are generated and the importance of adventitious sounds in identifying potential respiratory pathologies in subjects. First, a description of the human respiratory system has been provided, covering its anatomy, physiology, and pathology, allowing an understanding of the nature of respiratory sounds and the different components involved in their generation, as well as the most relevant obstructive lung pathologies related to wheezing and crackling sounds. Second, the fundamental principles of the auscultation procedure have been described, highlighting its advantages and limitations, as well as the various commercial options available for recording auscultated sounds. Finally, an overview of the different types of respiratory sounds that can occur during the mechanics of breathing has been provided, categorizing them into normal and adventitious respiratory sounds and distin-

guishing the various adventitious sounds that can be produced during the auscultation of a subject with a respiratory pathology.

To conclude this chapter, it is important to note that a comprehensive presentation of the different types of adventitious sounds has been carried out, with the aim of demonstrating the differences and similarities between them. However, this Thesis is focused solely on the analysis of wheezing and crackles sounds.

CHAPTER 3

Literature review

3.1 Classification and detection of respiratory sounds

In this pivotal chapter, we delve into the intricate landscape of the state of the art within the scope of our thesis. As we embark on this exploration, our primary focus will be directed towards the detection of wheezes and crackles, two critical elements that form the heartbeat of our research. This chapter serves as a comprehensive overview, meticulously crafted to provide a nuanced understanding of the existing methodologies, advancements, and challenges that define the current state of research in our chosen domain.

While traversing through this narrative, we will unravel the complexities of contemporary approaches and technologies employed in the broader field of respiratory sound analysis. Our intention is to establish a solid foundation, offering a contextual backdrop against which the significance of wheezes and crackles detection can be fully appreciated. As we navigate through the literature, we will discern key insights and discerning trends, laying the groundwork for the subsequent chapters that will dive into the specifics of our own contributions and methodologies.

On the journey through the cutting-edge advancements and evolving paradigms that shape the landscape of respiratory sound analysis, our compass is pointed specifically towards the wheezes and crackles detection. This chapter sets the stage for a deeper exploration into the heart of our thesis, where the convergence of technology and medical science converges to address the intricacies of respiratory health.

3.1.1 Preprocessing

In signal processing, preprocessing refers to a set of operations or techniques applied to raw input data (signals) before further analysis, feature extraction, or processing. The goal of preprocessing is to enhance the quality of the data, remove noise or irrelevant information, and make the data more suitable for subsequent processing steps. Preprocessing is often performed to improve the performance and accuracy of algorithms and models that work with signal data. Common preprocessing tasks in signal processing

may include:

3.1.1.1 Noise Removal

Noise removal [135, 136, 137, 138, 139, 140, 141, 142], in the context of signal processing, refers to the process of reducing or eliminating unwanted or irrelevant components from a signal, leaving behind the desired information. Noise is any undesirable interference or disturbance that gets mixed with the original signal, making it difficult to analyse or interpret. The goal of noise removal is to enhance the quality and clarity of the signal for more accurate and meaningful analysis. Here are some key aspects of noise removal:

A. Types of Noise [135, 136]: Noise can take various forms, such as random electrical fluctuations (electronic noise), interference from other signals (interference noise), or unwanted background sounds (acoustic noise). The nature of the noise dictates the methods used for its removal.

B. Techniques for Noise Removal [137, 138]:

- Filtering: Filtering methods use digital filters to attenuate specific frequency components of the signal, effectively reducing noise at those frequencies. Common filters include low-pass filters to remove high-frequency noise and high-pass filters to eliminate low-frequency noise.
- Wavelet Denoising: Wavelet transforms are used to decompose the signal into various frequency components. By thresholding and removing the high-frequency components, wavelet denoising can effectively reduce noise.
- Statistical Methods: Statistical techniques, such as averaging or median filtering, can help reduce random noise. These methods compute statistical properties of the signal and use them to distinguish between the signal and noise.
- Adaptive Noise Cancellation: This approach involves the use of a reference signal (representing the noise) to cancel out noise components from the original signal.

C. Trade-Offs [139, 140]: Noise removal is a trade-off between removing noise and preserving the desired signal. Aggressive noise removal may lead to loss of valuable signal information, while minimal noise removal may not eliminate noise adequately.

D. Application Areas [140, 141]: Noise removal is crucial in various fields, including audio and speech processing, medical imaging, communication systems, and more. In audio, for instance, it's used to remove background noise from audio recordings or enhance speech clarity.

E. Evaluating Success [141, 142]: The success of noise removal is often assessed using metrics such as signal-to-noise ratio (SNR) or mean squared error (MSE). These metrics quantify the improvement in signal quality after noise reduction.

Overall, noise removal is a fundamental step in signal processing to improve the accuracy of subsequent analysis, interpretation, and decision-making. The choice of noise removal method depends on the nature of the signal and the specific characteristics of the noise present in the data. For example, in [143], Ulukaya et al. aim to employ a denoising approach for synthetically generated crackling sounds, addressing the challenges posed by varying noise levels while preserving the critical components that significantly influence crackle parameters. The study highlights the limitations of classical wavelet-based denoising algorithms, particularly their vulnerability to abrupt noise changes, resulting in the production of fluctuations reminiscent of Gibbs phenomena. Similarly, total variation-based algorithms [144], known for mitigating some of the issues associated with classical wavelet-based methods, encounter difficulties when applied to signals characterized by both smooth and non-smooth, piece wise-smooth characteristics, such as crackles. These limitations manifest as unwanted flat regions in the denoised signals. The proposed method in [143], which combines wavelet and total variation-based denoising, demonstrates success in eliminating undesired artefacts originating from both classical wavelet and total variation denoising methods. To assess the effectiveness of the proposed approach, the research team conducts a comparative analysis with classical wavelet-based denoising methods. This evaluation involves the use of the root mean square error metric under varying levels of white Gaussian noise, ranging from 0 to 20 dB signal-to-noise ratio (SNR).

3.1.1.2 Signal Amplification or Scaling

Signal amplification [145, 146, 147, 148, 149, 150] or scaling in the context of signal processing refers to the process of increasing or decreasing the magnitude of a signal while maintaining its relative characteristics. It involves adjusting the amplitude of a signal to make it suitable for further processing, analysis, or transmission. Amplification is typically used to enhance the visibility, audibility, or detectability of the signal, or to adapt it to specific equipment or applications. Here are some key aspects of signal amplification or scaling:

A. Amplitude Adjustment [149, 150]: Signal amplification involves modifying the amplitude (magnitude) of a signal. This adjustment can be linear, where the signal is scaled proportionally, or it can be non-linear, where different scaling factors are applied to different parts of the signal.

B. Gain and Attenuation [150, 147, 148]: Amplification can be categorized into gain and attenuation. Gain increases the amplitude of the signal, making it larger, while attenuation reduces the amplitude, making it smaller.

C. Applications [150, 145, 146]:

- Audio Systems: In audio processing, amplification is commonly used to make

audio signals more audible, as in the case of amplifiers and volume controls.

- Sensor Signal Conditioning: Signals from various sensors, such as thermocouples or strain gauges, often need amplification to match the input range of data acquisition equipment.

- Data Transmission: In communication systems, signals may be amplified before transmission to overcome losses in signal strength over long distances.

- Signal-to-Noise Ratio (SNR): Amplification can affect the signal-to-noise ratio (SNR) of a signal. Increasing the amplitude of a signal also amplifies any noise present, potentially reducing the SNR. Care must be taken to balance amplification and noise control.

D. Digital Signal Processing [150, 147, 148]: In digital signal processing, signal scaling often involves multiplying digital samples by a scaling factor. This can be applied to adjust the range or amplitude of signals.

E. Dynamic Range Compression [150, 147, 148]: In some cases, signal amplification may be non-linear to adjust the dynamic range of a signal. This is commonly used in audio processing to avoid clipping (distortion) of loud sounds.

F. Scaling Factors [150, 147, 148]: Amplification is achieved using scaling factors, which are numerical values applied to the signal's samples. A scaling factor greater than 1 increases the amplitude (gain), while a factor between 0 and 1 reduces the amplitude (attenuation).

G. Quantization Effects [150]: In digital signal processing, scaling can introduce quantization effects. These effects are related to the limited precision of digital representations and should be considered when scaling digital signals.

Signal amplification is a fundamental operation in signal processing, and it plays a crucial role in various fields, including audio engineering, telecommunications, instrumentation, and data analysis. The choice of amplification method and scaling factor depends on the specific requirements of the application and the characteristics of the signals being processed. For example in [150], the authors propose a method for crackle detection. The approach involves the extraction of various feature sets through time-frequency and time-scale analyses. These feature sets are then input into support vector machines, both as individual sets and as an ensemble of networks. Furthermore, a preprocessing step is implemented to enhance the model's performance. This preprocessing involves the removal of frequency bands that do not contribute relevant information, achieved through the application of the dual tree complex wavelet transform. This transform is known for its shift invariance, limited redundancy, and improved characteristics compared to the discrete wavelet transform. The study offers a comparative

analysis of the results obtained from individual feature sets and ensembles of feature sets, considering both preprocessed and non-preprocessed data.

3.1.1.3 Signal Segmentation

Signal segmentation [151, 152, 150, 153, 154, 155, 156, 157] in the context of signal processing refers to the process of dividing a continuous signal into smaller, non-overlapping segments or frames. Each segment typically represents a portion of the original signal that is of interest for further analysis. Signal segmentation is a fundamental step in various signal processing applications and serves several important purposes:

A. Feature Extraction [151, 152, 150]: Segmentation helps in extracting relevant features from the signal. By dividing the signal into smaller segments, specific characteristics or patterns within each segment can be analysed independently. This is particularly useful in tasks like pattern recognition, where features extracted from individual segments are used for classification.

B. Temporal Analysis [150, 154, 155]: Many signals, such as speech or biomedical signals, exhibit variations over time. Segmentation allows the study of signal dynamics in different time intervals. This is important for understanding how a signal changes over time and identifying trends or transient phenomena.

C. Reduction of Complexity [150]: In some cases, processing an entire continuous signal can be computationally expensive. Segmenting the signal into smaller parts simplifies the analysis, especially when dealing with large datasets. It allows focusing computational resources on specific segments of interest.

D. Signal Isolation [150, 155, 156]: In multi-signal environments, segmenting a mixed signal can help isolate individual sources or components. For example, in audio processing, segmenting a recording with multiple speakers can isolate each speaker's speech.

E. Noise Reduction [150, 157]: Segmentation can aid in the identification and removal of noise or interference. By focusing on specific segments, it becomes easier to distinguish signal components from unwanted noise.

F. Event Detection [150, 157, 152]: Signal segmentation is crucial for detecting specific events or occurrences within a signal. For example, in video processing, segmenting a video stream into frames allows the detection of key events in each frame.

G. Time-Frequency Analysis [150]: In the analysis of non-stationary signals, segmentation is often used to apply time-frequency transforms (e.g., Short-Time Fourier Transform or Wavelet Transform) to each segment separately. This helps reveal the signal's frequency content at different time intervals.

H. Pattern Recognition [150, 157, 152]: Segmentation is commonly employed in

pattern recognition tasks, such as speech recognition, where the signal is divided into phoneme-sized segments for analysis and classification.

The choice of segmentation method and the size of the segments depend on the specific application and the characteristics of the signal being processed. Segmentation can be fixed (i.e., segments of equal size) or adaptive, where the segmentation boundaries are determined based on specific criteria or the signal's properties. Overall, signal segmentation is a critical preprocessing step in signal processing, enabling more focused and effective analysis of signals in various fields, including audio and speech processing, image and video analysis, biomedical signal processing, and many others. For example, in [158], the authors, extracted various features through time–frequency and time–scale analysis from pulmonary signals. To assess the impact of using different window and wavelet types in crackle detection, the researchers test various windows and wavelets, including Gaussian, Blackman, Hanning, Hamming, Bartlett, Triangular, and Rectangular windows for time–frequency analysis, as well as Morlet, Mexican Hat, and Paul wavelets for time–scale analysis. The extracted feature sets, both individually and as an ensemble of networks, are input into three distinct machine learning algorithms: Support Vector Machines, k-Nearest Neighbour, and Multilayer Perceptron. Furthermore, to enhance the model's performance, the researchers employ a dual-tree complex wavelet transform, which is a shift invariant transform with limited redundancy compared to the conventional discrete wavelet transform, to remove frequency bands containing no-crackle information before conducting the time–frequency/scale analysis. They extensively evaluate and compare the comparative results of individual feature sets and ensemble sets extracted using different window and wavelet types for both pre-processed and non-pre-processed data with various machine learning algorithms.

3.1.1.4 Data Normalization

Data normalization [159, 160, 161, 162, 163, 164, 165] is a fundamental data preprocessing technique used in various fields, including statistics, machine learning, and signal processing. It involves transforming data into a standard format or scale to make it suitable for analysis and modelling. The primary goal of data normalization is to remove variations in the data that can arise due to differences in measurement units, scales, or the presence of outliers. This process ensures that the data is consistent, which is essential for accurate and meaningful analysis. Common Methods of Data Normalization:

A. Min-Max Scaling [160, 161, 162]: This method scales data to a specific range, usually [0, 1] or [-1, 1]. It transforms each data point to a new value based on its

minimum and maximum values in the dataset. The formula for min-max scaling is:

$$X_{normalized} = (X - X_{min})/(X_{max} - X_{min}) \tag{3.1}$$

where X_{min} and X_{max} are the minimum and maximum value of the dataset.

B. Z-Score Standardization [164, 165]: Also known as mean normalization, this method scales data to have a mean (average) of 0 and a standard deviation of 1. It's suitable for normally distributed data. The formula for Z-score standardization is:

$$Z = (x - \mu)/\sigma \tag{3.2}$$

where μ is the mean and σ is the standard deviation of the data.

C. Robust Scaling [159, 160, 161, 162]: Robust scaling is used when the data contains outliers. It scales the data based on the median and the interquartile range (IQR) instead of the mean and standard deviation. This makes it less sensitive to extreme values.

$$X_{normalized} = (X - X_{median})/IQR \tag{3.3}$$

D. Log Transformation [164, 165]: In cases where data is skewed, applying a logarithmic transformation can help normalize it. This is commonly used with data exhibiting exponential growth.

Applications of Data Normalization:

- Machine Learning: In supervised learning, data normalization is a crucial preprocessing step before training models like support vector machines, k-nearest neighbours, and artificial neural networks.
- Statistical Analysis: It's used in descriptive statistics and inferential statistics to standardize data for meaningful comparisons.
- Signal Processing: In audio and image processing, normalization can help bring signals to a consistent amplitude or dynamic range.
- Finance: Normalizing financial data helps in comparing the performance of assets with varying scales and units.
- Biomedical Research: Normalizing biomedical data can make it easier to compare measurements from different experiments and studies.

In summary, data normalization is a process that transforms data into a standardized format to facilitate meaningful analysis, model training, and comparisons across diverse datasets. The choice of normalization method depends on the data characteristics and the requirements of the analysis or modelling task.

Young et al. [166] affirms in their study that the sensitivity of the Littman 3200 stethoscope and the fact the Power Spectral Density (PSD) is normalised means that even healthy patients can show detectable acoustic events that are usually attributed to movement between the diaphragm and skin, bumping of the chords, coughing and talking. The cyclic nature of crackles could be a key step in being able to distinguish between a patient with crackles and a healthy patient. Another possible advancement would be to use a non-normalised version of the PSD. With the current normalised approach, individual recordings cannot be compared and contrasted. With the ability to mathematically quantify the level, intensity or dominance of crackles in a recording in such a way that it can be compared to other samples, it would be possible to provide a mathematical basis for assessment. This could be used to understand both intra- and inter-patient variability. An example of this would be the ability to take recordings at each step of a recruitment manoeuvre and validate the increase in the level of recruitment.

3.1.1.5 Resampling

Resampling [167, 168, 169, 170, 171] is a signal processing technique that involves changing the sampling rate of a signal, which essentially means altering the number of samples taken per unit of time. It is commonly used in various fields, including digital signal processing, audio processing, image processing, and data analysis, for a range of purposes such as adjusting the signal's frequency, duration, or resolution. The main reasons for using a resampling preprocessing method are:

A. Rate Conversion [167, 168]: One of the primary reasons for resampling is to change the sampling rate of a signal. This may be done to convert a signal from one sampling rate to another, which can be necessary when integrating signals from different sources, devices, or systems with different sampling rates.

B. Aliasing Correction [168, 169]: Resampling can also be used to mitigate aliasing issues. Aliasing occurs when high-frequency components in a signal are erroneously represented as lower frequencies due to undersampling. By resampling at a higher rate (oversampling) and then applying a low-pass filter, aliasing can be reduced or eliminated.

C. Signal Compression or Expansion [169, 170]: Resampling can be used to compress or expand the duration of a signal, which may be useful in various applications like audio time-stretching or time-compression effects.

E. Interpolation [167, 171]: In image processing and spatial signal processing, resampling is used for interpolation, where new samples are generated to estimate signal values at positions that are not directly sampled.

Common Methods of Resampling:

- Upsampling (Interpolation): Upsampling involves increasing the sampling rate by inserting new samples between the original ones. The new samples are generated through interpolation techniques, such as linear interpolation, cubic spline interpolation, or sinc interpolation.
- Downsampling (Decimation): Downsampling reduces the sampling rate by discarding samples. To avoid aliasing when downsampling, it is common to first filter the signal with a low-pass filter to remove high-frequency components. Then, samples are retained at the desired lower rate.
- Resampling by a Rational Factor: In some cases, the sampling rate is changed by a rational factor, such as 2/3, which requires more complex interpolation and decimation techniques.
- Zero Padding: In zero-padding, zeros are inserted between the original samples to increase the sampling rate. This is a simple method but doesn't add new information to the signal.

The applications of resampling are numerous in the investigation field. In audio applications, resampling is used for changing play-back speed, pitch shifting, and sample rate conversion. This allows for creating slow-motion or fast-motion effects in audio and ensuring compatibility between different audio devices. Resampling is also used to resize images, either to make them smaller or larger, as well as for rotation and image warping. In data analysis, resampling is used for time series data, financial data, and other datasets, where the temporal resolution needs to be adjusted to match a specific analysis requirement. In digital signal processing, resampling may be required when signals with different sampling rates need to be synchronized and processed together. For example, in [172], the authors, in their paper, introduce a system designed to detect and classify abnormal lung sound events. The system's training and testing procedures were conducted using the publicly accessible ICBHI 2017 challenge dataset [173], and they employed the metrics outlined in the challenge to ensure that their framework and results can be readily compared with other research efforts. In order to do this, the authors perform a first preprocessing method based in resampling the raw data. Since the dataset is composed by lung sounds at different sampling rates, these are resampled to 6000 Hz.

3.1.1.6 Data Interpolation

Data interpolation [174, 175, 176, 177, 178] is a technique used to estimate or generate data points within a given set of discrete data points. In other words, it involves filling in the gaps between known data points with estimated values. Data interpolation is commonly used in various fields, including signal processing, image processing,

geographic information systems (GIS), computer graphics, and scientific computing, among others. Data interpolation is typically applied when you have a set of discrete data points (data sampled at specific locations or time intervals). These data points might represent measurements, observations, or samples.

Interpolation Methods: There are various interpolation methods available, and the choice of method depends on the nature of the data and the specific requirements of the application. Some common interpolation methods include:

A. Linear Interpolation [177, 178]: This method connects adjacent data points with straight lines and estimates values at intermediate points based on the linear relationship.

B. Polynomial Interpolation [178]: Higher-order polynomials, such as quadratic or cubic, are used to fit curves through the data points.

C. Spline Interpolation [175, 176]: Spline functions (piecewise polynomials) are used to create a smooth curve that passes through the data points.

D. Nearest Neighbour Interpolation [174, 175]: This method assigns the value of the nearest data point to any point requiring estimation.

D. Bilinear and Bicubic Interpolation [178]: These are commonly used in image processing to estimate pixel values based on nearby pixel values in images.

Extrapolation vs. Interpolation: Interpolation deals with estimating values within the range of known data points. Extrapolation, on the other hand, extends the estimation beyond the known data range, which can be more challenging and less accurate.

Applications of Data Interpolation:

- Image Processing: Data interpolation is used in resizing images. When you zoom in or out of an image, interpolation methods estimate pixel values at the new locations.

- Geographic Information Systems (GIS): GIS applications often require interpolation to estimate values between known geographic data points, such as elevation or temperature.

- Signal Processing: In time-series data, data points may be missing or irregularly sampled. Interpolation is used to create a regularly sampled time series for further analysis.

- Computer Graphics: Interpolation is used in 3D computer graphics for rendering curves and surfaces, as well as for animations and frame-rate conversions.

- Scientific and Engineering Data: Interpolation is employed in various scientific fields to estimate values between measured data points, such as in environmental monitoring, physics experiments, and simulations.

In summary, data interpolation is a valuable technique for estimating data values between known data points. It is widely used in fields where continuous or regular data is required from discrete, irregularly sampled, or incomplete datasets. The choice of interpolation method should align with the specific application and data characteristics. In seismic detection field, data interpolation is a widely used technique. For example, in [179], the authors present an algorithm for seismic data interpolation utilizing generative adversarial networks (GANs). This method involves self-learning, where feature vectors of the training data are extracted without the need for preprocessing to generate training labels. Unlike conventional interpolation techniques that often rely on assumptions regarding the linearity of events or data sparsity, this algorithm operates without such assumptions. Training labels are generated by randomly removing traces from various receiver indices in the original datasets to simulate the absence of traces.

3.1.1.7 Baseline Correction

Baseline correction [180, 181, 182, 183, 184, 185, 158, 150] is a signal processing technique used to remove or adjust the baseline or background signal in data, especially in analytical chemistry, spectroscopy, chromatography, and other fields. The baseline is the relatively flat and featureless region of a signal that may obscure or interfere with the analysis of the actual data, which often contains meaningful peaks or variations. Baseline correction is essential for improving the accuracy and interpretability of the data as it is the portion of the signal that represents the background noise, drift, or other unwanted variations. In many cases, it's relatively flat or slowly varying.The primary goal of baseline correction is to isolate and enhance the signal of interest while removing or minimizing the baseline's influence. This helps in the accurate detection, quantification, and analysis of features in the data. Baseline correction can be performed using various techniques, including:

A. Polynomial Fitting [183, 184]: This method fits a polynomial function to the baseline and subtracts it from the original signal.

B. Smoothing Techniques [180, 181]: Smoothing filters (e.g., moving averages) are applied to suppress baseline variations.

C. Wavelet Transform [181, 182]: Wavelet denoising can help remove baseline interference.

D. Iterative Methods [183]: Iteratively estimate and subtract the baseline.

E. Minimum/Maximum Correction [158, 150]: Find and adjust the minimum or maximum values of the signal to form the corrected baseline.

F. Spline Fitting [185, 158]: Fit spline curves to the baseline and subtract them.

Applications of Baseline Correction:

- Spectroscopy: In techniques like nuclear magnetic resonance (NMR) and mass

spectrometry, baseline correction is crucial for accurate peak identification and quantification.

- Chromatography: In high-performance liquid chromatography (HPLC) and gas chromatography (GC), baseline correction helps in identifying and quantifying the compounds in the sample.
- Electrochemical Measurements: Electrochemical experiments often involve the correction of baseline shifts due to various factors, ensuring precise analysis.
- Biomedical Signal Processing: In electroencephalography (EEG), electrocardiography (ECG), and other biomedical signals, baseline correction helps in isolating specific signal components.
- Image Processing: In image analysis, such as in microscopy, it can be used to correct uneven illumination or background noise.

In summary, baseline correction is a crucial preprocessing step in signal processing. It helps in enhancing the accuracy of data analysis by removing or reducing unwanted variations in the baseline, making it easier to identify and quantify features of interest in the data. The choice of correction method depends on the specific application and characteristics of the data. The specific preprocessing steps depend on the nature of the signal, the analysis objectives, and the characteristics of the data. Effective preprocessing can significantly impact the quality and interpretability of results in signal processing applications such as audio processing, image processing, and time series analysis. Techniques like the wavelet transform are widely employed as preprocessing technique in biomedical signal analisis [185, 158, 150]. Fraiwan et al, conduct a study where they explore the capability of deep learning, as exemplified by deep convolutional neural networks and long short-term memory units, in the recognition of various pulmonary diseases from lung sound signals. For this purpose, each signal is subjected to an initial preprocessing procedure to optimize the input for the deep learning network. The preprocessing steps encompass wavelet smoothing, displacement removal, and normalization.

Numerous preprocessing steps have been meticulously applied to the sound signals utilized within the context of this Thesis, with a particular focus on the sound data extracted from the ICBHI dataset. The inherent challenge of dealing with sound signals of varying lengths necessitated the implementation of various preprocessing techniques aimed at standardizing these diverse inputs. Here's a comprehensive breakdown of the preprocessing steps:

Diverse Signal Lengths: The sound signals come from different sources or databases and, as such, have varying duration. In the ICBHI dataset, the lengths of individual

breathing cycles within these signals can range widely, from as short as 0.2 seconds to as long as 16.2 seconds. On average, these cycles have a duration of approximately 2.7 seconds. In order to standardize this lengths, we cut all breathing cycles at 6 seconds. In those cases where the lengths of the cycle was superior, we kept only 6 seconds of it (and decided that we could afford the error generated in the detection process). On the other hand when the breathing cycle length was inferior to 6 seconds, we applied a zero padding to reach the desired length

Zero Padding: To ensure that all the respiratory cycles have a consistent length for further analysis, a technique called "zero-padding" is employed. Zero-padding involves adding extra samples (typically filled with zeros) to the end of a signal to reach a desired length. In this case, each respiratory cycle is extended by adding zeros until it reaches a total duration of 6 seconds. This padding ensures that all cycles have the same length, making it easier to process them consistently.

Downsampling: The respiratory events of primary interest in this analysis are wheezing and crackles. These events are characterized by sound components that don't extend beyond a certain frequency range, specifically, 2 kHz (kilohertz). To standardize the frequency content of the respiratory cycles, a process called "downsampling" is performed. Downsampling reduces the sampling rate of the signal, effectively reducing the amount of data in the high-frequency range. In this case, the respiratory cycles are downsampled to a common sampling rate of 4 kHz. This step is undertaken to ensure that all the signals have a consistent frequency representation for subsequent analysis.

In summary, the preprocessing steps are designed to make the sound signals consistent in terms of their duration (by zero-padding) and their frequency content (by downsampling). These standardized signals can then be used for further analysis and feature extraction in the thesis.

3.1.2 Feature extraction

Feature extraction is a process in data analysis and machine learning where relevant information or features are selected and transformed from raw data. It aims to reduce the dimensionality of the data while retaining important characteristics, making it more suitable for subsequent analysis or modeling. Feature extraction is commonly used in various domains, including image processing, natural language processing, and signal processing.

In the context of machine learning, feature extraction is crucial for improving the performance of models. It helps in reducing noise, improving computational efficiency, and enhancing the interpretability of the data. For example, in image processing, feature extraction might involve converting raw pixel data into more abstract features like edges, textures, or time frequency representations. Overall, feature extraction plays a

fundamental role in making complex data more understandable and usable for machine learning algorithms.

A time-frequency representation is a mathematical technique used to analyze and represent signals in both the time and frequency domains simultaneously [186]. It provides a way to visualize how a signal's frequency content changes over time. Time-frequency representations are particularly useful when dealing with non-stationary signals, where the signal's frequency components vary over time. Some common types of time-frequency representations include:

- Short-Time Fourier Transform (STFT) [187, 188, 189, 190]: The STFT divides a signal into short, overlapping windows and computes the Fourier transform for each window. This representation provides insight into the signal's frequency content at different time intervals. It is widely used in signal processing, audio analysis, and image processing.

- Continuous Wavelet Transform (CWT) [191, 192, 193, 194, 195]: CWT uses wavelet functions to analyze signals at multiple scales. This method can capture both high and low-frequency components in a signal and is especially useful for analyzing transient signals.

- Wigner-Ville Distribution (WVD) [189, 190, 188, 196, 197]: WVD is a distribution that provides a complete joint time-frequency representation of a signal. It offers precise information about the signal's instantaneous frequency but can suffer from cross-term interference when signals overlap in time and frequency.

- Gabor Transform [198, 199, 200, 201]: The Gabor transform is similar to the STFT but uses Gaussian windows. It combines time-domain and frequency-domain analysis and is known for its capability to provide precise time-frequency localization.

- Spectrogram [202, 191, 190, 200, 188]: The spectrogram is a popular time-frequency representation that uses the magnitude of the short-time Fourier transform to create a two-dimensional image of a signal's frequency content over time. It is commonly used in audio and speech processing.

- Mel-Frequency Cepstral Coefficients (MFCC) [203, 204, 205]: While not a traditional time-frequency representation, MFCCs are widely used in speech and audio processing. They involve a series of steps, including spectral analysis, Mel-frequency scaling, and the discrete cosine transform, to capture the essential features of audio signals.

- Constant-Q Transform (CQT) [206, 207, 208, 209]: CQT is another time-frequency representation that uses a logarithmic frequency scale to better match human auditory perception. It is commonly used in music signal analysis.

- Cochlogram [210]: A cochleogram is a time-frequency representation of a signal that mimics the human auditory system's response to sound. It divides the signal into frequency bands, with each band having a different bandwidth, similar to the cochlea in the human ear. Cochleograms are commonly used in speech and audio processing to capture important acoustic features for analysis and classification.

These time-frequency representations have specific advantages and are chosen based on the nature of the signal and the application's requirements. Researchers and engineers select the most suitable representation to gain insights into the frequency components of signals that evolve over time.

Now, a more detailed characterization of the time frequency representations used in the development of this Thesis is given.

3.1.2.1 Short-Time Frequency Transform (STFT)

The Short-Time Fourier Transform (STFT) [187, 188, 189, 190, 211, 212, 213, 214, 215, 216, 217, 218, 219] is a widely used time-frequency representation technique for signal analysis in the frequency domain. It is valuable for understanding how the frequency content of a signal changes over time. The STFT spectrogram is a matrix $\mathbf{X}$ that provides insight into the signal's frequency components at different time intervals.

A. Basic Concept:

The STFT is a technique for analyzing signals that helps us understand how their frequency content changes over time. It divides the signal into small, overlapping time segments, and for each segment, computes a Fourier Transform to examine its frequency components.

B. STFT Spectrogram:

The output of the STFT is called the STFT spectrogram, represented as $\mathbf{X}$. It is a complex-valued matrix with dimensions $\mathbf{X} \in \mathbb{C}^{K \times M}$, where:

- K represents the number of frequency bins.

- M represents the number of time frames.

C. Calculation of Spectrogram:

The STFT spectrogram is computed for an input signal $x(t)$, which is usually sampled at a specific rate f_s (sampling rate) in Hertz (Hz). The signal $x(t)$ is typically converted into a discrete-time signal $x(n)$, where n represents discrete time indices,

using the provided sampling rate. The STFT is calculated for overlapping segments of $x(n)$, where each segment is defined by an analysis window $w(n)$ with N samples.

D. Coefficients in Spectrogram:

The individual elements of the spectrogram, denoted as $X(k, m)$, represent the spectral content of the signal for a specific frequency bin and time frame. For each k-th frequency bin and m-th time frame, the coefficient $X(k, m)$ is calculated according to the following procedure:

$$X(k,m) = \sum_{n=0}^{N-1} x\left((m-1)\cdot J + n\right) w(n) e^{-j\frac{2\pi}{N}kn}, \tag{3.4}$$

where J is the time shift expressed in samples. Typically, only the magnitude spectrogram $\mathbf{X} = |\mathbf{X}| \in \mathbb{R}_{+}^{K\times L}$ is used for analyzing the spectral content, and the phase information is ignored. However, STFT spectrograms may not be ideal for analyzing respiratory sounds since they provide constant bandwidth, resulting in lower resolution at low frequencies where most of the relevant respiratory spectral content is located. Additionally, STFT may perform poorly when analyzing respiratory sounds in noisy environments [220].

In summary, the STFT is a powerful tool for analyzing the time-varying frequency content of a signal. It operates by breaking the signal into small time segments using an analysis window and then computing the Fourier Transform for each segment. The resulting spectrogram, represented by $\mathbf{X}$, provides a detailed representation of the signal's frequency components over time, which is valuable in various fields, including audio processing, speech recognition, and music analysis. The choice of parameters like the window size and hop size can affect the resolution in time and frequency and should be adjusted according to the specific requirements of the analysis.

In the literature of adventitious sounds signals analysis, many authors made use of the Short Time Fourer Translform (STFT). For example, in [221], the researchers employed the ICBHI 2017 database, which encompasses various sample frequencies, noise variations, and background sounds, to categorize lung sounds. The signals representing lung sounds underwent an initial conversion to spectrogram images through the application of a time–frequency method, specifically the short time Fourier transform (STFT). Two distinct deep learning approaches were implemented for the classification of lung sounds. In the initial approach, a pretrained deep convolutional neural networks (CNN) model served for feature extraction, followed by the utilization of a support vector machine (SVM) classifier for the actual classification process. In the second approach, the pretrained deep CNN model underwent fine-tuning (transfer learning) using spectrogram images to enhance its performance in lung sound classification. The effectiveness of these methods was evaluated through ten-fold cross-validation to determine their accuracies. In [222], the researchers present a snore detection algorithm that in-

tegrates a convolutional neural network (CNN) and a recurrent neural network (RNN). A dataset comprising audio recordings from 38 subjects undergoing a sleep study at a clinical center was employed in the study. The recordings were captured using a total of 5 strategically positioned microphones around the bed. Utilizing the CNN, features were extracted from the short time Fourier transform (STFT). Subsequently, the RNN processed the sequential output from the CNN, classifying audio events into snore and non-snore categories. Furthermore, the study explored the influence of microphone placement on the algorithm's performance, aiming to understand its impact on the accuracy of snore detection. And in [223], the authors introduce a method for event detection in single-channel lung sound recordings, specifically targeting the identification of crackles and breathing phase events (inspiration/expiration). Their proposed approach employs spectral features and bidirectional gated recurrent neural networks (BiGRNNs) for event detection. The experiments conducted utilize multi-channel lung sound recordings obtained from both lung-healthy subjects and patients diagnosed with idiopathic pulmonary fibrosis, gathered during a clinical trial. The processing of lung sound recordings involves a sampling frequency of fs = 16 kHz. All recordings undergo processing with a short time Fourier transform (STFT) using a Hamming window with a window size of 32 ms (512 samples) and 12 ms overlap (frame-shifts of 20 ms).

3.1.2.2 Mel-frequency cepstral coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCC) [203, 204, 205, 224, 225, 226, 227, 10, 228] is a prominent feature extraction technique employed in the fields of speech and audio signal processing. It plays a crucial role in capturing the essential characteristics of sound in a format that aligns more closely with human auditory perception. Let's break down the key concepts and steps involved in MFCC in detail:

A. Motivation for MFCC:

The primary goal of MFCC is to create a feature representation of audio signals that is both informative and compact, making it suitable for various applications like speech recognition and audio classification. Human auditory perception doesn't treat all frequencies equally. We are more sensitive to certain frequency ranges, and MFCC aims to mimic this perceptual aspect.

The first step in MFCC involves calculating the power spectrum of the audio signal. This step quantifies the strength of different frequency components in the signal. It is typically obtained by taking the absolute square of the Discrete Fourier Transform (DFT) of short overlapping windows of the audio signal.

The second step and key innovation of MFCC is the application of Mel-frequency scaling to the power spectrum. This step transforms the frequency axis from a linear scale to a non-linear Mel scale.

B. Mel Filter Banks:

In practical MFCC computation, the next step involves creating Mel filter banks. These filter banks are triangular-shaped filters spaced evenly on the Mel scale. Each filter bank covers a specific range of frequencies and is used to capture the energy content within that frequency range.

C. Filter Bank Responses:

For each frame of the power spectrum, the filter bank responses are computed by taking the dot product of the power spectrum with each filter bank. This operation provides a representation of the signal's energy distribution in Mel frequency bins.

D. Logarithmic Compression:

The filter bank responses are usually subjected to a logarithmic compression. This is done to mimic the logarithmic perception of loudness by the human ear. Taking the logarithm of the filter bank responses helps in emphasizing smaller variations in energy.

E. Discrete Cosine Transform (DCT):

The final step involves applying the Discrete Cosine Transform to the logarithmically compressed filter bank responses. This step converts the data into a more compact and efficient representation. The resulting coefficients are the MFCCs, and they are typically used as features for various audio processing tasks.

F. What is the Mel Scale?:

The Mel scale is a perceptually motivated frequency scale that is designed to emulate how humans perceive sound. It is especially valuable when dealing with audio signals meant for human consumption. The Mel scale is defined as follows:

$$\text{Mel}(f) = 1127 \cdot \ln(1 + f/700) \tag{3.5}$$

where f is the frequency in Hz. The Mel-scaled power spectrum is then passed through a filterbank $H_m(k)$ composed of M triangular filters that mimic the frequency selectivity of the human auditory system as,

$$H_m(k) = \sum_{i=1}^{M} |X(k)|^2 \cdot H_m^i(k), \tag{3.6}$$

where $|X(k)|$ is the magnitude of the Fourier transform at frequency k, and $H_m^i(k)$ is the i-th triangular filter in the Mel filter bank centered at Mel frequency m_i. The filters are spaced uniformly on the Mel scale (see Eq. (3.5)), and their bandwidths increase with increasing frequency. The output of each filter is then squared and summed over frequency to obtain a measure of the energy in each filter as,

$$S_m = \log \left(\sum_{k=1}^{K} |X(k)|^2 \cdot H_m(k) \right), \tag{3.7}$$

where K is the number of frequency bins in the Fourier transform. Finally, the result is transformed using the discrete cosine transform (DCT) which decorrelates the filterbank energies and produces a set of coefficients that are often used as features for machine learning architectures. The DCT transform can be expressed as:

$$Y_n = \sqrt{\frac{2}{M}} \sum_{m=0}^{M-1} S_m \cos\left(\frac{\pi n}{M}\left(m + \frac{1}{2}\right)\right), \tag{3.8}$$

where S_m is the logarithmic scaling of the m-th filter bank output and Y_n is the n-th MFCC coefficient. The first few MFCC coefficients tend to capture the spectral envelope or shape of the signal, while the higher coefficients capture finer spectral details. The number of MFCC coefficients is typically chosen based on the application, and can range from a few to several dozen.

In conclusion, MFCC is a feature extraction technique in audio signal processing that leverages the Mel-frequency scaling to better match human auditory perception. By capturing the distribution of energy across Mel frequency bins and using the Discrete Cosine Transform, MFCCs provide a powerful and compact representation of audio signals, making them suitable for a wide range of applications, including speech recognition, music analysis, and more. In the reseach field of this Thesis, we can find many authors that make use of the MFCC as feature extraction method [229, 128, 230, 231]. For example in [229] a technique for classifying breathing sonority using Convolutional Neural Networks (CNN) is introduced. The researchers employed a training process wherein each audio sample underwent visual representation, facilitating the identification of classification features. The same methodologies used for high-precision image categories were applied to classify resources. The Mel frequency cepstral coefficients method (MFCCs) was utilized for resource extraction from every audio file in the dataset, resulting in an image representation for each audio sample. The described approach demonstrated success in categorizing respiratory diseases within the six classes available in the database, namely COPD (Chronic Pulmonary Obstructive Disease), Healthy, URTI (Upper Respiratory Tract Infection), Bronchiectasis, Pneumonia, and Bronchiolitis, achieving results exceeding 93 percent accuracy. Also in [128] the authors concentrates on the analysis of two distinct sounds, specifically crackles and pleural friction rub lung sounds, employing the Mel Frequency Cepstral Coefficients (MFCC) speech analysis technique. The calculation of MFCCs was carried out for both types of lung sounds, and four fundamental statistical parameters of MFCC were determined. Among these parameters, the standard deviation of MFCC exhibited the highest linear separability. Consequently, it is proposed that the standard deviation of MFCC can serve as a promising feature for the classification of adventitious lung sounds associated with pulmonary crackles and pleural friction rubs.

3.1.2.3 Constant-Q Transform (CQT)

The Constant-Q Transform (CQT) [206, 207, 208, 209] is a frequency-domain analysis technique commonly employed to examine signals with non-uniform frequency content, particularly musical signals. This method is especially useful for accurately representing audio signals, such as musical or respiratory sounds, which consist of a wide range of frequencies that don't evenly span the entire spectrum. The CQT is distinguished from the Short-Time Fourier Transform (STFT) by its use of a logarithmic frequency scale that closely resembles the way the human auditory system perceives sound.

Here's a detailed explanation of the key points in the provided description:

A. Frequency-Domain Analysis:

The Constant-Q Transform is a method for analyzing the frequency content of a signal. It helps in understanding the distribution of different frequencies within the signal.

B. Non-Uniform Frequency Content:

Many real-world signals, especially audio signals like music or respiratory sounds, don't have a uniform distribution of frequencies. Some frequencies may be more prominent or important than others. The CQT is well-suited for handling such signals.

C. Comparison with STFT:

The Short-Time Fourier Transform (STFT) is another common tool for analyzing signals in the frequency domain. It divides the signal into short segments and computes a Fourier Transform for each segment. The CQT differs from the STFT in its approach to frequency representation.

D. Logarithmic Frequency Scale:

The central characteristic of CQT is the use of a logarithmic frequency scale. In contrast, the STFT employs a linear frequency scale. The choice of the logarithmic scale is essential because it mimics the way humans perceive sound. In the logarithmic scale, each octave (a doubling of frequency) takes up the same amount of space on the scale. This closely aligns with how the human ear distinguishes frequencies.

E. Human Auditory System Comparison:

The reason for using a logarithmic scale is to make the frequency representation more akin to how our auditory system processes sound. Human hearing is more sensitive to relative changes in pitch (frequency) rather than absolute changes. Logarithmic frequency representation captures this characteristic more accurately.

F. Benefits for Respiratory Sounds:

Respiratory sounds often exhibit non-uniform frequency content. They may contain a mix of high-pitched wheezes, low-frequency rumbles, and various other components. A logarithmic frequency scale can better capture these nuances.

G. Mathematical Definition:

The description alludes to a mathematical definition of the CQT for an input signal denoted as $x(n)$. The mathematical expression for the CQT is provided in Eq. 3.9:

$$X(k,n) = \sum_{j=n-\lfloor N_K/2 \rfloor}^{j=n+\lfloor N_K/2 \rfloor} x(j)a^*(j-n-N_k/2), \tag{3.9}$$

where k is the frequency index in the CQT domain, $\lfloor \cdot \rfloor$ denotes towards negative infinity and $a^*(n)$ are the time-frequency atoms defined by

$$a_k(n) = \frac{1}{N_k} w\left(\frac{n}{N_K}\right) \exp\left[-i2\pi n \frac{f_k}{f_s}\right], \tag{3.10}$$

where f_k is the center frequency of bin k, f_s is the sampling rate, $w(n)$ is the window function (e.g. Hann or Blackman Harris). The window lengths $N_k \in \mathbb{R}$ are inversely proportional to f_k in order to have the same Q-factor for all the frequency bins k. The Q factor of bin k is given by

$$Q_k = \frac{f_k}{\Delta f_k}, \tag{3.11}$$

where Δf_k denotes the $-3dB$ bandwidth of the frequency response of the atom $a_k(n)$ and the range of f_k obey

$$f_k = f_1 2^{\frac{k-1}{b}} \tag{3.12}$$

where f_1 is the center frequency of the lowest frequency bin and b is the number of bins per octave. In fact, parameter b determines the time-frequency resolution trade-off of the CQT.

Unlike the STFT, where the frequency resolution is constant across all frequency bins, the CQT has a higher frequency resolution for lower frequencies and a lower frequency resolution for higher frequencies. Another advantage of the CQT is that it can provide better time-frequency resolution compared to the STFT for signals with rapidly changing frequencies. Authors as [232, 233, 234] employ the CQT as feature extraction. Healthcare professionals routinely employ stethoscopes to auscultate the lungs of individuals, aiming to diagnose chronic obstructive pulmonary diseases (COPDs) and other lower respiratory infections such as asthma. Given the non-stationary and time-varying nature of most biomedical signals, time-frequency analysis serves as a comprehensive method to describe signals across various time-frequency planes. In the study by [233], the focus is on respiratory signals as the non-stationary signals of interest. Four time-frequency analysis methods—short-time Fourier transform (STFT), continuous wavelet transform (CWT), Wigner–Ville distribution (WVD), and constant Q-Gabor transform (CQT)—are considered for the analysis of lung sounds. To facilitate

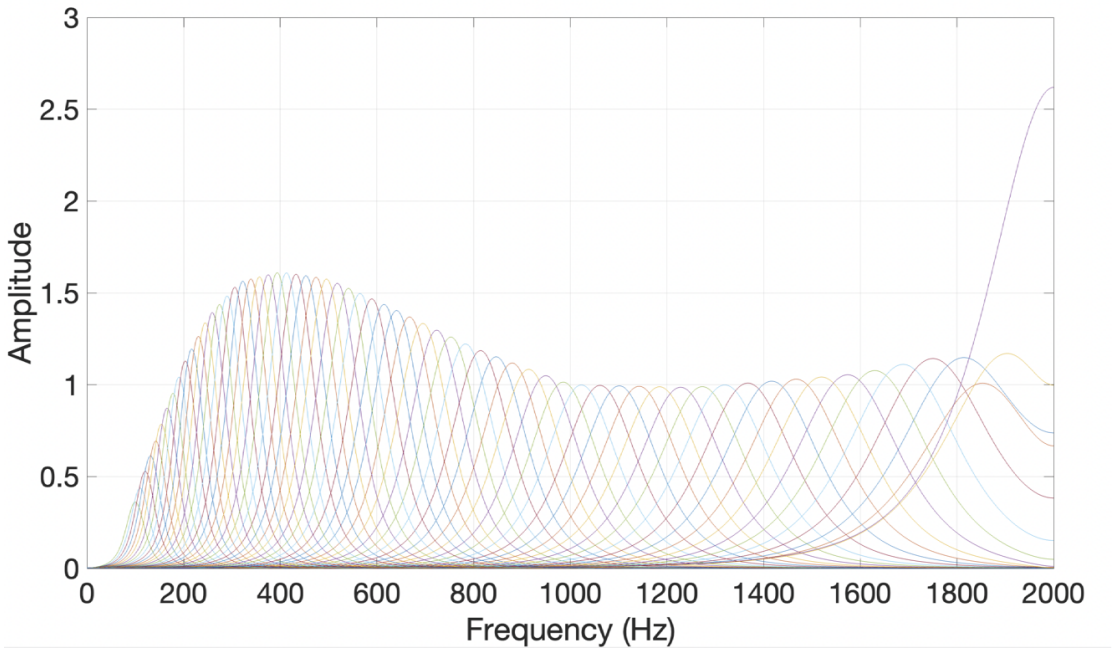


Fig. 3.1 Middle-ear gain normalization of the frequency response of 64-channel gammatone filter bank [1]. It can observed higher spectral resolution at low frequencies.

this analysis, lung sounds are decomposed into intrinsic mode functions (IMF) through empirical mode decomposition (EMD). The study further involves the classification of vesicular and adventitious sounds, such as crackles, wheezes, and rhonchi, utilizing a pre-trained GoogLeNet classifier.

3.1.2.4 Cochleogram

It is known that the gammatone filter (see Figure 3.1) is designed to replicate the frequency selectivity of the human cochlea by using non-uniform spectral resolution. This approach associates wider frequency bandwidths with higher frequencies to mimic the human ear's performance. This variable resolution provides a TF representation that can extract more accurate spectral content from the input signal due to higher robustness against noise and acoustic changes [235, 1, 220]. To compute the cochleogram, a gammatone filter bank is used. The impulse response of the gammatone filter $g(t)$ is obtained by multiplying a gamma distribution and a sinusoidal function as follows,

$$g(t) = t^{o-1} e^{2\pi b(f_c)t} cos(2\pi f_c t), t > 0 \tag{3.13}$$

where the filter order o and the exponential decay coefficient $b(f_c)$ associated with the center frequency of the filter f_c Hz determine the bandwidth. The center frequencies are equally spaced on the equivalent rectangular bandwidth (ERB) scale,

$$b(f_c) = 1,019 \cdot \text{ERB}(f_c) \tag{3.14}$$

$$\text{ERB}(f_c) = 24,7 \cdot \left(4,37 \cdot \frac{f_c}{1000} + 1\right) \tag{3.15}$$

After filtering the signal using the gammatone filter, the implementation of which can be found in [1], a representation similar to the spectrogram is obtained by adding the energy in the windowed signal for each frequency channel as follows,

$$C(k,m) = \sum_{n=0}^{N-1} \left|\hat{X}(k,n)\right| w(n), \tag{3.16}$$

where $\hat{X}(k,n)$ is the gammatone filtered signal, $k = 1, ..., K$ is the number of gammatone filters and $C(k,m)$ represents the coefficient corresponding to the center frequency $f_c(k)$ for the m-th frame and $w(n)$ refers to the windowed signal. In this work, we used $K = 64$ gammatone filters with the central frequencies $f_c(k)$ distributed between 100 Hz and $\frac{f_s}{2}$ Hz, respectively, on the linear frequency scale since most adventitious respiratory sounds, mainly wheezing and crackles, contain the predominant content in this spectral range. The order is set to $o = 4$ because it provides satisfactory results in replicating the human auditory filter, as demonstrated in [220].

As an example, Figure 3.2 shows a comparison of TF representations computed by means of STFT, Mel-scaled spectrogram, CQT, and cochleogram. Among these representations, the cochleogram stands out for its ability to provide a highly accurate depiction of adventitious sounds. In fact, this gammatone filtering technique with non-uniform resolution proves to be particularly effective in modeling the low spectral respiratory content.

In the realm of biomedical signal processing, there is a notable absence of literature utilizing cochleograms as a feature extraction method for the detection of pulmonary abnormalities. Consequently, the primary objective of this thesis is to explore the potential of incorporating new time-frequency representations as input for a neural network in the context of pulmonary abnormality detection. Through our investigation, we discovered that the cochleogram emerged as the most effective approach when compared to other time-frequency representations. This finding underscores the significance of employing cochleograms in the field, revealing their superiority in capturing relevant information for the detection of pulmonary abnormalities. However, in other biomedicl fields, the literature include several authors that proceeded to use this novel time frequency representation [236, 237]. For example, in [237], the study involved recording the five waves constituting the averaged neural response (cochlear audiometry or electrocochleogram) to click acoustic stimuli using pinna and scalp electrodes in both human and animal subjects. Similarities in latencies and waveforms of the waves in cats and humans were observed. Simultaneous recordings in cats with pinna-scalp electrodes and intracranial recordings from auditory nuclei in the same animal led to the following conclusions

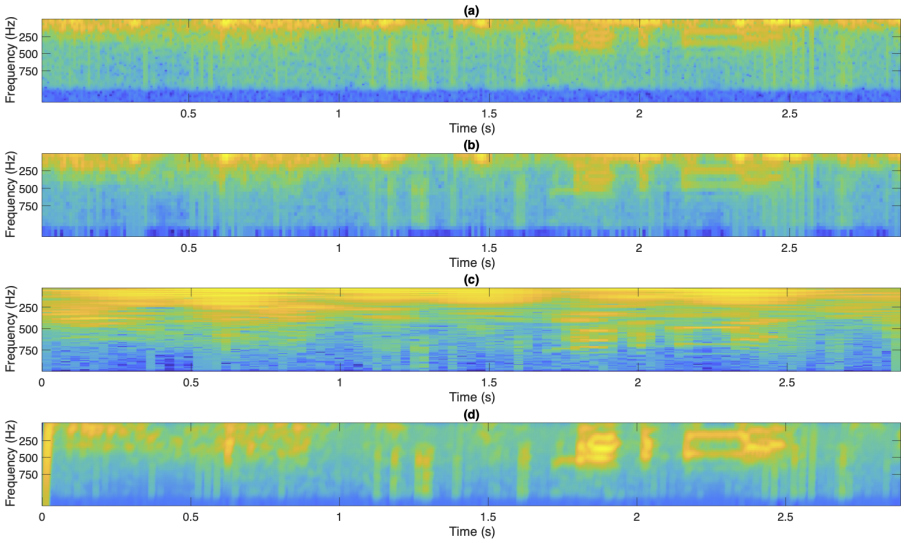


Fig. 3.2 Magnitude, in logarithmic scale, of the TF representations analyzing a respiratory cycle with a time duration of 2.9 seconds associated to the patient number 103 from ICBHI [2]. The respiratory cycle is composed by one wheeze sound located in the temporal range [2.1-2.6] seconds. STFT spectrogram (a), Mel-scaled spectrogram (b), Constant-Q (c) and Cochleogram (d).

about the sources of these waves in cats, likely applicable to humans as well: 1. the first response wave is generated by the first-order cochlear nerve fibers; 2. the second wave is mainly generated in the cochlear nucleus; 3. the third wave is generated in the superior olivary complex; 4. the fourth and fifth waves are generated in the inferior colliculus. Consequently, this recording system offers a convenient method for measuring the neural activity of the cochlea, the brain stem auditory nuclei, and the cerebral cortex in human subjects.

3.1.3 Classifier

A classifier is a machine learning or statistical model that is used to categorize or label data into different classes or categories based on patterns or features present in the data. Its primary purpose is to assign each data point to a specific class or category, typically in a binary (two classes) or multi-class (more than two classes) fashion.

The history of classifiers is intertwined with the evolution of machine learning and artificial intelligence [238, 239, 240]. Classifiers are algorithms or models designed to categorize or classify input data into different classes or categories. Here's a brief overview of the historical milestones in the development of classifiers:

Statistical Classifiers (1950s): Early classifiers were based on statistical methods. The discriminant analysis, which involves finding the combination of features that best separates classes, was one of the initial approaches.

Perceptron (1957): The perceptron, proposed by Frank Rosenblatt, was an early neural network model designed for binary classification tasks. It marked the beginning of exploring artificial neural networks for pattern recognition.

Decision Trees (1960s): The concept of decision trees emerged, where a tree-like model is constructed to make decisions based on input features. This approach provides a visual representation of decision-making processes.

Nearest Neighbor Algorithms (1967): The k-nearest neighbors (k-NN) algorithm was introduced, a simple yet effective approach that classifies data points based on the majority class of their k-nearest neighbors.

Support Vector Machines (SVM) (1990s): SVMs gained popularity as effective classifiers, especially in binary classification tasks. They aim to find a hyperplane that best separates data points of different classes.

Ensemble Methods (2001): Ensemble methods, such as Random Forests and AdaBoost, gained prominence. These methods combine the predictions of multiple classifiers to improve overall accuracy and robustness.

Naive Bayes Classifier (2000s): The Naive Bayes classifier, based on Bayes' theorem with the assumption of independence between features, became widely used in text classification and spam filtering.

Deep Learning and Neural Networks Resurgence (2010s): With the advent of deep learning, neural networks, especially deep convolutional neural networks (CNNs), demonstrated exceptional performance in image classification tasks. The ImageNet Large Scale Visual Recognition Challenge played a significant role in showcasing the capabilities of deep neural networks.

Transfer Learning (2010s): Transfer learning became a popular technique, allowing pre-trained models to be fine-tuned for specific tasks. This approach significantly reduced the need for large labeled datasets.

XGBoost (2014): XGBoost, an optimized gradient boosting library, gained widespread use in machine learning competitions due to its efficiency and high performance.

AutoML (2010s): Automated Machine Learning (AutoML) tools and platforms emerged, simplifying the process of model selection, hyperparameter tuning, and feature engineering.

Explainable AI (2020s): Recent advancements focus on developing interpretable and explainable classifiers to enhance transparency and accountability in AI systems.

The history of classifiers reflects a continuous evolution driven by advancements in algorithms, computing power, and the increasing availability of diverse and large datasets. As AI continues to progress, classifiers will likely undergo further refinement and innovation.

Classifiers are widely used in various applications, including:

- Image Classification [241, 242, 5, 243, 244]: Classifying images into categories,

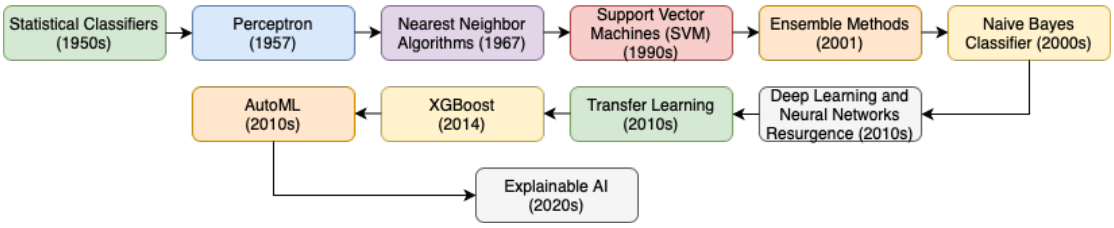


Fig. 3.3 History of Classifiers

such as recognizing whether an image contains a cat or a dog.

- Text Classification [245, 246, 247, 248, 249]: Assigning labels or categories to text data, such as spam detection in emails or sentiment analysis of customer reviews.
- Object Detection [250, 251, 252, 253, 254, 255]: Identifying and localizing objects within images or video frames.
- Natural Language Processing (NLP) [256, 257, 258, 259, 260]: Categorizing text documents, including tasks like topic classification and intent recognition.
- Medical Diagnosis [261, 262, 263]: Classifying medical images or patient data to detect diseases or conditions.
- Anomaly Detection [264, 265, 266]: Identifying anomalies or outliers in data, which can be crucial for fraud detection or network security.
- Recommendation Systems [267, 268]: Recommending products, services, or content to users based on their preferences and behavior.
- Quality Control [269, 270]: Identifying defects in manufacturing processes or products.

There are various types of classifiers, and they employ different algorithms and techniques, such as:

- Logistic Regression [271, 272]: A simple linear classifier suitable for binary classification.
- Decision Trees [273, 274]: Trees of decisions that can be used for both classification and regression tasks.
- Random Forest [275, 276]: An ensemble of decision trees, offering higher accuracy and robustness.
- Support Vector Machines (SVM) [277, 278, 279]: Effective for both binary and multi-class classification by finding a hyperplane that best separates classes.

- Naive Bayes [280, 249]: A probabilistic classifier based on Bayes' theorem, often used in text classification tasks.
- Neural Networks [281, 282]: Deep learning models with multiple layers, particularly effective for complex tasks like image and speech recognition.
- K-Nearest Neighbors (K-NN) [283, 284]: Classifies data points based on the majority class among their k-nearest neighbors.

The choice of classifier depends on the nature of the data, the specific task at hand, and the required level of accuracy. Classifier performance is typically evaluated using metrics like accuracy, precision, recall, F1-score, and the receiver operating characteristic (ROC) curve, among others.

3.1.3.1 Support Vector Machine (SVM)

Support Vector Machines (SVMs) [277, 278, 279] are a class of supervised machine learning algorithms used for classification and regression tasks. They are especially well-known for their effectiveness in binary classification problems and have been widely used in various fields, including image recognition, text classification, and bioinformatics. Here's a detailed explanation of Support Vector Machines:

A. Basic Concept:

SVMs are used to find a hyperplane that best separates data points into different classes. In the case of binary classification, this hyperplane aims to maximize the margin between the two classes. The margin is defined as the distance between the hyperplane and the nearest data points from each class. These nearest data points are called "support vectors."

B. Linear vs. Non-linear Classification:

SVMs can perform both linear and non-linear classification. In the linear case, the hyperplane is a straight line that separates the data into two classes. In non-linear cases, SVMs use a kernel trick to transform the data into a higher-dimensional space, making it possible to separate classes that are not linearly separable.

C. Hyperplane and Decision Boundary:

In a binary classification problem, a hyperplane is represented by the equation:

$$w^T x + b = 0 \tag{3.17}$$

Here, "w" is the weight vector, "x" is the input feature vector, and "b" is a bias term. The decision boundary is defined as:

$$w^T x + b > 0 \tag{3.18}$$

for one class

$$w^T x + b < 0 \tag{3.19}$$

for the other class

D. Margin Maximization:

- SVMs aim to find the hyperplane that maximizes the margin, which is the distance between the hyperplane and the nearest data points (the support vectors) from each class.
- The goal is to maximize the margin while minimizing the classification error.

E. Soft Margin SVM:

- In real-world scenarios, it's common for data to be noisy or for classes to overlap to some extent. In such cases, it may not be possible to find a perfect hyperplane.
- Soft Margin SVM allows for a certain degree of misclassification, introducing a regularization parameter "C" to control the trade-off between maximizing the margin and allowing misclassification.

F. Kernel Trick:

For non-linear classification, SVMs use the kernel trick to implicitly map data to a higher-dimensional space where classes can be separated by a hyperplane. Common kernels include the polynomial kernel and radial basis function (RBF) kernel.

G. Training:

- The training process involves finding the optimal hyperplane that best separates the data. This is typically done using optimization techniques like Quadratic Programming (QP).
- The support vectors are the data points that are closest to the hyperplane, and they play a crucial role in determining the position of the hyperplane.

H. Prediction: To make predictions, SVMs classify new data points by evaluating which side of the hyperplane they fall on.

I. Multi-class Classification:

While SVMs are originally designed for binary classification, they can be extended to multi-class classification using techniques like one-vs-all (OvA) or one-vs-one (OvO) classification.

J. Evaluation:

The performance of an SVM classifier is often assessed using metrics like accuracy, precision, recall, F1-score, and the receiver operating characteristic (ROC) curve.

Support Vector Machines are a powerful and versatile tool for solving classification problems. Their ability to handle non-linear data through the kernel trick and their robustness in high-dimensional spaces make them a popular choice in various applications where accurate classification is crucial.

3.1.3.2 Neural Networks

The advent of neural networks has made once unimaginable tasks remarkably convenient. Activities like image recognition, speech recognition, and uncovering intricate relationships within datasets have become significantly more accessible. Gratitude is extended to the distinguished researchers in this field whose discoveries and findings have enabled us to harness the true power of neural networks. Presently, artificial intelligence is applied across diverse domains, including virtual assistants, medical research, self-driving cars, and online retail stores. The progress in artificial intelligence and machine learning originated from a foundational mathematical model that paved the way for the future development of artificial neural networks. This mathematical model was conceived with the primary objective of constructing a machine capable of emulating human thought processes. The concept of instructing AI to mimic the navigational patterns of our brains dates back to the inception of computers. The longstanding aspiration of machines becoming ideal companions is now more tangible with the advent of artificial neural networks.

Neural networks serve as the foundational elements in today's technological breakthroughs within the realm of Deep Learning. Conceptually, a neural network operates as a massively parallel processing unit, capable of acquiring and storing knowledge, subsequently applying this knowledge to make predictions. Mirroring the brain's learning process, a neural network acquires knowledge from its environment. Synaptic weights, representing connection strengths, are then adjusted to store this acquired knowledge. Throughout the learning process, these synaptic weights are systematically modified to achieve the desired objectives. The idea of the brain as a distributed system was introduced by neuro-psychologist Karl Lashley in his 1950 thesis. Neural networks are often compared to the human brain due to their operation as non-linear parallel information-processing systems, swiftly conducting computations such as pattern recognition and perception. This characteristic renders them highly effective in tasks like speech, audio, and image recognition where inputs and signals inherently exhibit nonlinearity. The pioneers of neural networks, McCulloch and Pitts, published a research article in 1943 (see Figure 3.4) outlining a model with two inputs and a single output. This model activated a neuron if one input was active, the weights for each input were equal, and the output was binary, determined by a computed threshold level. Hebb's 1949 book, 'The Organization of Behaviour,' introduced the notion that the brain's connectivity contin-

uously changes in response to task alterations, becoming foundational for the development of computational models of learning and adaptive systems. Fifteen years later, Rosenblatt's perceptron emerged in 1958 as the next neuron model, separating data linearly into two classes. Despite random interconnections and a trial-and-error approach for weight adjustment, research stagnated for the next 15 years. This pause resulted from Minsky and Papert's 1969 mathematical analysis of the perceptron, revealing its limitations, such as the inability to represent crucial problems like the exclusive-or function (XOR), and the computational challenges associated with large neural networks. A new era began in 1986 with the development of the back-propagation algorithm by Rumelhart, Hinton, and Williams. This algorithm solved problems like XOR, marking the onset of the second generation of neural networks. The same year saw the publication of the influential two-volume book, "Parallel Distributed Processing: Explorations in the Microstructures of Cognition," edited by Rumelhart and McClelland, emphasizing the use of back-propagation as the most popular learning algorithm for training multilayer perceptrons.

The progression of artificial neural networks is experiencing exponential growth, and their future appears even more promising with the integration of augmented reality, machine learning, artificial intelligence, and big data. The synergy of artificial neural networks with other technologies has enhanced their utility across various applications. Biomedical signal processing represent one such application where artificial neural networks play a pivotal role. The healthcare sector stands to experience significant advantages from advancements in the future of artificial neural networks. Studies indicate that the combination of artificial neural networks and artificial intelligence can be employed in the diagnosis of critical illnesses like cancer, offering recommendations for effective treatments. Additionally, the potential exists for neural networks and artificial intelligence to contribute to the discovery of new drugs for treating life-threatening diseases.

The journey of image classification networks has been an extraordinary evolution, characterized by substantial advancements in the realms of artificial intelligence and computer vision. Below is a concise overview of the pivotal milestones in the development of image classification networks:

- Neural Networks (1950s — 1980s): The concept of artificial neural networks (ANNs) originated in the 1950s, but practical application faced computational constraints. Frank Rosenblatt proposed the perceptron, a fundamental neural network architecture, in the late 1950s. However, the limitations of single-layer perceptrons for intricate tasks led to waning interest in neural networks by the late 1960s.

- Back-propagation (1980s — 1990s): The rediscovery of the back-propagation

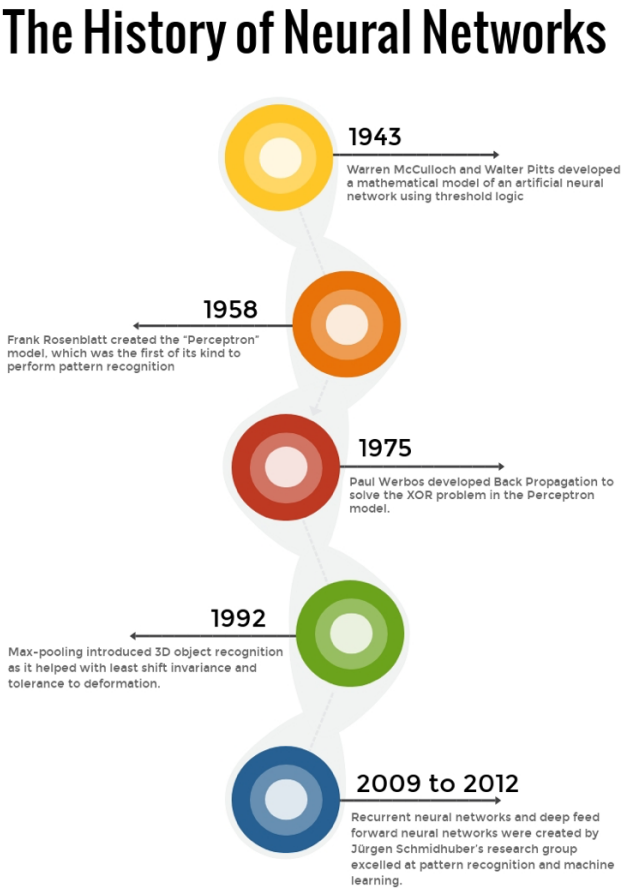


Fig. 3.4 The historical evolution of neural networks. Figure origin: `https://www.allerin.com/blog/the-evolution-of-neural-networks`

algorithm in the 1980s enabled efficient training of multi-layer neural networks, renewing interest in the field. Challenges like vanishing gradients and overfitting, however, constrained their success for image classification.

- LeNet-5 (1998): Yann LeCun's LeNet-5, introduced in 1998, pioneered the convolutional neural network (CNN) architecture for handwritten digit recognition. It featured convolutional and pooling layers, pivotal components of modern image classification networks.

- Deep Learning Resurgence (2010s): The mid-2010s marked a breakthrough for image classification with the rise of deep learning, propelled by larger datasets and more powerful GPUs.

- AlexNet (2012): AlexNet, developed by Alex Krizhevsky et al., won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. It showcased the potential of deep learning for image classification with multiple convolutional layers.

- VGGNet (2014): The Visual Geometry Group's VGGNet emphasized deeper architectures, setting a benchmark for network depth exploration.

- GoogLeNet (2014): GoogLeNet, or Inception v1, introduced inception modules, optimizing computation and leading to networks with increased depth and width.

- ResNet (2015): Residual Networks addressed the vanishing gradient problem with residual connections, allowing successful training of very deep networks and demonstrating improved performance.

- DenseNet (2017): DenseNet introduced dense connectivity patterns, promoting feature reuse and enhancing training efficiency and accuracy in image classification.

- Transfer Learning and Pretrained Models (2010s): Transfer learning gained popularity, employing networks pretrained on large datasets like ImageNet, reducing the need for extensive datasets and expediting model development.

- Efficient Networks (2019): In response to computational challenges, EfficientNet (2019) proposed a scalable architecture achieving state-of-the-art performance with fewer parameters, enhancing feasibility for various applications.

- Transformers in Vision (2020s): Initially designed for natural language processing, Transformers were adapted for computer vision tasks like image classification. Vision Transformers (ViTs) and hybrid models emerged as new contenders.

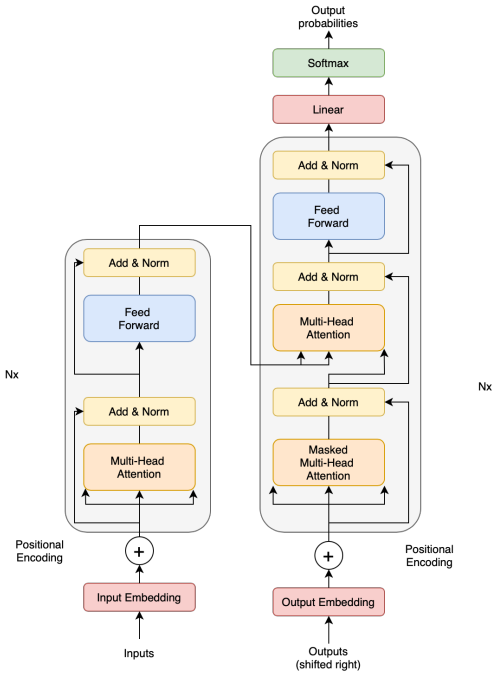


Fig. 3.5 The architecture based on the Vision Transformer model [3]

- Continued Research and Innovation (2020s): Ongoing research explores model efficiency, interpretability, robustness, and generalization in image classification. Recent developments include architectures like Swin Transformer and models utilizing self-supervised learning.

The evolution of image classification networks illustrates a continuous process of innovation, progressing from basic neural networks to sophisticated deep learning architectures. This iterative journey consistently pushes the boundaries of what AI can achieve in visual understanding and recognition. In this thesis, we made use of several of these neural networks, as the Vision Transformer (ViT), AlexNet, ResNet50 and VGG16. These architectures are explained in more detail here.

A. Vision Transformer-based classifier The Vision Transformer (ViT) [3, 285], Figure 3.5, is a revolutionary deep learning architecture designed to tackle computer vision tasks, such as image classification, object detection, and segmentation. It was introduced in the paper ”An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale” by Dosovitskiy et al. in 2020. ViT represents a significant departure from traditional Convolutional Neural Networks (CNNs), which have dominated the field of computer vision for many years. Here, I’ll explain ViT in detail:

Background:

- ViT is inspired by the success of Transformer models in natural language processing, especially in tasks like machine translation. The original Transformer model, introduced in the ”Attention Is All You Need” paper, utilized self-attention mechanisms for sequential data. ViT extends this concept to non-sequential data, such

as images.

Image Patching:

- The primary idea of ViT is to divide an input image into small, fixed-size non-overlapping patches. Each patch is treated as a token, much like a word in natural language processing. These patches are linearly embedded to create token embeddings, which are the input for the ViT model.

Positional Encodings:

- Since images do not inherently have an order like words in a sentence, positional information needs to be incorporated. Positional encodings, similar to those used in the original Transformer, are added to the token embeddings to give the model information about the spatial arrangement of the patches.

Transformer Encoder:

- The core of ViT is a stack of Transformer encoder layers. Each encoder layer consists of two sub-layers: a Multi-Head Self-Attention (MHA) layer and a Position-wise Feed-Forward (FFN) layer.
- The MHA layer allows the model to focus on different patches when making predictions for a given patch, capturing global dependencies across the image.
- The FFN layer is a simple feed-forward neural network applied independently to each patch.
- Residual connections and layer normalization are employed in each sub-layer to facilitate training.

Classification Head:

- In the original ViT paper, the final token embedding (corresponding to the entire image) is used as the representation for classification. This embedding is then passed through a linear classification head to make predictions.
- It's worth noting that ViT is pre-trained on a large corpus of images using techniques similar to BERT for language models.

Pre-training and Fine-tuning:

- ViT is pre-trained on a large dataset for self-supervised learning tasks (e.g., predicting the order of shuffled patches). After pre-training, it can be fine-tuned on specific downstream tasks, such as object detection, image segmentation, or image classification.

Advantages:

- ViT has several advantages, including the ability to capture long-range dependencies in images, making it robust to object scaling and positioning. It can also generalize to various tasks without significant architectural changes.
- Additionally, ViT has fewer architectural components compared to traditional CNNs, which can make it easier to design, train, and fine-tune.

Challenges:

- One of the challenges with ViT is its computational cost, especially for large images. Also, for some tasks, it may require a substantial amount of labeled data to perform well in fine-tuning.

In summary, the Vision Transformer (ViT) is a groundbreaking deep learning architecture for computer vision tasks. It introduces the Transformer architecture to images, treating them as sequences of patches and applying self-attention mechanisms to capture complex relationships across the entire image. ViT has shown impressive results on various computer vision tasks and has opened up new possibilities in the field of image analysis and understanding.

Our model design closely follows the original Transformer architecture, as presented in the Vision Transformer (ViT) work by Dosovitskiy et al. (2020) [3]. In this design, images are divided into patches of a predefined size, and each patch is linearly embedded. We then incorporate position embeddings into these patch embeddings. The resulting sequence of vectors is processed through a standard Transformer encoder, adapted for image-based tasks. To facilitate classification, we include a learnable ”classification token” within the sequence, which shares similarities with the original Transformer architecture designed for natural language processing tasks [260].

To manage the computational cost, ViT computes relationships among pixels within fixed-sized patches of the image. These patches are linearly embedded, and position embeddings are added to them. The sequence of vectors is then passed through a conventional Transformer encoder, consisting of a stack of six identical layers. Each layer comprises two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. To ensure information flow and aid in gradient propagation, we introduce residual connections around each sub-layer, followed by layer normalization. Consequently, the output of each sub-layer is computed as LayerNorm(x + Sublayer(x)), with Sublayer(x) representing the function implemented by the sub-layer itself. Additionally, all sub-layers in the model, including the embedding layers, produce outputs with a dimension of dmodel = 512.

The decoder is constructed with a stack of six identical layers, mirroring the encoder’s structure. Each decoder layer contains two sub-layers similar to those in the

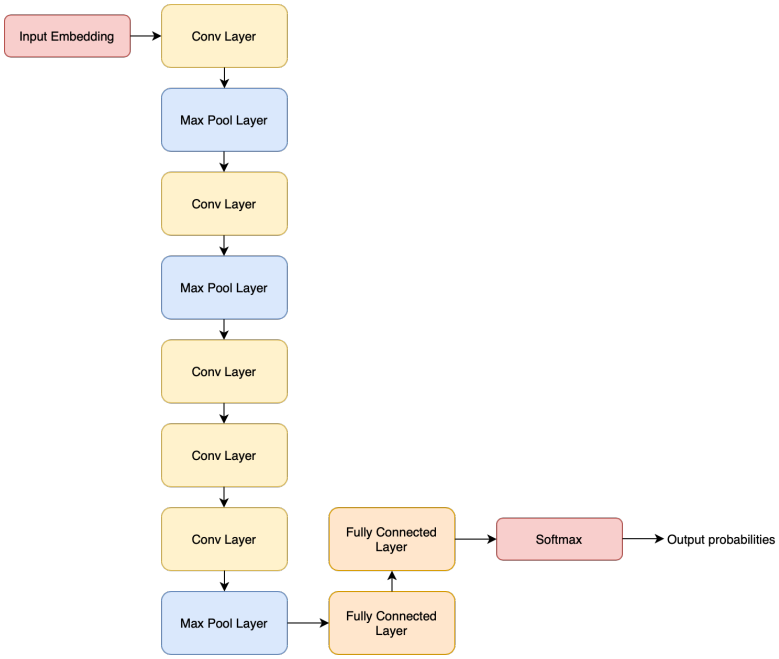


Fig. 3.6 The architecture based on the AlexNet model [4]

encoder. However, the decoder introduces an additional third sub-layer that conducts multi-head attention over the output of the encoder stack. Like the encoder, the decoder also integrates residual connections around each sub-layer, followed by layer normalization.

Regarding the input data, 80% of the study data was allocated for model development, with the remaining 20% reserved for a test set. The development data was further divided into training and validation subsets, with an 80-20 split. Stratified splits were used to maintain a balanced distribution of positive and negative cases in each subset. The datasets were further balanced by randomly selecting negative cases to match the number of positive cases.

To assess the models' sensitivity to the choice of training instances, we performed 10-fold cross-validation on models with optimized hyperparameters. To evaluate the impact of random initialization, the final model training was repeated five times. Subsequently, the classification performance figures presented in this work represent the test results obtained on the held-out test set.

B. AlexNet

AlexNet [241, 286] is a deep convolutional neural network (CNN) architecture that played a pivotal role in the advancement of image classification and object recognition tasks. Developed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, it was the winning entry in the ImageNet Large Scale Visual Recognition Challenge in 2012. Here's a detailed explanation of the AlexNet network (see Figure 3.6):

Architecture:

- AlexNet is a deep neural network consisting of eight layers: five convolutional layers and three fully connected layers. It was one of the first CNNs to demonstrate the effectiveness of deep learning on image classification tasks.

Convolutional Layers:

- The first convolutional layer takes the input image, which is typically of size 224x224 pixels. It applies 96 filters of size 11x11 with a stride of 4 and a rectified linear unit (ReLU) activation function. This is followed by max-pooling with a 3x3 filter. The subsequent convolutional layers are similarly structured with different numbers of filters (256, 384, 384, and 256, respectively), ReLU activation functions, and max-pooling.

Fully Connected Layers:

- After the convolutional layers, AlexNet has three fully connected layers. These layers have 4096 neurons each, with ReLU activation functions. The last fully connected layer has 1000 neurons, corresponding to the 1000 classes in the ImageNet dataset. These layers are followed by a softmax activation function to obtain class probabilities.

Local Response Normalization (LRN):

- AlexNet incorporates a local response normalization layer after the first and second convolutional layers. This layer helps in normalizing the activations within a local neighborhood of each neuron, which enhances the network's response to variations in input.

Dropout:

- Dropout is applied to the fully connected layers to prevent overfitting. It randomly drops a fraction of neurons during training, reducing co-dependency among neurons and improving the network's generalization.

Rectified Linear Unit (ReLU):

- AlexNet uses the ReLU activation function in its convolutional and fully connected layers. ReLU introduces non-linearity into the network, which helps it capture complex features and accelerates training.

Max-Pooling:

- Max-pooling is applied after each convolutional layer, which reduces the spatial dimensions of the feature maps and extracts dominant features.

Cross-Channel Normalization:

- Cross-channel normalization is performed after max-pooling in some layers to enhance the model's generalization capabilities.

Output:

- The final output of AlexNet is a probability distribution over the 1000 classes in the ImageNet dataset. It uses a softmax activation function to obtain class probabilities.

Training:

- AlexNet was trained on the ImageNet dataset, which consists of millions of labeled images, to learn the features and parameters of the network.

Impact:

- AlexNet was a breakthrough in deep learning and had a profound impact on the field of computer vision. It demonstrated that deep neural networks can outperform traditional machine learning methods in image classification tasks, leading to a surge in the development of deep learning models for a wide range of applications.

AlexNet's success laid the foundation for the development of deeper and more sophisticated convolutional neural networks, ultimately leading to advancements in image recognition, object detection, and other computer vision tasks. It is considered one of the seminal models in the deep learning landscape.

C. ResNet50

ResNet50 [5, 287], short for "Residual Network with 50 layers," is a powerful convolutional neural network (CNN) architecture. It is part of the ResNet family, which was introduced by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun in their 2015 paper "Deep Residual Learning for Image Recognition." ResNet-50 is one of the most well-known variants of the ResNet architecture, and it is specifically designed for image classification tasks. Here's a detailed explanation of ResNet-50 (see Figure 3.7):

Residual Learning:

- The key innovation of ResNet is residual learning. In traditional deep networks, as the network depth increases, performance can degrade due to the vanishing gradient problem. Residual networks address this by introducing residual blocks. Instead of learning the desired output of a layer, ResNet learns the residual, the difference between the desired output and the actual output. The network is trained to make the residual close to zero. This approach simplifies the training process and enables the successful training of very deep networks.

Architecture:

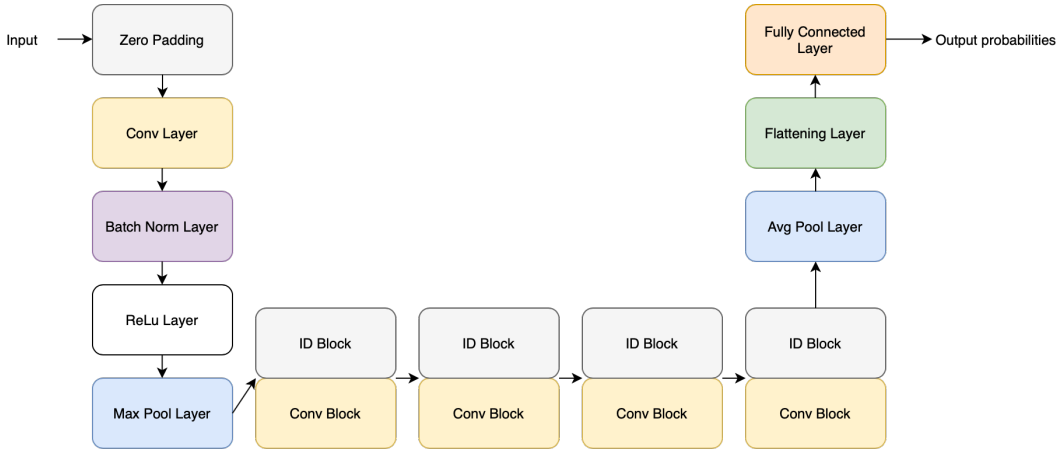


Fig. 3.7 The architecture based on the ResNet50 model [5]

- ResNet50 is a deep network consisting of 50 layers. These layers are organized into blocks, and each block contains several convolutional layers, batch normalization, and ReLU activation functions. The network is divided into five stages, with varying numbers of residual blocks in each stage. These blocks vary in terms of their depth and the number of filters.

Identity and Convolutional Blocks:

- Within each block, there are two main types of blocks: identity blocks and convolutional blocks. Identity blocks are used when the input and output dimensions are the same. They consist of two 3x3 convolutional layers. Convolutional blocks are used when the input and output dimensions differ. They include a 1x1 convolution layer to match the dimensions and are followed by two 3x3 convolutional layers.

Bottleneck Design:

- To make the network computationally efficient, ResNet-50 utilizes a bottleneck design. It employs 1x1, 3x3, and 1x1 convolutions within each convolutional block. The 1x1 convolutions are used to reduce and then restore the dimensions, which reduces the number of parameters and computational load.

Global Average Pooling:

- After the convolutional layers, global average pooling is applied to reduce the spatial dimensions to a 1x1 feature map for each channel. This feature map is then flattened and passed to the final fully connected layer for classification.

Output:

- The output layer typically consists of 1000 neurons (for ImageNet classification), each corresponding to a different class. A softmax activation function converts the network's final logits into class probabilities.

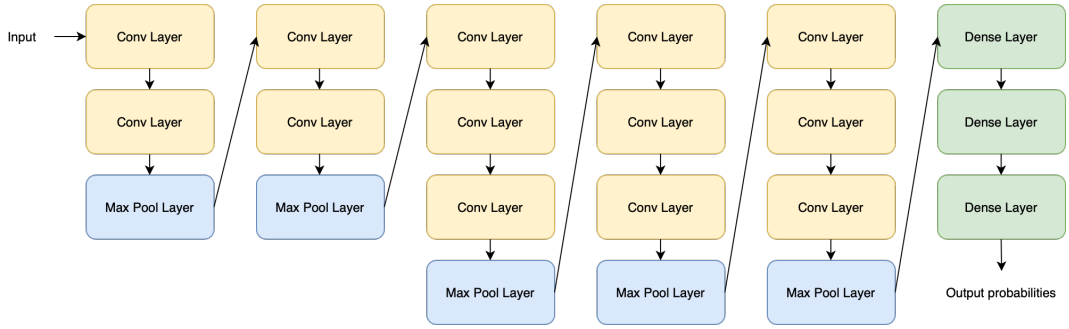


Fig. 3.8 The architecture based on the VGG16 model [6]

Training:

- ResNet-50 is trained on large datasets, such as ImageNet, using techniques like stochastic gradient descent (SGD) with momentum and learning rate schedules. Data augmentation and batch normalization are commonly used to improve the training process.

Transfer Learning:

- Due to its impressive performance and generalization capabilities, ResNet-50 is often used as a feature extractor or for transfer learning in various computer vision tasks.

Applications:

- ResNet-50 is a versatile architecture used for image classification, object detection, and various other computer vision tasks. It has set performance benchmarks in these domains and continues to be a popular choice for deep learning practitioners.

ResNet-50 is a significant advancement in the field of deep learning, and its success has paved the way for even deeper and more complex neural networks. Its residual learning concept and efficient bottleneck design have greatly improved the training and performance of deep convolutional networks.

D. VGG16

VGG16 [242, 288, 6] is a convolutional neural network (CNN) architecture that was developed by the Visual Geometry Group (VGG) at the University of Oxford. It is known for its simplicity and effectiveness in image classification tasks. VGG16 is a variant of the VGG family of models and is specifically designed for image recognition. Here's a detailed explanation of the VGG16 architecture (see Figure 3.8):

Architecture:

- VGG16 is characterized by its deep architecture with 16 layers. These layers consist of 13 convolutional layers and 3 fully connected layers. The network has a straightforward and uniform structure.

Convolutional Layers:

- VGG16 primarily uses 3x3 convolutional filters, which is a smaller filter size compared to some other architectures. These smaller filters allow the network to capture finer details in the input images. The network stacks multiple convolutional layers together before applying max-pooling to reduce the spatial dimensions.

Pooling Layers:

- After each set of convolutional layers, max-pooling layers are applied to reduce the spatial dimensions of the feature maps. VGG16 uses 2x2 max-pooling with a stride of 2x2, which effectively reduces the size by half.

Fully Connected Layers:

- VGG16 has three fully connected layers. The first two have 4096 neurons each, followed by a third fully connected layer with 1000 neurons, which is the output layer for the 1000 classes in the ImageNet dataset (a common benchmark for image classification). The final fully connected layer uses softmax activation to produce class probabilities.

Activation Function:

- Throughout the network, VGG16 employs the rectified linear unit (ReLU) activation function. ReLU introduces non-linearity into the model and accelerates training.

Uniform Structure:

- One of VGG16's distinguishing characteristics is its uniform structure. The network repeats a sequence of convolutional and pooling layers several times, creating a consistent and regular architecture. VGG16 uses a series of 'convolutional blocks' with varying depths, followed by fully connected layers. The repetition of blocks allows it to capture features at multiple scales and complexities.

Image Input Size:

- VGG16 expects input images to be of size 224x224 pixels, which is common for many CNN architectures.

Pre-Trained Model:

- VGG16 is often used as a pre-trained model for transfer learning. Pre-training is typically done on a large dataset, like ImageNet, and then fine-tuned on a specific task with a smaller dataset.

Performance and Applications:

- VGG16 achieved excellent performance in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2014. It played a key role in advancing image classification and object recognition. VGG16 has been used in various computer vision tasks, such as image classification, object detection, and feature extraction. It serves as a strong baseline for many image-related problems.

Model Variants:

- The VGG family includes variants with different depths, such as VGG16 and VGG19. VGG16's simplicity and effectiveness have made it a popular choice in computer vision, but the deeper variants, like VGG19, offer improved performance at the cost of increased computational complexity.

In summary, VGG16 is a widely recognized and influential deep learning architecture that has played a pivotal role in image recognition and classification. Its simplicity, uniform structure, and strong performance make it a valuable tool for various computer vision applications.

3.1.4 Statistical tests

In this section, we delve into the intricacies of statistical tests, with a particular focus on two powerful non-parametric methods: the Wilcoxon signed-rank test and the Mann-Whitney U test. As we navigate through the statistical landscape, these tests emerge as indispensable tools for analyzing data when assumptions of normality are not met or when dealing with ordinal and non-normally distributed data. The Wilcoxon test proves invaluable for paired samples, offering a robust alternative to the parametric paired t-test. Meanwhile, the Mann-Whitney U test stands as a key player in comparing independent samples without relying on the normality assumption. Join us in exploring the rationale, application, and interpretation of these tests, as we unravel their significance in statistical analysis within the context of our research.

3.1.4.1 Wilcoxon test

A Wilcoxon test, also known as the Wilcoxon signed-rank test [289], is a non-parametric statistical hypothesis test used to compare two related samples or paired data. It's particularly useful when the data do not follow a normal distribution, and it's employed to determine if there is a significant difference between the two groups. The test is named after Frank Wilcoxon, who introduced it in 1945. Here's a detailed explanation of the Wilcoxon test:

A. Assumptions:

The Wilcoxon test is a non-parametric test, which means it does not assume that the data follows a specific probability distribution, such as the normal distribution. This makes it robust to data with non-normal distributions.

B. Use Case:

The Wilcoxon test is typically used when you have paired data or two samples that are not independent. For example, it can be used to compare the performance of the same group of individuals before and after a treatment or to compare two different methods applied to the same subjects.

C. Hypotheses:

The test evaluates the null hypothesis (H0) that there is no significant difference between the two samples, and the alternative hypothesis (H1) that there is a significant difference.

D. Rank Transformation:

To perform the test, you first calculate the absolute differences between the pairs of data points and then rank these absolute differences from the smallest to the largest. Ties (data points with the same values) are handled by assigning them the average of the ranks they would occupy if they were unique values.

E. Test Statistic:

The test statistic is calculated based on the sum of the ranks of the signed differences. If you're comparing two groups of data, one will have positive differences and the other negative. The test statistic (W) is the sum of the ranks of the signed differences.

F. Comparison to Critical Values:

The test statistic is compared to critical values from the Wilcoxon signed-rank table or calculated based on sample size and chosen significance level (alpha). If the calculated W value is greater than the critical value, you reject the null hypothesis.

G. Significance Level (Alpha):

The significance level (alpha) is chosen in advance, typically set at 0.05 or 0.01, depending on the desired level of significance.

H. Interpretation:

If the test statistic falls below the critical value, you fail to reject the null hypothesis, suggesting that there is no significant difference between the paired samples. If the test statistic is greater than the critical value, you reject the null hypothesis, indicating a significant difference between the paired samples.

I. Effect Size:

The Wilcoxon test does not provide an effect size directly, but you can compute it by calculating a statistic such as Cohen's d to quantify the magnitude of the observed difference.

J. Assumptions:

The Wilcoxon test does assume that the data pairs are independent of each other and that the differences are symmetrically distributed around zero.

K. Software:

Statistical software packages, like R, Python (with libraries like SciPy), and specialized statistical software, can perform the Wilcoxon test automatically.

The Wilcoxon test is a valuable tool in statistics when you need to compare two related samples or paired data and cannot rely on the assumption of a normal distribution. It is commonly used in various fields, including medicine, social sciences, and engineering, for analyzing data with paired observations.

3.1.4.2 Mann-Whitney U test

The Mann-Whitney U test, also known as the Mann-Whitney-Wilcoxon test [290], is a non-parametric statistical test used to compare two independent samples or groups to determine if they are significantly different from each other. This test is often applied when the data do not follow a normal distribution and is particularly useful for ordinal or interval data. The Mann-Whitney test was introduced by Henry Mann and Donald Ransom Whitney in 1947. Here's a detailed explanation of the Mann-Whitney U test:

A. Assumptions:

The Mann-Whitney U test is a non-parametric test, which means it does not rely on the assumption of a normal distribution. It is suitable for data with non-normal distributions.

B. Use Case:

The test is typically used to compare two independent groups to assess whether they come from the same population or if one group tends to have higher or lower values than the other. It is often used in situations where you have two separate samples and want to determine if they differ in some way.

C. Hypotheses:

The Mann-Whitney test evaluates the null hypothesis (H0) that there is no significant difference between the two groups, and the alternative hypothesis (H1) that there is a significant difference between the two groups.

D. Rank Transformation:

To perform the test, you combine the two samples and rank all the data points from lowest to highest. Ties (data points with the same values) are handled by assigning them the average of the ranks they would occupy if they were unique values.

E. Test Statistic (U statistic):

The Mann-Whitney U test calculates a test statistic called the U statistic. The U statistic is based on the ranks assigned to data points in the two groups and their sum. It measures the magnitude of the differences between the two groups.

F. Calculation of U Statistic:

The U statistic is calculated separately for each group. The smaller U value represents the group with the lower values. The test statistic for the Mann-Whitney test is then the U value of the smaller group (the group with lower values).

G. Comparison to Critical Values:

The test statistic (U value) is compared to critical values from the Mann-Whitney U table or calculated based on sample sizes and chosen significance level (alpha). If the U value is less than or equal to the critical value, you fail to reject the null hypothesis, suggesting no significant difference between the groups. If the U value is greater than the critical value, you reject the null hypothesis, indicating a significant difference between the groups.

H. Significance Level (Alpha):

The significance level (alpha) is chosen in advance and represents the level of significance for the test. Common values are 0.05 or 0.01.

I. Effect Size:

The Mann-Whitney U test does not provide an effect size directly, but you can compute it using statistics like the common language effect size (CL) to quantify the magnitude of the observed difference.

J. Software:

Statistical software packages like R, Python (with libraries like SciPy), and specialized statistical software can perform the Mann-Whitney U test automatically.

The Mann-Whitney U test is a valuable tool in statistics when you need to compare two independent samples without relying on the assumption of a normal distribution. It is commonly used in various fields, such as psychology, biology, and social sciences, for analyzing data from two independent groups to determine if they differ significantly.

3.2 Databases

The research field related to the analysis of adventitious sounds faces a significant issue concerning the lack of standardized adventitious sound databases. This means that each author uses their own database of adventitious sounds, and as a result, each study employs a different database from other proposals. Consequently, the performance evaluation of various state-of-the-art methods is achieved by implementing different algorithms and subsequently assessing them using their own respective databases.

Given this problem, and following an extensive literature review, it has been concluded that authors in the field of adventitious sounds analysis typically design their databases using three possible methodologies:

- On one hand, some authors create their databases from respiratory sound signals obtained from different patients with obstructive lung pathologies who exhibit

adventitous sounds. From a medical perspective, this is the most effective, reliable, and robust way to build a database of adventitious sounds. However, not all authors have access to a medical team that provides this possibility. Additionally, creating databases from real patients is a time-consuming process. First, patient consent must be obtained. Second, multiple recordings need to be made until adventitous sounds are generated. Lastly, it's important to note the significant amount of time medical professionals must invest in analyzing and detecting every adventitious sounds.

- On the other hand, the second alternative is the sharing of databases among authors, including in the publication the corresponding reference to the work that defines the database in question and acknowledgments to the authors who have facilitated the use of that database.

- Finally, authors who cannot pursue the previous option create their own databases from online repositories of adventitious sounds and books that include adventitious sounds as an additional resource. In this regard, the rest of this section presents the main online repositories and the bibliography most commonly used by authors in this research field for creating adventitious sound databases. Additionally, it describes a database of wheezing and crackling sounds that has emerged recently.

3.2.1 Online repositories of respiratory sounds

There are several online repositories that provide access to different types of respiratory sounds, including both normal and adventitious respiratory sounds, such as wheezing. Undoubtedly, this is the quickest option that any researcher focusing on wheezing sound analysis can use to build their own database. Table 3.1 presents the main repositories of respiratory sounds that researchers often utilize in this field. Among these repositories, we can find companies specializing in stethoscope design (Thinklabs, Littmann, or Stethographics), software dedicated to respiratory signal processing (RALE), and more.

3.2.1.1 ICBHI Database

Recently, a database of respiratory sounds has been published [2], mainly composed of normal respiratory sounds, wheezes, and crackles. This database was originally created to support the scientific challenge organized by the International Conference on Biomedical Health Informatics (ICBHI) 2017. Since 2019, the public and private dataset of the ICBHI challenge has been available for free [173], in recent years, there has been a significant increase in research based on different machine learning approach, such as Recurrent neural networks (RNN) [60], Hybrid neural networks

Database or Author Name - Ref.	Country	Participants Number(Total (M/F); HC)	Abnormal Lung Sounds Labeled	(Pathologies Labeled	Availability
ICBHI Challenge [173]	Portugal and Greece	-	Crackles and Wheezes	Asthma, COPD,Bronchiolitis, Laryngeal web, Bronchogenic carcinoma, Lung fibrosis, Cystic fibrosis	Available online
R.A.L.E. Lung Sounds 3.2 [291]	Canada	70 (-); 17	Crackles, Wheezes, Squawk, Stridor, Rhonchi	Asthma, COPD,Bronchiolitis, Laryngeal web, Bronchogenic carcinoma, Lung fibrosis, Cystic fibrosis.	Available online
KAUH database [292]	Jordan	120 (43/69); 35	Crackles, Wheezes, Crepitations, Bronchial sounds, Crackles + Wheezes, Crackles + Bronchial	Asthma, Pneumonia, COPD, Bronchitis, Heart failure, Lung fibrosis, Pleural effusion	Available online
Respiratory Database@TR [293]	Turkey	77 (64/13); 30	Crackles, Wheezes	Asthma, COPD	Available online
Thinklabs Lung Sounds Library [294]	United States	-	Crackles, Wheezes, Pleural rub, Rhonchi, Stridor	Asthma, Bronchiolitis, COPD, Laryngomalacia, Pulmonary edema	Available online
East Tennessee State University Pulmonary Breath Sounds [295]	United States	-	Crackles, Pleural rub, Stridor, Wheezing, Rhonchus		Available online
ASTRA database	France	-	-	-	CD-ROM
Auscultation Skills: Breath and Heart Sounds [296]	United States	-	-	-	CD-ROM
Fundamentals of Lung and Heart Sounds [297]	United States	-	-	-	CD-ROM
Heart and Lung Sounds Reference Library, Wrigley [298]	United States	-	Bronchial, Bronchovesicular, Rhonchi, Pneumonia, Wheezes, Bronchophony, Crackles, Stridor,	-	CD-ROM
Understanding Lung Sounds, Lehrer [299]	United States	-	Crackles, Wheezes	-	CD-ROM
Bahoura 1999 [300]	France	-	-		Undefined
Hsiao 2020 [301]	Taiwan	22 (12/10); -	Crackles, Wheezes	-	Undefined
Bogazici University Lung Acoustics Laboratory	Turkey	-	-	Bronchiectasis, Interstitial lung disease	Undefined
CORA database [302]	Ukraine	-	-	Bronchitis, COPD	Undefined
Stethographics Lung Sound Samples 2	United States	-	-	-	Undefined
3M Littmann Lung Sounds Library [90]	United States	-	-	-	Undefined
Mediscuss Respiratory Sounds 2	-	-	-	-	Undefined

Table 3.1 A comprehensive overview of availables databases online. Table provided by [12].

Number of cycles	Total
Crackles	1.864
Wheezes	886
Crackles + Wheezes	506
Normal	3.642
Total number of cycles	6.898

Table 3.2 Cycle breakdown of ICBHI 2017 challenge dataset.

[303, 219, 304, 305, 306] and above all Convolutional neural networks (CNN) [307, 212, 308, 309, 224, 215, 214, 310, 311, 225, 216, 217, 218, 219, 312, 226, 313, 314, 315, 227, 316, 317, 318, 228, 319, 320, 321, 322, 323, 324]. As described in [269], the ICBHI database was constructed with the aim of supporting scientific contributions dedicated to the classification of respiratory sounds (normal respiratory sounds, wheezes, and crackles) and eliminating the lack of a standardized database.

The ICBHI database contains audio samples collected independently by two research teams over several years. Most of the database consists of audio recordings made by the research team at the Faculty of Health Sciences of the University of Aveiro (ESSUA), collected at the Research and Rehabilitation Respiratory Laboratory (Lab3R) of ESSUA and at the Infante D. Pedro Hospital in Aveiro, Portugal. The second research team, consisting of Aristotle University of Thessaloniki (AUTH) and the University of Coimbra (UC), acquired respiratory sounds at the General Hospital of Papanikolaou, Thessaloniki, Greece, and at the General Hospital of Imathia (Health Unit of Naousa), Greece.

The database consists of 5.5 hours of respiratory sounds, with a total of 6898 respiratory cycles (one respiratory cycle consists of the inspiration and expiration stages), of which 1864 contain crackles, 886 contain wheezes, and 506 contain both crackles and wheezes (see Table 3.2). These sounds were obtained from 920 audio recordings of 126 different subjects. The recordings were collected using a variety of equipment, including three stethoscopes (3M Littmann Classic II SE Stethoscope, 3M Littmann 3200 Electronic Stethoscope, and WelchAllyn Meditron Master Elite Electronic Stethoscope) and a condenser microphone (AKG C417L Microphone), with durations ranging from 10 to 90 seconds. This database also indicates the thoracic locations from which the recordings were acquired. It should be noted that respiratory cycles were labeled by medical experts in the field, who determined, for each respiratory cycle, whether there were wheezes, crackles, a combination of them, or no adventitious respiratory sound. In addition to the demographic information of the patients (age, gender, weight, height, and BMI), the database also includes the diagnosis for each of them.

This database has been created to assess typical acoustic scenarios that occur in the real world. Therefore, in the respiratory audio signals, various types of sounds that act

Table 3.3 Overview of the simulated respiratory sounds database. K_C: number of crackles per signal. $NOTS$: number of signals per SNR. N_S: number of signals generated taking into account all SNRs evaluated.

Scenario	Type	Model	K_C	$NOTS$	Noise	Diagnosis	SNR	N_S
Simulated	FCS [326]	ATS	10	15	N_R	-	[−10 dB, 10 dB]	315
	FCS [127]	Hoevers	10	15	N_R	-	[−10 dB, 10 dB]	315
	FCS [325]	Cohen	10	15	N_R	-	[−10 dB, 10 dB]	315
	CCS [326]	ATS	10	15	N_R	-	[−10 dB, 10 dB]	315
	CCS [127]	Hoevers	10	15	N_R	-	[−10 dB, 10 dB]	315
	CCS [325]	Cohen	10	15	N_R	-	[−10 dB, 10 dB]	315
Real	FCS [9, 7]	-	10	15	N_R	IPF	[−10 dB, 10 dB]	315
	CCS [9, 7]	-	10	15	N_R	BE	[−10 dB, 10 dB]	315

as background noise can be heard, such as people talking, children crying, etc.

3.2.2 Simulated repositories of respiratory sounds

In [9, 7] a new concept of database is introduced. The authors present five simulated scenarios for the creation of a dataset composed of 2520 signals based in crackles adventitious sounds; i) simulated signals related to fine and coarse (FCSs and CCSs) according to [127, 325, 326]; (ii) real signals related to fine and coarse (FCSs and CCSs) extracted from lung sound recordings. Specifically, real FCSs were extracted from a patient with idiopathic pulmonary fibrosis (IPF), and real CCSs were selected from a patient with bronchiectasis (BE) [7]. Each simulated and real signal has been mixed with noise N_R that shows the same spectral energy distribution as typically found in breath noise from a healthy subject measured over the lung bases on the right-hand side of the back as occurs in [7]. Moreover, several signal-to-noise ratios (SNRs) ranged from -10 to 10 dB in steps of 1 dB have been created to evaluate the robustness of the proposed method detecting crackles. For each SNR, 15 simulated signals of every scenario have been evaluated, considering the effect of random variations of the local SNR around any given crackle. In this manner, a set of 315 signals have been generated considering all the SNRs for each type of simulated or real crackle signal. Although the dataset ψ is detailed in Table 3.3, more details can be found in [9, 7].

3.3 Metrics

In any field of research, establishing an appropriate evaluation methodology is essential to assess the reliability of proposed solutions. This section describes the main objective metrics used to measure the performance of algorithms dedicated to addressing the primary tasks related to adventitious sound analysis: separation, detection, and classifi-

cation of adventitious sounds during the auscultation process. These metrics allow for the evaluation of the detection performance of algorithms designed to detect the presence or absence of adventitious sounds in respiratory audio signals. In other words, those algorithms that provide discrimination or classification between respiratory signals from healthy subjects (without adventitious sounds) and those from sick subjects (with adventitious sounds).

Specifically, the most relevant metrics [94, 33] for measuring the ability to discriminate between healthy and sick subjects are as follows:

- Accuracy (Acc) measures the number of correctly classified adventitious sounds and normal respiratory sounds cycles from the total number of test samples.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.20}$$

- Sensitivity (Sen) is defined as the number of correctly detected adventitious sounds class from the total number of predicted adventitious sound events.

$$Sen = \frac{TP}{TP + FN} \tag{3.21}$$

- Precision (Pre) is defined as the positive predictive value (PPV) where a true positive is considered as the target event, when the test makes a positive forecast, and the subject has a positive result.

$$Prec = \frac{TP}{TP + FP} \tag{3.22}$$

- Specificity (Spe) represents the correctly labeled normal respiratory sound events (TN) from the total number of normal respiratory sound events (TN + FP).

$$Spe = \frac{TN}{TN + FP} \tag{3.23}$$

- Score (Sco) represents a general measure of the quality of the classifier as an average of the sensitivity and specificity metrics.

$$Sco = \frac{Sen + Spe}{2} \tag{3.24}$$

3.4 Summary of the State of the Art for wheezing and crackles detection

Throughout this chapter, we conducted a comprehensive review of the literature pertaining to various techniques employed in the preprocessing, feature extraction, and

classification stages for the identification and categorization of respiratory sounds. Table 3.4 provides a comprehensive comparison between the proposed method and several state-of-the-art techniques for evaluating the classification performance in the context of respiratory sound analysis, specifically targeting the differentiation of four classes: normal, wheezes, crackles, and the combination of crackles and wheezes. The evaluation is conducted on the ICBHI database, with a focus on respiratory cycles (RC) represented by the temporal length in seconds, including zero padding to ensure fixed-duration respiratory cycles.

Table 3.4 gives a discerning examination of various machine learning algorithms, elucidated in previous studies within the field, is evident. The selected algorithms represent diverse methodologies employed for the analysis of respiratory sound. These algorithms are systematically categorized across multiple works, each contributing to the overarching landscape of respiratory sound analysis.

The table systematically enumerates the methodologies employed in these studies, emphasizing the distinct time-frequency representations utilized, such as Short-Time Fourier Transform (STFT), Mel-Frequency Cepstral Coefficients (Mel), Wavelet, Scalogram, Spectrogram, and Cochleagram. Additionally, it specifies the temporal length of respiratory cycles (RC) in seconds, accounting for zero-padding to achieve a standardized duration.

The machine learning techniques deployed are thoroughly outlined, encompassing a spectrum of approaches including Hidden Markov Models (HMM), Recurrent Neural Networks (RNN), Support Vector Machines (SVM), Residual Networks (ResNet), and Convolutional Neural Networks (CNN), among others. These techniques serve as pivotal components in the development and implementation of models for respiratory sound classification.

Moreover, the table provides insights into the training and testing methodologies, delineating the ratios employed for data partitioning, such as 80/20 splits or specific fold configurations in k-fold cross-validation settings.

Critical performance metrics, including Sensitivity (Sen), Specificity (Spe), Score (Sco), and Accuracy (Acc), are meticulously documented for each respective algorithm and study. These metrics serve as crucial indicators of the efficacy of the implemented methodologies in accurately classifying respiratory sounds across the normal, wheezes, crackles, and combined crackles and wheezes classes.

Furthermore, the table underscores key findings and outcomes from each study, offering a comprehensive overview of the diversity in approaches and corresponding classification results.

Authors	Time-frequency representation		RC (s)	Technique	Train/Test	Results (%)			
	Type	Parameters				Sen	Spe	Sco	Acc
[211]	STFT	30 ms	-	HMM	60/40	–	–	39.6	–
[212]	STFT	500 ms	-	RNN	- (5-fold)	58.4	73.0	65.7	–
[213]	STFT	512 ms	-	HMM SVM	60/40	20.81	78.5	49.65	49.43
[224]	Mel	250 ms	–	RNN	80/20	**64.0**	**84.0**	**74.0**	–
[214]	STFT, Wavelet	20 ms, $D_2 - D_7, A_7$	–	bi-ResNet	– (10-fold)	31.1	69.2	50.2	52.8
[215]	STFT, Scalogram	40 ms	–	CNN	60/40	28.0	81.0	54.0	–
[216]	STFT	$64 - 128 - 524$ ms	–	CNN SVM	– (10-fold)	–	–	–	65.5
[217]	STFT	20 ms	–	ResNet NL	60/40	41.3	63.2	52.3	–
[225]	Mel	60 ms	–	CNN RNN	80/20	–	58.01	–	–
[218]	STFT	100 ms	2.5	ResNet SE SA	70/30	17.8	81.3	49.6	–
[226]	Mel	–	–	CNN	60/40	–	–	–	**80.4**
[219]	STFT	40 ms	–	CNN bi-LSTM	– (5-fold)	63.0	83.0	73.0	–
[312]	Wavelet	30 ms	–	DAG HMM	–	–	–	–	50.1
[227]	Mel	–	7	CNN	60/40	40.1	72.3	56.2	–
[10]*	STFT	32 ms 64 filters	6	CNN	80/20 (10-fold)	51.61	65.45	58.53	60.61
	Mel			CNN		47.83	63.33	55.58	57.56
	STFT + Mel			CNN		46.97	63.97	55.47	57.33
[228]	STFT, Log-mel	32 ms, 50 bins	8	ResNet	60/40	37.2	79.3	58.3	–

Table 3.4 Comparison between the state-of-the art methods evaluating the four-classes (normal vs. wheezes vs. crackles vs. crackles+wheezes) classification performance in the ICBHI database. Respiratory cycle (RC) represents the temporal length (in seconds) including zero padding to create respiratory cycles of fixed duration. bi-ResNet: bilinear ResNet, NL: non-local, SE: Squeeze-and-Excitation, SA: Spatial Attention, bi-LSTM: bi-directional LSTM, DAG: Directed Acyclic Graph. The rest of the acronyms have been previously mentioned. The references followed by $*$ means that the method has been implemented in this Thesis following the authors description. The results for other methods have been directly extracted from the corresponding works. In bold letter is indicated the maximum value for each metric.

3.5 Conclusions

In this chapter, a state-of-the-art review has been presented, focusing on the works that have addressed the primary tasks of interest in the analysis of adventitious sounds: the preprocessing phase, feature extraction and the classifiers typically employed in previous works, as well as the ones, used in this Thesis.

Next, the issues related to the shortage of standardized databases in this scientific field have been presented. Additionally, the main metrics used in the evaluation of algorithm performance have been described, considering the primary tasks related to adventitious sound analysis.

CHAPTER 4

Automatic Robust Crackle Detection and Localization Approach Using AR-Based Spectral Estimation and Support Vector Machine

4.1 Abstract

Auscultation primarily relies upon the acoustic expertise of individual doctors in identifying, through the use of a stethoscope, the presence of abnormal sounds such as crackles because the recognition of these sound patterns has critical importance in the context of early detection and diagnosis of respiratory pathologies. In this paper, we propose a novel method combining autoregressive (AR)-based spectral features and a support vector machine (SVM) classifier to detect the presence of crackle events and their temporal location within the input signal. A preprocessing stage is performed to discard information out of the band of interest and define the segments for short-time signal analysis. The AR parameters are estimated for each segment to be classified by means of support vector machine (SVM) classifier into crackles and normal lung sounds using a set of synthetic crackle waveforms that have been modeled to train the classifier. A dataset composed of simulated and real coarse and fine crackles sound signals was created with several signal-to-noise (SNR) ratios to evaluate the robustness of the proposed method. Each simulated and real signal was mixed with noise that shows the same spectral energy distribution as typically found in breath noise from a healthy subject. This study makes a significant contribution by achieving competitive results. The proposed method yields values ranging from 80% in the lowest signal-to-noise ratio scenario to a perfect 100% in the highest signal-to-noise ratio scenario. Notably, these results surpass those of other methods presented by a margin of at least 15%. The combination of an autoregressive (AR) model with a support vector machine (SVM) classifier offers an effective solution for detecting the presented events. This approach exhibits enhanced robustness against variations in the signal-to-noise ratio that the input signals may encounter.

4.2 Contribution

In this work, our primary focus lies in the exploration of the autoregressive (AR)-based frequency features that form the foundation for characterizing the spectral envelope of a breathing signal. We delve into the intricate world of these features and their potential in enhancing the performance of machine learning classifiers. Notably, our research introduces a novel approach involving the utilization of complex-valued poles derived from the AR model as inputs for a Support Vector Machine (SVM) classifier employing the Radial Basis Function (RBF) kernel.

The motivation behind this approach stems from the need to advance the state-of-the-art in the domain of signal processing and pattern recognition. We recognize that the spectral characteristics of breathing signals are rich with information, and harnessing this data can substantially enhance the accuracy and efficiency of event detection.

This evaluation focuses on the examination and assessment of three distinct methods: the Iterative Envelope Mean-Fractal Dimension (IEM-FD), the Time-Varying Autoregressive (TVAR) approach, and a novel method that is introduced as part of this research. These methods serve as the central subjects of investigation, and their individual performance and capabilities are meticulously scrutinized in the context of the study's objectives and research goals.

- IEM-FD[7]: The Iterative Envelope Mean-Fractal Dimension method is one of the key players in this evaluation. It brings to the table a well-established approach for analyzing and characterizing signals, particularly in the context of identifying and understanding certain patterns or features within the data. This method represents a benchmark against which the other methods will be compared.

- TVAR [8]: The Time-Varying Autoregressive method, another prominent contender in this assessment, offers its own set of techniques and tools for analyzing time-series data. It is known for its adaptability in modeling and capturing temporal variations, making it an essential component of the comparative analysis.

- Proposed New Method: Alongside the established methods, a novel approach, developed as part of this research, is introduced. This new method presents innovative and potentially groundbreaking techniques for addressing the specific challenges or goals outlined in the study. It is a product of the researcher's creativity and problem-solving abilities, making it a key highlight of the evaluation.

Each of these methods contributes its unique strengths, algorithms, and insights to the overall evaluation, enabling a comprehensive understanding of their relative performance and suitability for the research objectives. The comparison among these methods

aims to shed light on their respective advantages, limitations, and potential contributions to the field of study.

Furthermore, our study goes one step further by integrating the AR-based features with a state-of-the-art Convolutional Neural Network (CNN) architecture, as referenced in Rocha et al. (2020) [10]. This integration serves as a testament to the versatility and adaptability of our approach, as it demonstrates compatibility with cutting-edge deep learning models. The synergistic combination of AR-based features and a CNN architecture has the potential to unlock new horizons in event detection and classification tasks, propelling the field into the era of advanced data-driven analysis.

It's worth noting that the choice of classifier is critical, particularly in scenarios with limited data quantities. In this context, SVM stands out as a preferable choice due to its remarkable capacity to generalize from limited examples. Deep learning models, on the other hand, often come with a significant number of tunable weights, making them susceptible to overfitting when the number of weights approaches or exceeds the number of available training examples. In contrast, SVM, especially when coupled with a well-suited kernel function like the RBF, offers a robust, efficient, and easily interpretable solution, making it less prone to overfitting in classification problems. This inherent advantage underscores the significance of our approach in addressing real-world data challenges and presenting a viable alternative to more complex deep learning models.

4.2.1 Modelling of Simulated Crackle Sounds

As mentioned above, we based our model training and testing in an simulated database that we created in order to simulate the real life conditions of crackle sounds. This approach involves creating artificial crackle waveforms to generate training data for the classifier. Breaking down the process step by step it can be seen as:

- Generating Synthetic Crackle Sounds: To train the classifier, synthetic crackle sounds are produced. These sounds mimic the acoustic properties of real crackles, but they are artificially generated for the purpose of creating a training dataset.
- Crackle Waveform Creation: The core of this process involves creating a crackle waveform denoted as $y(t)$. Equations (4.1) to (4.4) define the mathematical representation of this waveform. These equations contain the parameters necessary for generating the crackle sound.
- Assumptions about the Crackle Sound: Several assumptions are made to create the synthetic crackle sound. These assumptions include:
  - Two Cycles: The crackle sound is assumed to consist of two cycles. The parameter t_{2CD} represents the duration of these two cycles.

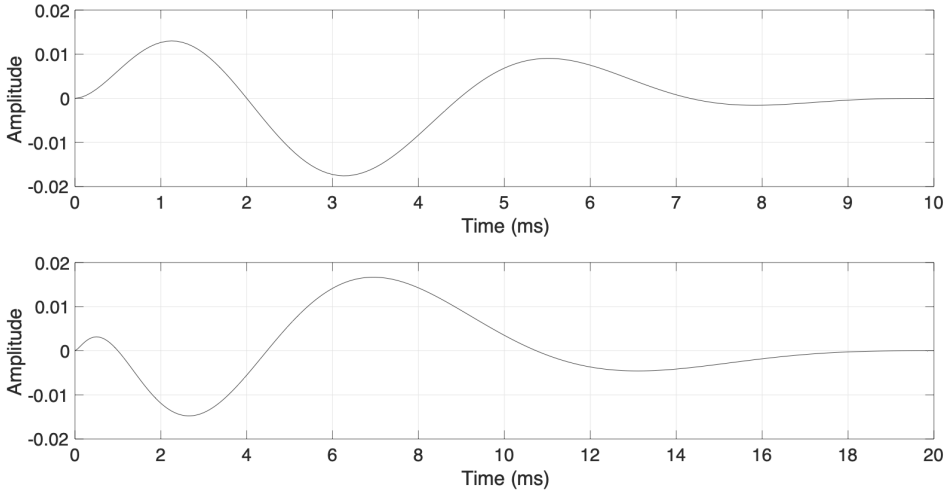


Fig. 4.1 Two simulated crackles, normalized in energy, are modelled: (t_{IDW}, t_{2CD}) = (2 ms, 10 ms) in the top plot and (t_{IDW}, t_{2CD}) = (1 ms, 20 ms) in the bottom plot.

  - Zero Crossing Location: The point where the first cycle of the crackle waveform $y(t)$ reaches zero amplitude is explicitly defined by the parameter t_{IDW}.
  - Power Concentration: Most of the power in the crackle waveform is concentrated near the beginning of the waveform. This means that the crackle sound starts with a burst of energy.

- Modulating Function: To shift the majority of the power to the beginning of the waveform (as assumed in point iii above), a modulating function denoted as $m(t)$ is generated. This function likely shapes the amplitude envelope of the crackle waveform to concentrate the power at the beginning. This is done to match the characteristic of real crackles, where the initial part of the sound is more intense.

A visual representation of these artificial crackles is shown in Figure 4.1 that likely shows an example of the simulated crackles in the time domain. This visual representation provides a clear illustration of what the synthetic crackle sounds look like.

These synthetic sounds are then used to train a classifier, allowing it to learn the distinguishing features of crackles, which can be applied to identify real crackles in medical or acoustic applications, for instance. The choice of parameters and the modulating function used in the generation of these synthetic sounds are essential in ensuring that the simulated data accurately represents real-world crackles.

$$t_0 = \frac{t_{IDW}}{t_{2CD}} \tag{4.1}$$

$$y_0(t) = sin\left(4\pi t^{\alpha}\right), \alpha = \frac{log_{10}\left(0.25\right)}{log_{10}\left(t_0\right)} \tag{4.2}$$

$$m(t) = \frac{1}{2}\left(1 + cos\left(2\pi\left(t^{\frac{1}{2}} - \frac{1}{2}\right)\right)\right) \tag{4.3}$$

$$y(t) = y_0(t)m(t) \tag{4.4}$$

The methodology used in this research study that describes the creation of a dataset of 187 crackle waveforms for analysis in detail:

- Crackle Waveform Creation: The primary objective of this work is to create a dataset of crackle waveforms for analysis. To achieve this, a set of K (K = 187) crackle waveforms is generated. The characteristics of each waveform are determined by two key parameters: t_{IDW} and t_{2CD}.
- Parameter Selection Strategy: The choice of parameter values for t_{IDW} and t_{2CD} is determined by a conservative strategy. In this context, "conservative" suggests a systematic and thorough approach. To cover a wide range of possibilities, all possible combinations between the parameters t_{IDW} and t_{2CD} are considered.
- Parameter Ranges and Steps:
  - t_{IDW}: The parameter t_{IDW} is allowed to vary within the range of 0.5 ms to 1.5 ms. This range is chosen to capture variations in the location where the first cycle of the waveform equals zero. The step size used for this parameter is 0.1 ms, indicating that the parameter value is incremented in 0.1 ms intervals. This fine granularity allows for a comprehensive exploration of possible values.
  - t_{2CD}: The parameter t_{2CD} varies within the range of 3.3 ms to 20 ms. This parameter represents the duration of two cycles of the crackle waveform. The step size used is 1 ms, indicating that the parameter value is adjusted in 1 ms increments. This also offers a systematic exploration of possibilities.
  - Normalization of Signals: Each generated signal, represented by $y(t)$, is subjected to energy normalization. The normalization ensures that the sum of the squares of the signal values is equal to 1.0. This normalization step is important for ensuring that the signals are on a consistent scale and to make them directly comparable for analysis.

The normalization step ensures consistency in signal energy, and the spectral patterns provide a visual representation of the dataset's characteristics, which is valuable for subsequent analysis and interpretation. This methodology is a fundamental step in understanding and characterizing crackle sounds for various applications, such as medical diagnosis or acoustic analysis. The magnitude Fourier transform (spectral pattern)

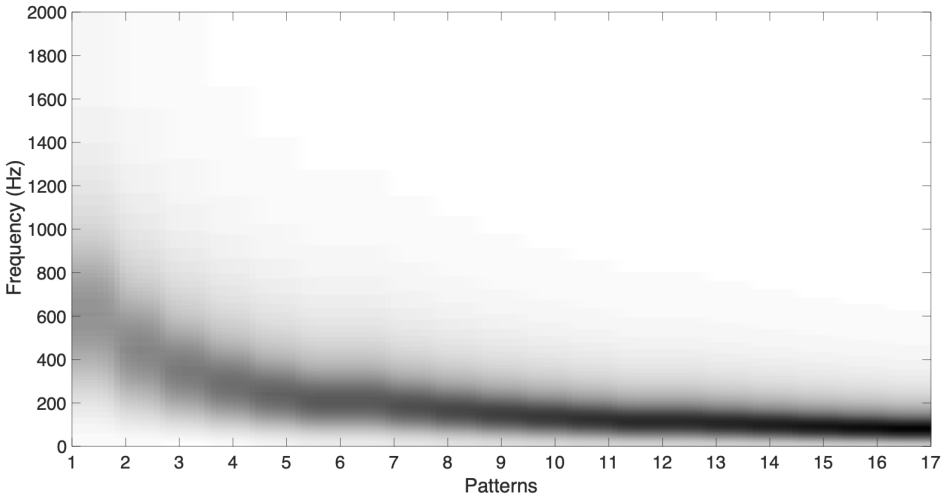


Fig. 4.2 Magnitude spectrogram of the first eighteen spectral patterns combining the parameters t_{IDW} and t_{2CD} as previously mentioned. Higher energy is indicated by darker colour.

of the set of K crackle waveforms is displayed in Figure 4.2. This visual representation likely provides insights into the frequency domain characteristics of the crackle waveforms, which can be informative for further analysis and classification.

4.3 Experimental results

This section of the document discusses a systematic comparison of the results obtained from different methods: IEM-FD, TVAR, and a newly proposed method. The purpose of this comparison is to evaluate the performance of these methods in a specific context. Here's a more detailed breakdown of this process:

- Methods Under Comparison: Three methods are under scrutiny in this evaluation: IEM-FD, TVAR, and the new method being proposed. These methods are used for some form of signal separation or analysis.
- Programming Language: All of the separation techniques, including the mentioned methods, have been implemented using the MATLAB (R2020b) programming language. MATLAB is a commonly used platform for signal processing and numerical analysis.
- Evaluation Metrics: The document refers to Chapter 3 for the description of the metrics used to evaluate the implemented methods. These metrics are crucial for quantifying and comparing the performance of the methods. They likely include measures that assess the accuracy, precision, and reliability of the results.

- Comparison Criteria: The comparison involves an evaluation of the agreement between the ”ground-truth” and the ”estimated” temporal location and length of each crackle. The ”ground-truth” represents the actual or reference values, while the ”estimated” values are what the methods produce. The evaluation is done within a tolerance window of 0.6 ms. This means that results are considered accurate if they fall within this 0.6 ms range of the true values.

- Data Visualization: The overall results of the implemented methods are visualized in Figure 4.3. This figure employs a ”boxplot” to represent the data distribution. A boxplot summarizes data using a five-number summary: minimum, first quartile (Q1), median, second quartile (Q2), and maximum. It provides valuable information about the spread of data, the presence of outliers or atypical values, and the overall distribution characteristics.

- Interpreting the Boxplot: The boxplot is a powerful tool for understanding data patterns. It helps in identifying whether the data is symmetric or skewed, how tightly the values are clustered, and whether there are any extreme or unusual data points. It is a concise way to visualize the central tendency and variability of the data, making it easier to make informed comparisons between the methods.

In summary, this section provides a rigorous evaluation of three different methods in a specific context, using well-defined metrics and visualization tools. It allows for a comprehensive assessment of their performance and the identification of strengths and weaknesses. This type of systematic comparison is common in scientific research and data analysis to ensure that the chosen method is suitable for the intended application.

4.3.1 Comparison of Different Methods

To ascertain the effectiveness of our proposed methodology, we conducted a thorough comparative analysis against well-established techniques in the field. Our evaluation included benchmarking the AR-based features against the Iterative Envelope Mean-Fractal Dimension (IEM-FD) [7] and the Time-Varying Autoregressive (TVAR) [8] methods. The results of this comparison unveiled significant improvements in performance metrics, thereby establishing the efficacy of our approach in characterizing and detecting events, particularly in the context of identifying crackles in respiratory signals.

Figure 4.3 provides a visual representation of the results in terms of accuracy (Acc), sensitivity (S_e), and precision (P_r) for three different methods: the Iterative Envelope Mean-Fractal Dimension (IEM-FD) method on the left side, the Time-Varying Autoregressive (TVAR) method in the center, and the newly proposed method on the right

side. This figure allows for a direct comparison of the performance of these methods in these key metrics.

In addition to the boxplot representation, Figure 4.4 presents a graphical display of accuracy results for all the compared methods as a function of Signal-to-Noise Ratio (SNR) values, which vary between -10 dB and 10 dB. This plot provides an insight into how each method's accuracy changes with different levels of noise in the signal.

In broad terms, the results indicate that the proposed method outperforms the IEM-FD and TVAR methods in several important aspects:

- Accuracy (Acc): The proposed method demonstrates significantly higher accuracy compared to the other methods. In practical terms, this means it is more successful in correctly classifying or identifying certain features in the data.

- Precision (P_r): The precision of the proposed method is notably superior to that of the IEM-FD and TVAR methods. This signifies that the proposed method produces fewer false positives, indicating its ability to make precise and reliable identifications.

- Sensitivity (S_e): Moreover, the proposed method also excels in terms of sensitivity, surpassing both the IEM-FD and TVAR methods. This indicates that the proposed method is more effective at identifying true positives, highlighting its capacity to capture important features or events.

In summary, the visual representations in these figures clearly illustrate that the newly proposed method offers substantial advantages in terms of accuracy, precision, and sensitivity when compared to the established IEM-FD and TVAR methods. These findings emphasize the efficacy and potential of the proposed approach in the context of the study's objectives and its applicability to real-world scenarios.

A detailed analysis is offers a nuanced understanding of the performance of the IEM-FD, TVAR, and proposed methods, revealing their strengths and weaknesses in terms of accuracy, sensitivity, and precision, and highlighting the robustness and reliability of the proposed approach, particularly in variable SNR conditions.

IEM-FD Method [9]:

- Median: The median accuracy for the IEM-FD method is 45.45%. This value represents the middle point of the dataset when arranged in ascending order.

- First Quartile (Q1): The first quartile, Q1, is at 32.25%. It signifies that 25% of the data points have an accuracy score lower than this value.

- Third Quartile (Q3): The third quartile, is 50%. Half of the data points have an accuracy score below this value. It signifies that 50% of the data points have an accuracy score higher than this value.

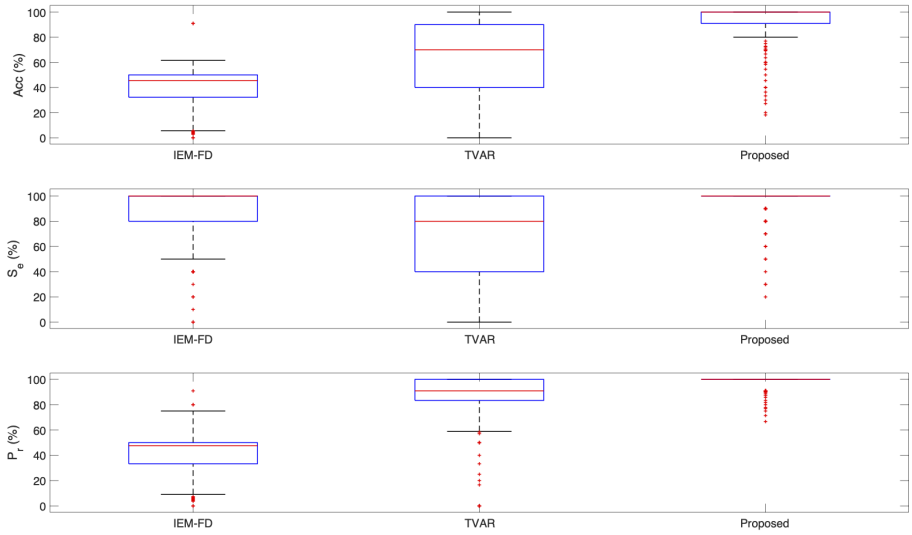


Fig. 4.3 Accuracy, sensitivity and precision average results evaluating all scenarios and SNRs in the dataset ψ by IEM-FD [7], TVAR [8], and the proposed method.

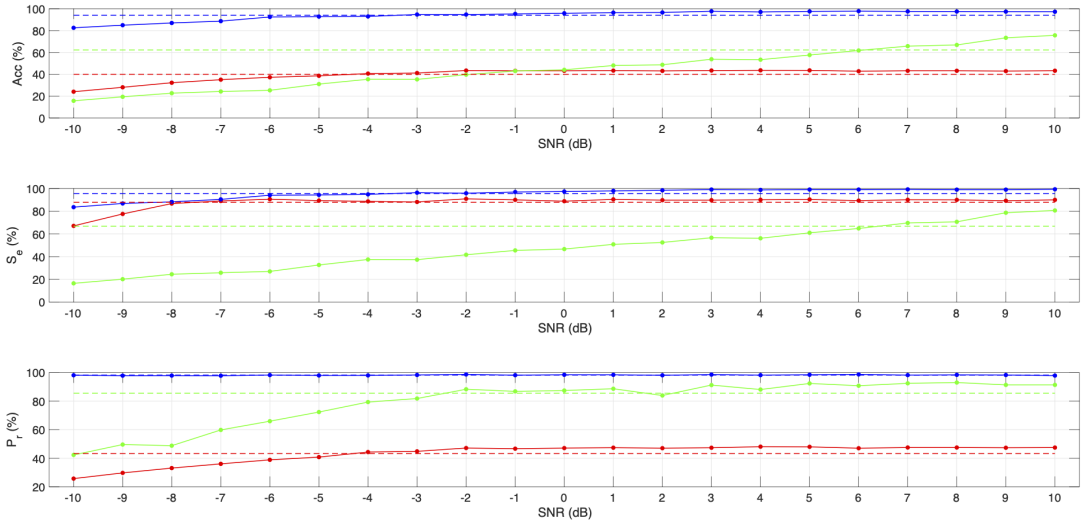


Fig. 4.4 Accuracy, sensitivity, and precision average results evaluating all scenarios for each SNR in the database ψ by IEM-FD [9] (red color), TVAR [8] (green color), and the proposed method (blue color), where the dashed lines represent the mean value for each metric and method.

- Maximum: The maximum accuracy achieved by the IEM-FD method is 61.54%, indicating the best performance observed in this study.
- Minimum: The minimum accuracy recorded is 5.66%, highlighting the lowest level of accuracy attained by the IEM-FD method.

These statistics reveal the IEM-FD method's performance characteristics. The method's overall accuracy is centered around the median value, and there is a wide range of performance results, from a minimum of 5.66% to a maximum of 61.54%. It's noted that the method struggles with false positives, as indicated by its precision values in Figure 4.3.

TVAR Method [8]:

- Median: The TVAR method's median accuracy is 70%, representing the middle point of the dataset.
- First Quartile (Q1): The first quartile, Q1, is at 40%. It signifies that 40% of the data points have an accuracy score lower than this value.
- Third Quartile (Q3): The third quartile, Q3, is at 90%. It signifies that 90% of the data points have an accuracy score higher than this value.
- Maximum: The TVAR method reaches a maximum accuracy of 100%, representing the best results achieved.
- Minimum: The minimum accuracy obtained by the TVAR method is 0%, illustrating the poorest performance in the dataset.

The TVAR method displays notable variability in its results. While it achieves high accuracy values (e.g., median of 70%), the range between a minimum of 0% and a maximum of 100% implies significant performance fluctuations. This suggests that the TVAR method's effectiveness strongly depends on the input signal conditions, especially noticeable in Figure 4.4, where it exhibits the worst performance under low Signal-to-Noise Ratio (SNR) conditions.

Proposed Method:

- Median: The proposed method boasts a median accuracy of 100%, signifying that the middle point of its dataset is at the maximum achievable accuracy.
- First Quartile (Q1): The first quartile, Q1, is at 90.9%, indicating that a substantial portion of data points achieve high accuracy values.
- Third Quartile (Q3): The third quartile, Q3, remains at 100%, revealing that half of the data points attain the maximum accuracy.

- Maximum: The proposed method attains a maximum accuracy of 100%, showing consistent high performance across various scenarios.
- Minimum: The minimum accuracy for the proposed method is 80%, demonstrating that even the lowest performance still achieves a relatively high accuracy score.

The proposed method excels in terms of consistency and robustness, with a narrow range of accuracy scores. This method demonstrates a strong performance regardless of the input signal's SNR conditions, as highlighted in Figure 4.4.

Now, let's take a closer look at a specific subtype of crackles. Figure 4.5 provides a comprehensive breakdown of the results, presenting the metrics of accuracy (Acc), sensitivity (S_e), and precision (P_e) in relation to the type of crackle, distinguishing between coarse (on the right side of the figure) and fine (on the left side of the figure), and their respective Signal-to-Noise Ratio (SNR) values. The observations made from this analysis shed light on the performance characteristics of the proposed method as compared to other methods.

First and foremost, it's evident that the proposed method showcases superior performance in both cases—coarse and fine crackle detection. It consistently outperforms the other methods in these specific contexts.

Upon closer inspection, comparing the results between the two sides of the figure, we notice that the proposed method does exhibit a relative underperformance when dealing with coarse crackles under low SNR scenarios. This suggests that its performance might be sensitive to the presence of background noise in the signal. A similar behavior is observed with the TVAR method. As mentioned earlier, the results from the TVAR method tend to be limited by the number of false positives, resulting in a noticeable underperformance in terms of sensitivity (S_e).

Interestingly, the results appear to be more stable in the case of the IEM-FD method when comparing its performance in fine versus coarse crackle detection. However, it is essential to note that the results of the IEM-FD method, while relatively stable, still fall short of the performance achieved by the compared approaches, particularly due to the occurrence of false negatives (i.e., crackle events being detected as normal). This leads to a clear underperformance in terms of precision (P_r).

In essence, accuracy (Acc) is a composite metric that takes into account both sensitivity (S_e) and precision (P_r) values. It serves as a general measure of overall performance, encapsulating the trade-off between correctly identifying true events and minimizing both false positives and false negatives.

These observations provide a comprehensive perspective on the nuanced performance of the proposed method, its strengths, and the specific scenarios where it excels, all within the context of fine and coarse crackle detection under various SNR conditions.

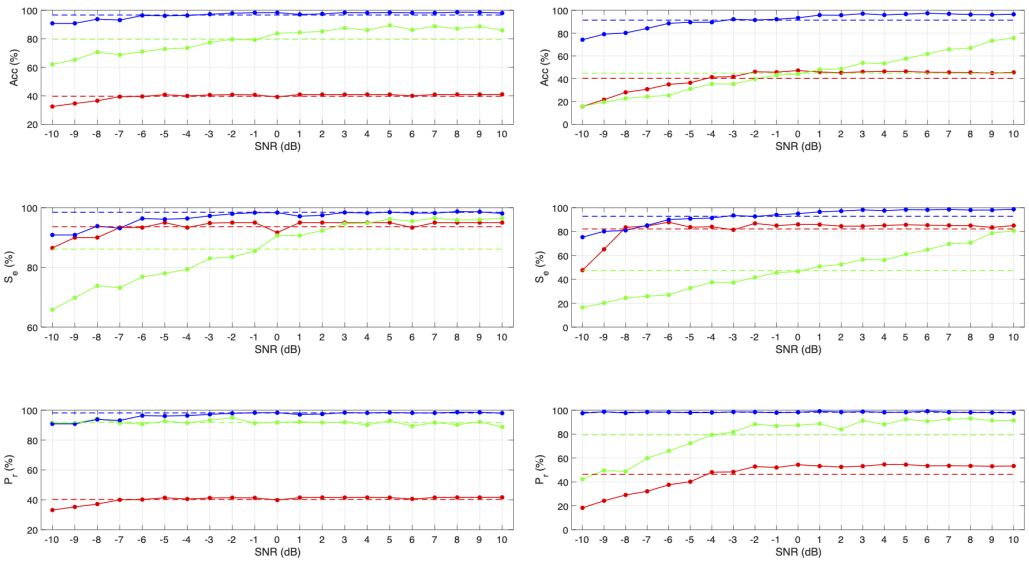


Fig. 4.5 Accuracy, sensitivity, and precision average results evaluating all scenarios for each type (fine crackles on the left side and coarse crackles on the right side) of crackles and SNRs from database ψ by IEM-FD [9] (red color), TVAR [8] (green color), and the proposed method (blue color), where the dashed lines represent the mean value for each metric and method.

Table 4.1 serves as an essential reference point, providing a comprehensive overview of the average performance metrics, including accuracy (Acc), sensitivity (S_e), and precision (P_r), for three distinct methods: the Iterative Envelope Mean-Fractal Dimension (IEM-FD) method, the Time-Varying Autoregressive (TVAR) method, and the innovative approach proposed in this study. These metrics are derived from testing samples within an artificial database that encompasses a wide range of Signal-to-Noise Ratios (SNR) spanning from -10 dB to +10 dB. Notably, this table extends its analysis further by categorizing the test samples according to the type of crackles they contain. The categorization includes five distinct types of crackles: ATS, Hoevers, Cohen, IPF, and BE, denoted here as simulated scenarios 1 to 6, as well as real scenarios 1 and 2.

Here, in this detailed breakdown of results, the superiority of the proposed method becomes evident. It consistently demonstrates significantly higher percentage values in comparison to both the IEM-FD and TVAR methods across the board. These findings emphasize the remarkable efficacy and potential of the proposed approach in accurately identifying and characterizing various types of crackles, whether real or simulated, across a wide range of SNR conditions.

However, it is essential to note an intriguing observation. All three methods exhibit a notable underperformance when faced with real scenario 2, where the audio contains coarse crackles, often associated with conditions such as bronchitis, superimposed on simulated breathing sounds. This suggests that the key differentiation between simu-

Scenario	Type	Kc	NOTS	Noise	Diagnosis	SNR	Accuracy [Acc]			Sensitivity [S_e]			Precision [P_r]		
							IEM-FD	TVAR	Proposed	IEM-FD	TVAR	Proposed	IEM-FD	TVAR	Proposed
Simulated 1	FCS [326]	10	15	Nr	-	[−10 dB, 10 dB]	46.94%	69.23%	98.36%	97.46%	73.02%	99.94%	46.94%	93.33%	98.42%
Simulated 2	FCS [127]	10	15	Nr	-	[−10 dB, 10 dB]	33.21%	81.36%	95.85%	100%	92.60%	97.81%	33.21%	87.27%	97.93%
Simulated 3	FCS [325]	10	15	Nr	-	[−10 dB, 10 dB]	47.72%	76.21%	98.03%	97.14%	81.56%	99.94%	47.72%	91.98%	98.08%
Simulated 4	CCS [326]	10	15	Nr	-	[−10 dB, 10 dB]	41.86%	46.00%	96.57%	93.17%	48.54%	98.25%	42.15%	82.64%	98.27%
Simulated 5	CCS [127]	10	15	Nr	-	[−10 dB, 10 dB]	47.63%	71.86%	98.63%	97.46%	76.73%	99.90%	47.63%	91.68%	98.71%
Simulated 6	CCS [325]	10	15	Nr	-	[−10 dB, 10 dB]	41.12%	38.68%	95.47%	91.49%	40.76%	96.86%	41.46%	76.70%	98.54%
Real 1	FCS [9, 7]	10	15	Nr	IPF	[−10 dB, 10 dB]	30.70%	92.04%	94.59%	79.96%	97.30%	96.10%	33.26%	94.39%	98.32%
Real 2	CCS [9, 7]	10	15	Nr	BE	[−10 dB, 10 dB]	30.64%	22.70%	74.78%	46.54%	23.87%	76.03%	53.86%	66.33%	97.72%

Table 4.1 Detailed results in terms of accuracy, sensitivity, and precision (mean values per crackle type) K_C: number of crackles per signal. $NOTS$: number of signals per SNR. N_S: number of signals generated taking into account all SNRs evaluated.

lated and real fine crackles allows for a modeling approach that is valid for both scenarios. In contrast, the parameters necessary for effectively detecting and characterizing coarse crackles might vary between simulated and real cases, leading to a less accurate performance under these specific circumstances.

In summary, Table 4.1 provides a detailed and comprehensive evaluation of the proposed method and its peers, showcasing the former's exceptional performance across diverse scenarios, while also shedding light on the challenges associated with real-world conditions that involve coarse crackles. These insights are invaluable in advancing our understanding of crackle detection and refining the methods applied in this critical field.

Having completed a thorough evaluation of the proposed method's robustness across a diverse array of signal types and Signal-to-Noise Ratio (SNR) conditions, as well as a comprehensive comparison with traditional, classical approaches, our research endeavors extend further. We seek to embark on a new phase of evaluation, one that involves benchmarking the results against cutting-edge techniques rooted in deep learning.

Deep learning methodologies have recently emerged as a focal point of interest and investigation within the field of adventitious sound detection. This modern approach harnesses the power of artificial neural networks, offering a unique and data-driven perspective on the challenges associated with detecting and characterizing these crucial acoustic events. These methods have gained significant traction due to their potential to revolutionize the way we approach the analysis of adventitious sounds, presenting an opportunity to overcome traditional limitations and further enhance the accuracy and efficiency of detection systems.

Numerous studies and researchers have delved into the application of deep learning techniques for adventitious sound detection. These efforts have spanned a spectrum of applications, from lung sound analysis to noise reduction and the classification of specific adventitious sound types. The literature showcases a wealth of pioneering work, including but not limited to the contributions of Bardou et al. [308], Aykanat et al. [307], Kochetov et al. [212], Liu et al. [309], Minami et al. [215], Ma et al. [214], Ngo et al. [310], and Demir et al. [216].

The incorporation of deep learning into the realm of adventitious sound detection

marks a significant shift in the landscape of research and development. These innovative techniques have the potential to bring about substantial improvements in accuracy, efficiency, and the adaptability of detection systems, ultimately contributing to more effective diagnostic tools and enhanced patient care. Therefore, our intention is to engage in a rigorous comparison with these state-of-the-art deep learning approaches, aiming to gain a comprehensive understanding of their relative strengths and limitations, and ultimately, to facilitate the continued advancement of this vital field of study.

The study conducted by Rocha et al. in 2020 [10] underscored the significant utility of Convolutional Neural Networks (CNNs) as cutting-edge solutions that have garnered wide-ranging acclaim across diverse research domains. Building upon the insights and architecture detailed in Rocha et al.'s work, we extended our investigation by subjecting the same array of scenarios, previously described, to a rigorous testing regimen. In this phase, we embarked on a meticulous comparative analysis, scrutinizing the outcomes produced by two distinct classifiers: the Support Vector Machine (SVM) classifier and the CNN classifier.

The comparative results, meticulously presented in Table 4.2, unveil a compelling performance contrast in terms of three pivotal metrics: accuracy (Acc), sensitivity (S_e), and precision (P_r). These metrics encapsulate the core measures of success in our assessment across the eight scenarios integral to our proposed method.

Upon a thorough examination of the data in Table 4.2, it becomes evident that both classifiers, namely SVM and CNN, exhibit remarkable performance, consistently achieving accuracy scores exceeding 90% in all scenarios, with the exception of the real coarse crackle scenario. This notable observation underscores the robustness and reliability of these classifiers in accurately identifying and characterizing various adventitious sound events, which is particularly promising from a clinical perspective.

However, a discernible nuance emerges from the comparative analysis. While both classifiers deliver commendable results, a subtle performance differentiation becomes apparent. The SVM-based classifier exhibits a slight advantage in terms of accuracy (Acc) and sensitivity (S_e) in some scenarios. Conversely, the CNN classifier demonstrates a marginal edge in terms of precision (P_r) in certain instances. This nuanced contrast indicates that the proposed system, harnessing both classifiers, possesses a certain degree of robustness when it comes to discerning the presence of crackle events. This characteristic holds considerable significance in the context of clinical applications, where the accurate identification of such events is paramount.

It is crucial to note that the SVM-based classifier carries certain distinct advantages. It is notably simpler to train, boasts fewer parameters, and is, therefore, more resistant to overfitting, a common concern in machine learning. Furthermore, its parameters are more interpretable and transparent, rendering it an appealing choice for contexts where interpretability and simplicity are prized.

The incorporation of CNN and SVM classifiers into our methodology brings about a multifaceted set of results, shedding light on their respective strengths and advantages. This nuanced comparison serves as a pivotal stepping stone in our quest to enhance the accuracy and efficacy of adventitious sound detection systems, with the ultimate aim of benefiting clinical practice and patient care.

It is essential to underscore the noteworthy performance decline that becomes evident when transitioning from simulated scenarios to real-world situations, as meticulously examined in this study. A closer examination of Table 4.2 unmistakably reveals discernible decreases in all three critical performance metrics—accuracy (Acc), sensitivity (S_e), and precision (P_r) across these real-world scenarios. These findings provide invaluable insights into the potential limitations inherent in the proposed methodology.

In the context of real patient scenarios, it is imperative to acknowledge the inherent variability and unpredictability that characterize the properties of input signals. This variability stems from numerous factors, including individual patient characteristics, ambient conditions, and the dynamic nature of physiological processes. These factors, in turn, can result in consequential reductions in the performance outcomes of detection and characterization systems, as is observed in this study.

The research endeavors on the horizon are strategically poised to address these challenges. The forthcoming investigations will focus on data characteristic extraction and the modeling of time-frequency behaviors inherent to real-world adventitious sound signals. These efforts are envisioned to yield substantial progress in enhancing the generalization and adaptability of the proposed methodology. By capturing and accounting for the authentic behaviors and intricacies exhibited by such sound signals in the dynamic and often unpredictable real-world contexts, we aim to fortify the reliability and robustness of our detection system.

In summary, the observed performance decline in real-world scenarios serves as a poignant reminder of the complexity and variability inherent in clinical practice. However, this realization fuels our determination to refine our methodologies, improve adaptability, and ultimately contribute to more accurate and dependable adventitious sound detection systems in the challenging realm of healthcare and medical diagnostics.

Scenario	Type	Kc	NOTS	NOISE	Diagnosis	SNR	Accuracy (%) [Acc]		Sensitivity (%) [S_e]		Precision (%) [P_r]	
							SVM	CNN	SVM	CNN	SVM	CNN
Simulated 1	FCS [326]	10	15	Nr	-	[−10 dB, 10 dB]	98.36	97.61	99.94	97.94	98.42	99.49
Simulated 2	FCS [127]	10	15	Nr	-	[−10 dB, 10 dB]	95.85	97.93	97.81	98.28	97.93	99.53
Simulated 3	FCS [325]	10	15	Nr	-	[−10 dB, 10 dB]	98.03	97.44	99.94	97.72	98.08	99.50
Simulated 4	CCS [326]	10	15	Nr	-	[−10 dB, 10 dB]	96.57	92.98	98.25	92.73	98.27	99.60
Simulated 5	CCS [127]	10	15	Nr	-	[−10 dB, 10 dB]	98.63	98.25	99.90	98.54	98.71	99.55
Simulated 6	CCS [325]	10	15	Nr	-	[−10 dB, 10 dB]	95.47	94.34	96.86	94.21	98.54	99.58
Real 1	FCS [9, 7]	10	15	Nr	IPF	[−10 dB, 10 dB]	94.58	91.42	96.10	93.11	98.32	99.34
Real 2	CCS [9, 7]	10	15	Nr	BE	[−10 dB, 10 dB]	74.78	72.34	76.03	74.00	97.72	98.12

Table 4.2 Comparison of the results of the proposed method as input of an SVM and CNN in terms of accuracy, sensitivity, and precision (mean values per crackle type) K_C: number of crackles per signal. $NOTS$: number of signals per SNR. N_S: number of signals generated considering all SNRs evaluated.

CHAPTER 5

Cochleogram-based adventitious sounds classification using convolutional neural networks

5.1 Abstract

The World Health Organization (WHO) establishes as a top priority the early detection of respiratory diseases. This detection could be performed by means of recognizing the presence of acoustic bio- markers (adventitious sounds) from auscultation because it is still the main technique applied in any health center to assess the status of the respiratory system due to its non-invasive, low-cost, easy to apply, fast to diagnose and safe nature. Despite the novel deep learning approaches applied in this biomedical field, there is a notable lack of research that rigorously focuses on different time–frequency representations to determine the most suitable transformation to feed data into Convolutional Neural Network (CNN) architectures. In this paper, we propose the use of the cochleogram, based on modeling the frequency selectivity of the human cochlea, as an improved time–frequency representation to optimize the learning process of a CNN model in the classification of respiratory adventitious sounds. Our proposal is evaluated using the largest and most challenging public database of respiratory sounds. The cochleogram obtains the best binary classification results among the compared methods with an average accuracy of 85.1% in wheezes and 73.8% in crackles, and a competitive performance evaluating a multiclass classification scenario in comparison with other well-known state-of-the-art deep learning models. The cochleogram provides a suitable time–frequency representation since it is able to model respiratory adventitious content more accurately by means of non-uniform spectral resolution and due to its increased robustness to noise and acoustic changes. This fact implies a significant improvement in the learning process of CNN models applied in the classification of respiratory adventitious sounds.

5.2 Contribution

While recent research endeavors have harnessed advanced machine learning techniques for adventitious sound classification, a conspicuous gap in the literature remains con-

cerning a comprehensive evaluation of the various time-frequency (TF) representations that underpin these systems. It is well recognized that the choice of TF representation significantly influences the ability to replicate the auditory perception of the human ear, a phenomenon that has been articulated in previous works [327, 328]. In this study, we delve into an exploration of the impact of classical TF representations, such as the Short-Time Fourier Transform (STFT) and the Mel-scaled spectrogram. In addition, we introduce a novel perspective by proposing the utilization of a human auditory-inspired non-linear representation known as the cochleogram. It's worth noting that the cochleogram has found application in diverse scientific domains, including audio analysis [1, 329], and heart sound detection [220]. However, to the best of our knowledge, its potential in classifying adventitious respiratory sounds remains largely unexplored.

The significance of this non-linear representation lies in its unique attribute—non-uniformity. The cochleogram's non-uniform nature has demonstrated higher robustness in the face of noise and acoustic variations when compared to classical linear or speech-based TF representations [220]. This characteristic makes it a promising candidate for enhancing the reliability and adaptability of adventitious sound classification systems.

To empirically substantiate the advantages of adopting the cochleogram derived from the human auditory system as opposed to conventional TF representations like the STFT and Mel-scaled spectrogram, we introduce a Convolutional Neural Network (CNN) architecture closely resembling the one presented in [10]. Our objective is to rigorously assess the impact of different input TF representations on the task of classifying adventitious respiratory sounds. To carry out this evaluation, we leverage the largest and most challenging public database of breath sounds, the ICBHI database [2, 173]. This database serves as a fertile ground for our investigations.

In particular, our study involves an examination of the baseline CNN model's performance in detecting the presence of crackles and wheezing, both in a binary and a multiclass classification scenario. The outcomes of these evaluations are pivotal in unraveling the potential advantages and limitations of employing the cochleogram as the input TF representation for respiratory sound classification.

Moreover, we extend our exploration to other state-of-the-art CNN models, including AlexNet [241], ResNet50 [5], and VGG16 [242]. This broader analysis allows us to gain insights into how different TF representations influence the performance of diverse CNN architectures and to assess the generalizability of our findings across various model paradigms.

In summary, this study is positioned at the intersection of auditory science, machine learning, and respiratory healthcare. It strives to elucidate the potential of the cochleogram as a valuable tool in the classification of adventitious respiratory sounds, with the overarching aim of advancing the accuracy and reliability of diagnostic systems in this critical domain.

5.3 Experimental results

The initial phase of our experimentation involved a comprehensive preliminary analysis to determine the most suitable window lengths for each time-frequency (TF) representation. We systematically evaluated a range of window lengths to discern their impact on the classification task. Specifically, we assessed the following window lengths: $N = [8, 16, 32, 64, 128, 256]$ milliseconds. To maintain consistency in our analysis, we employed a Blackman-Harris window and a time shift between windows set at a 75% overlap size. This windowing configuration aligns with the optimal setup identified in a prior study [10] and serves as a standard for our experimentation.

In terms of the training and testing conditions, we adhered to a robust and widely-recognized approach. We implemented a 10-fold cross-validation procedure, which was repeated five times to ensure robustness and mitigate variability [330]. Each iteration of this process involved dividing the dataset into training and testing subsets, following a 75%-25% distribution, where 75% of the data constituted the training subset, and 25% served as the testing subset. Within the training set, a further 25 percent was designated for validation purposes.

It's worth noting that the distribution of crackles and wheezes within the dataset is unbalanced, a fact detailed in Table 3.2 in Chapter 3. To address this imbalance, we took great care to ensure that each fold of the cross-validation process contained a proportional representation of both crackles and wheezes events. This approach guarantees that the classification model's performance is rigorously evaluated across a balanced distribution of different sound events in each fold.

The practical implementation of our experimental work was carried out using TensorFlow and Keras, well-established machine learning libraries, on a computer equipped with an Intel(R) Core(TM) i7-5500 CPU operating at 2.4 GHz with four cores, an NVIDIA GeForce GTX1080Ti GPU, and 64 GB of RAM. This computational setup provided the necessary resources to execute our experiments efficiently.

To promote transparency and reproducibility in our research, we have made the code publicly accessible through a dedicated repository at the following URL: `https://github.com/loredanadariamang/CODE_EAMBES2022.git`. This repository serves as a valuable resource for fellow researchers and practitioners interested in delving deeper into our work, replicating our experiments, and building upon our findings in the realm of adventitious sound classification.

5.3.1 Optimal parameters estimation

As elucidated earlier, our investigation delved into the impact of varying window lengths on the classification performance of different time-frequency (TF) representations. To ascertain the optimal parameters for our model, we conducted a rigorous evaluation

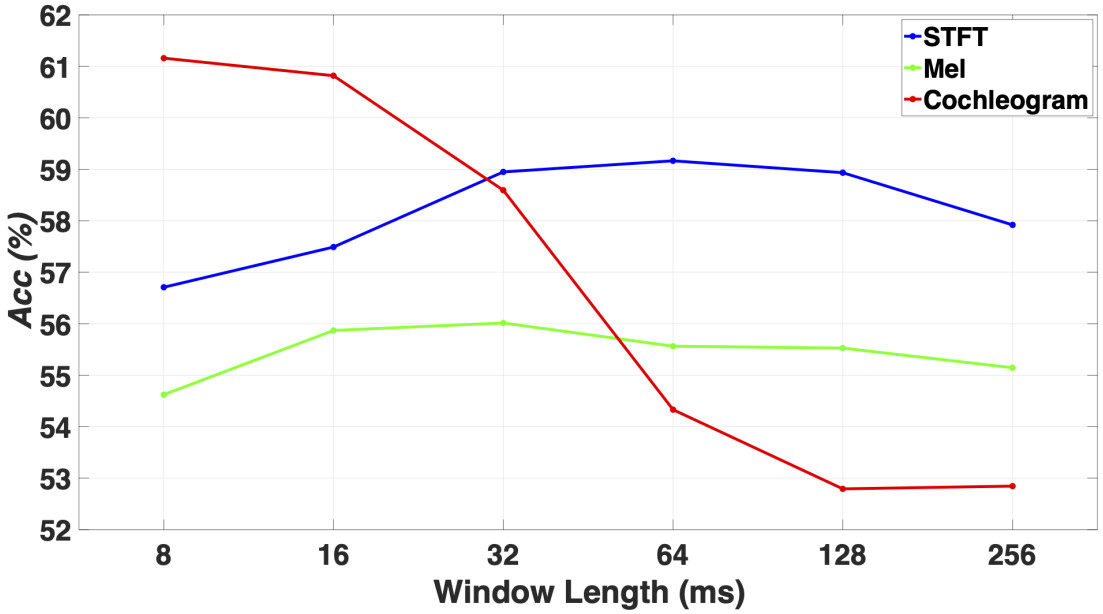


Fig. 5.1 Overall accuracy results evaluating 4 classes, in terms of mean values for the whole range of window lengths, using the ICBHI dataset.

considering four distinct sound classes: normal, crackles, wheezes, and a combination of both crackles and wheezes. The objective was to identify the window lengths that yielded the best classification performance for each TF representation.

The results of this analysis are visually depicted in Figure 5.1, which showcases the averaged accuracy values attained through 10-fold cross-validation for our baseline Convolutional Neural Network (CNN) model when different TF representations are employed as inputs.

Upon close examination of Figure 5.1, several noteworthy observations come to the fore. In the case of the Short-Time Fourier Transform (STFT)-based model, the highest performance is achieved within the window length range of 32 to 128 milliseconds. Remarkably, the optimal performance is achieved with a 64 millisecond window length. This finding aligns with the outcomes presented in a prior study [10] and underscores the consistency of these optimal window length values in different research contexts.

Turning our attention to the Mel-scaled model, it is evident that the peak performance is associated with a 32 millisecond window length. This observation mirrors the findings in the STFT-based model and, notably, aligns with the results presented in the previous work by Rocha et al. [10].

In essence, our systematic investigation has not only reaffirmed the effectiveness of certain window length values but has also provided valuable insights into the consistency of these optimal parameters across different research settings. These findings serve as crucial guidance for selecting the appropriate window length for a given TF representation when developing and deploying CNN-based models for adventitious sound classification.

In contrast to the optimal window lengths observed for traditional time-frequency (TF) representations, such as the Short-Time Fourier Transform (STFT) and Mel-scaled spectrogram, the human auditory system-based cochleogram exhibited a distinctive pattern of performance. This finding corroborates previous studies, specifically those conducted in [235], which have emphasized the importance of smaller window sizes for this particular TF representation.

As illustrated in Figure 5.1, the cochleogram reached its peak performance with an impressively small 8 millisecond window length. This optimal window length choice is pivotal as it aligns with the unique characteristics of adventitious sounds, particularly crackles. Crackles typically exhibit an explosive and transient nature with very short durations. The choice of an 8 millisecond window length is well-suited to model these characteristics. Larger window sizes were found to yield worse performance. This can be attributed to the loss of high temporal resolution with larger windows, making it challenging to accurately capture the short-lived and rapidly changing nature of crackle sounds. Essentially, larger windows dilute the temporal information, leading to increased confusion in the system's classification.

It's important to note that the cochleogram transformation operates using a logarithmic Equivalent Rectangular Bandwidth (ERB) scale, which inherently provides some temporal and spectral non-uniformity. However, in order to obtain a discrete TF representation, windowing is required for each resulting channel of the filtering process. This windowing process enables the extraction of meaningful information from the cochleogram.

An initial exploration was conducted based on the Multi-Resolution Cochleogram (MRCG) approach [1]. This approach combines a high temporal resolution cochleogram with several low temporal resolution cochleograms in an attempt to model both local and temporal context information. However, the results from this investigation indicated that using only the high temporal resolution cochleogram provided superior classification performance for respiratory adventitious sounds compared to the combination of cochleograms with varying temporal resolutions. This observation underlines the critical importance of maintaining a high temporal resolution to effectively model crackle sounds. As a result, we focused exclusively on the TF representation derived from a single high temporal resolution cochleogram throughout our study.

The choice of a smaller window length for the cochleogram represents a deliberate and highly effective strategy, capitalizing on its capacity to capture the unique characteristics of adventitious sounds, particularly crackles. This emphasis on high temporal resolution is pivotal for improving the modeling and accurate classification of such sounds in the context of respiratory healthcare and diagnostics.

Following the meticulous parameter optimization process, our research endeavors extended to encompass additional experiments aimed at thoroughly assessing the clas-

sification performance of the proposed time-frequency (TF) representation. With the optimal parameter settings in place, we embarked on this phase of the study to comprehensively evaluate the capabilities and effectiveness of the TF representation in the context of our adventitious sound classification system.

This subsequent set of experiments represents a crucial step in our research, as it serves to validate the viability and robustness of the proposed TF representation. By subjecting it to rigorous testing and classification tasks, we aim to ascertain its ability to accurately discern and classify different types of respiratory sounds, including crackles and wheezes, under various conditions and scenarios.

The outcome of these experiments is expected to provide valuable insights into the suitability of the proposed TF representation for practical applications in the field of respiratory healthcare. It will help establish the foundation for the development of reliable diagnostic tools that can aid healthcare professionals in the early detection and assessment of respiratory conditions, ultimately contributing to improved patient care and outcomes.

These post-optimization experiments represent a pivotal phase in our research, focusing on the comprehensive evaluation of the proposed TF representation and its potential impact on the field of adventitious sound classification and respiratory healthcare

5.3.2 Binary classification results

In this phase of our research, we turn our attention to the evaluation of the classification performance of the cochleogram in a binary classification context. Our primary objective is to develop a system capable of detecting the presence of two crucial types of adventitious respiratory sounds, namely crackles and wheezes. The input data for this task consists of monaural sound signals that correspond to individual respiratory cycles.

The classification performance of our baseline model, equipped with the ability to utilize three distinct time-frequency (TF) representations, is showcased in Figure 5.2. This visualization offers valuable insights into the effectiveness of these TF representations in accurately identifying the presence of crackles and wheezes within respiratory sound data.

The results of these evaluations reveal several noteworthy findings. Firstly, the cochleogram representation emerges as the standout performer, yielding the highest classification accuracy for both wheezes (with an average accuracy of 85.1%) and crackles (with an average accuracy of 73.8%). This underscores the efficacy of the cochleogram in capturing and distinguishing the key features of these adventitious sounds, making it a robust choice for the classification task.

Moreover, the Short-Time Fourier Transform (STFT) representation also demonstrates competitive results, achieving an average accuracy of 84.9% for wheezes and

Comparison	Mann-Whitney U Test (p-value)	Wilcoxon signed-rank test (p-value)	Significantly better
Crackles			
Cochleogram vs. STFT	$2.21e-10$	$3.78e-09$	yes
Cochleogram vs. Mel	$1.38e-17$	$7.55e-10$	yes
Wheezes			
Cochleogram vs. STFT	$1.91e-06$	$8.01e-06$	yes
Cochleogram vs. Mel	$1.49e-17$	$7.55e-10$	yes

Table 5.1 Man-Whitney U Test and Wilcoxon signed-rank test Results for the sets of results obtained in the Figure 5.2, using a significance level α = 0.05.

71.5% for crackles. The strong performance of the STFT representation highlights its suitability for capturing the necessary frequency and temporal characteristics required for accurate detection of these respiratory sound events. This outcome emphasizes the importance of employing an appropriate window length and hop size in conjunction with low-pass filtering to achieve this level of resolution.

Conversely, the Mel-scaled spectrogram representation lags behind in terms of accuracy, delivering the lowest results in our evaluation. For wheezes, it achieves an accuracy of 81.5%, while for crackles, it achieves an accuracy of 68.7%. These findings suggest that, although Mel-scaled spectrograms are well-suited for modeling music and speech signals, they may not adequately highlight the frequencies of interest when it comes to adventitious sounds. The inferior performance of the Mel-scaled representation underscores the critical role of selecting the right TF representation tailored to the specific characteristics of the sound events under consideration.

In conclusion, our evaluation not only underscores the superior performance of the cochleogram representation for the classification of wheezes and crackles but also highlights the competitive capabilities of the STFT representation. These insights are instrumental in the development of robust diagnostic tools for respiratory healthcare, with the potential to enhance the early detection and assessment of respiratory conditions, ultimately improving patient care and outcomes.

In our pursuit of a thorough and robust evaluation of the classification performance of different time-frequency (TF) representations, we also aimed to determine whether the observed differences in performance were statistically significant. To achieve this, we employed two widely recognized and reliable non-parametric tests: the Mann-Whitney U Test and the Wilcoxon signed-rank test, both of which are designed for comparing two sets of distributions [289, 290].

These tests are used to assess the statistical significance of the differences between the TF representations, specifically the cochleogram, Short-Time Fourier Transform (STFT) spectrogram, and Mel-scaled spectrogram. To carry out this analysis, we formulated both a null hypothesis (H_o) and an alternative hypothesis (H_1). The null hypothesis represents the status quo, asserting that the two sets of results being compared are statistically equal. It is assumed to be true unless there is substantial evidence to

Fig. 5.2 Overall accuracy results evaluating the ICBHI dataset for the Cochleogram (window length of 8ms), STFT (window length of 32 ms) and Mel-scaled (window length of 32 ms) spectrograms using both the optimal values for the window lengths and overlap sizes. Each box represents 50 data points, each of them associated to a 10-fold cross-validation of the database evaluated. The lower and upper lines of each box show the first and third quartile. The line in the middle of each box represents the median value. The diamond shape in the center of each box represents the average value. The lines extending above and below each box show the extent of the rest of the samples, excluding outliers. Finally, outliers are defined as points that are over 1.5 times the interquartile range from the sample median, which are depicted as crosses.

support the contrary.

The pivotal metric in these tests is the p-value, which plays a decisive role in determining whether we should accept or reject the null hypothesis. A p-value below a predefined significance level, denoted as α (commonly set at 0.05), suggests that there is significant evidence to reject the null hypothesis and support the alternative hypothesis. Conversely, a p-value above α indicates that there is insufficient evidence to reject the null hypothesis, thus maintaining the status quo.

The results of the Mann-Whitney U Test and the Wilcoxon signed-rank test, presented in Table 5.1, provide a critical piece of evidence. Notably, in all cases, the calculated p-values do not exceed the chosen significance level of $\alpha = 0.05$. This outcome empowers us to confidently reject the null hypothesis and conclude that the classification performance offered by the cochleogram is indeed significantly superior when compared to both the STFT and Mel-scaled spectrograms for the assessment of both types of adventitious sounds, namely, crackles and wheezes.

These statistical tests validate the substantial and statistically significant advantages provided by the cochleogram representation in the context of our classification task, reinforcing its position as the most effective TF representation for accurate detection of these critical adventitious respiratory sounds. These findings are crucial for advancing the development of robust diagnostic tools for respiratory healthcare and enhancing the quality of patient care.

5.3.3 Four-class Normal/Crackles/Wheezing/Both classification results

In this section, we delve into a comprehensive evaluation of the performance of the proposed cochleogram representation in the context of a multiclass classification scenario [10]. The primary objective of this scenario is to classify a given input breathing sound into one of two categories: ”healthy” (comprising normal respiratory sounds) and ”unhealthy” (encompassing respiratory cycles exhibiting crackles, wheezes, and the combination of both crackles and wheezes).

To carry out this evaluation, we leveraged the baseline Convolutional Neural Network (CNN) model described in [10]. In addition to assessing the cochleogram representation, we also sought to compare its performance against different time-frequency (TF) representations. Furthermore, we extended our analysis by considering a recent approach proposed in [11]. This approach introduced the concept of artificial noise addition (ANA), which involves augmenting unhealthy respiratory sounds with a type of noise similar to that found in the real world. The aim of ANA is to enhance the features associated with abnormal respiratory sounds (ARS) and bolster the system's robustness in handling these sounds.

In this multifaceted evaluation, we did not limit ourselves to the baseline CNN

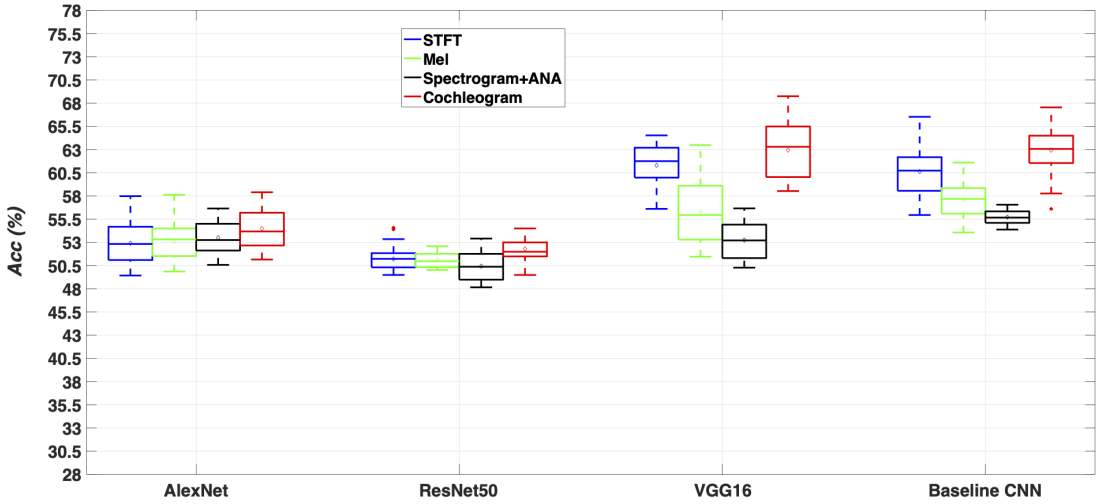


Fig. 5.3 Performance results, in terms of accuracy, for different CNN networks AlexNet, ResNet50, VGG16 and the CNN implemented in [10] of the STFT spectrogram, Mel-scaled spectrogram, the spectrogram+ANA features in [11] and Cochleogram evaluating four-classes scenario: normal vs. wheezes vs. crackles vs. wheezes+crackles. Each box represents 50 data points, each of them associated to a 10-fold cross validation of the database evaluated. The lower and upper lines of each box show the first and third quartile. The line in the middle of each box represents the median value. The diamond shape in the center of each box represents the average value. The lines extending above and below each box show the extent of the rest of the samples, excluding outliers. Finally, outliers are defined as points that are over 1.5 times the interquartile range from the sample median, which are depicted as crosses.

model. Instead, we explored several state-of-the-art deep learning architectures, including AlexNet [241], ResNet50 [5], and VGG16 [242]. These architectures are typically designed to work with images as input. To adapt the TF representation data for use with these architectures, we transformed the computed TF representation matrices into an image format. This transformation was achieved by applying the Viridis Color Map, a color mapping scheme that ensures a smooth gradient transition from blue to green to yellow [216].

Figure 5.3 indicates that the use of Cohleagram obtains the best accuracy results using any TF representation and CNN architecture evaluated in this work. In fact, VGG16 provides the best classification performance followed by our baseline CNN model whereas AlexNet and ResNet50 provide similar results in terms of accuracy at the expense of drastically reducing their classification rates. Similar performances were observed in [216] but using pretrained image models from the general purpose Imagenet dataset [331]. On the contrary, in this paper we focus on analyzing the advantages of using alternatives TF representation rather than exploring the possible transfer learning solutions.Regarding the state-of-the-art method in [11], it can be observed that the Spectrogram+ANA strategy obtains competitive results using AlexNet and ResNet50 but its

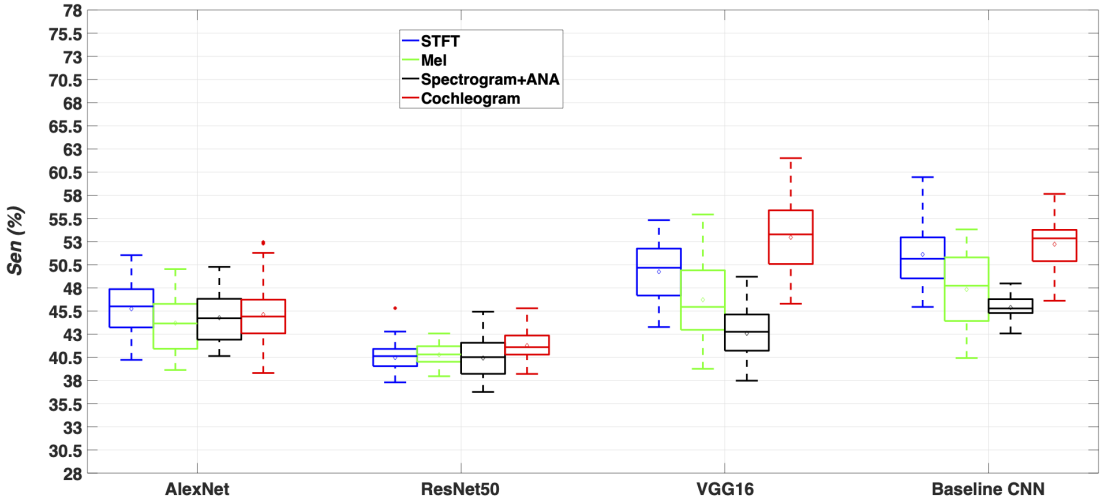


Fig. 5.4 Performance results of different CNN networks (AlexNet, ResNet50, VGG16 and the CNN implemented in [10]) of the STFT spectrogram, Mel-scaled spectrogram, the Spectrogram+ANA features [11] and Cochleogram evaluating four-classes scenario (normal vs. wheezes vs. crackles vs. wheezes+crackles) in the ICBHI database in terms of Sensibility (Sen).

performance drops drastically with respect to the other TF representations when using VGG16 and the Baseline CNN. It suggests that the performance achieved by the Spectrogram+ANA in this work differs from that obtained in [11] probably since the ICBHI database is more complex to analyze because it includes high sound interferences in most respiratory cycles to simulate real acoustic environments.

In Figure 5.4, 5.5, 5.6, 5.7, we present a comprehensive set of classification standard metrics to offer a more in-depth understanding of the methods' performance, thereby facilitating comparisons with other approaches evaluated using the ICBHI dataset. These metrics, namely sensitivity (Figure 5.4), specificity (Figure 5.5), score (Figure 5.6), and precision (Figure 5.7), were explained in Chapter 3. They play a crucial role in shedding light on various aspects of the performance.

Let's delve into the insights provided by these metrics:

- Sensitivity (Figure 5.4): Sensitivity measures the ability of a method to correctly identify the presence of adventitious respiratory sounds (ARS). In this context, higher sensitivity values indicate that the methods are more effective at detecting ARS. Sensitivity essentially quantifies the true positive rate, revealing how well the system reacts to ARS events.

- Specificity (Figure 5.5): Specificity assesses the method's capacity to accurately classify normal (healthy) respiratory sounds. Higher specificity values imply better performance in correctly identifying normal events. Specificity quantifies the true negative rate and indicates the system's proficiency in distinguishing normal sounds from abnormal ones.

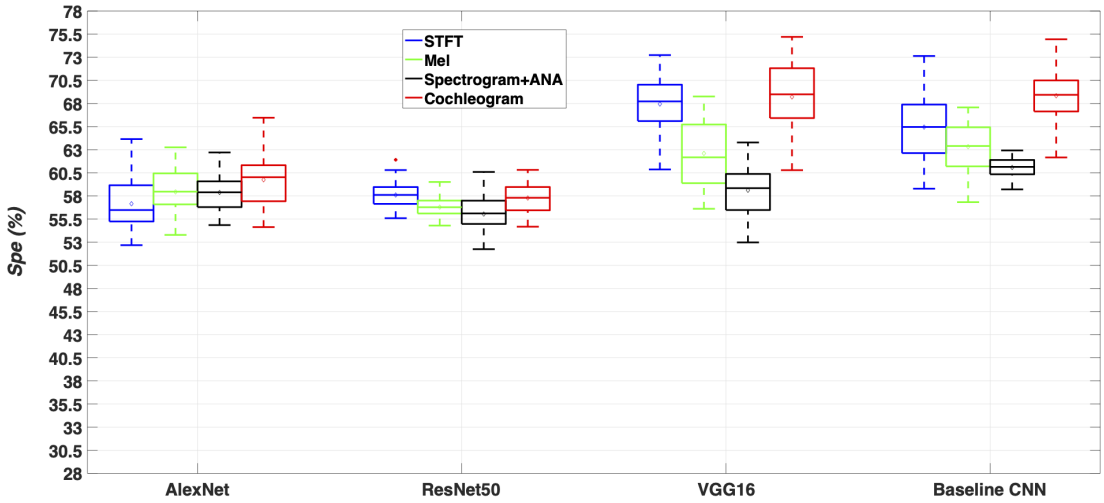


Fig. 5.5 Performance results of different CNN networks (AlexNet, ResNet50, VGG16 and the CNN implemented in [10]) of the STFT spectrogram, Mel-scaled spectrogram, the Spectrogram+ANA features [11] and Cochleogram evaluating four-classes scenario (normal vs. wheezes vs. crackles vs. wheezes+crackles) in the ICBHI database in terms of Specificity (Spe).

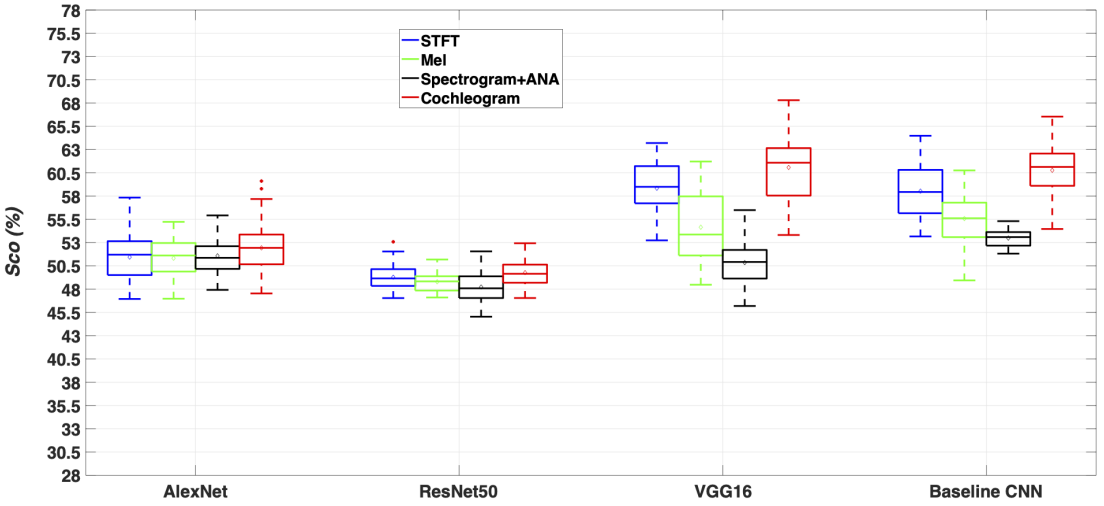


Fig. 5.6 Performance results of different CNN networks (AlexNet, ResNet50, VGG16 and the CNN implemented in [10]) of the STFT spectrogram, Mel-scaled spectrogram, the Spectrogram+ANA features [11] and Cochleogram evaluating four-classes scenario (normal vs. wheezes vs. crackles vs. wheezes+crackles) in the ICBHI database in terms of Score (Sco).

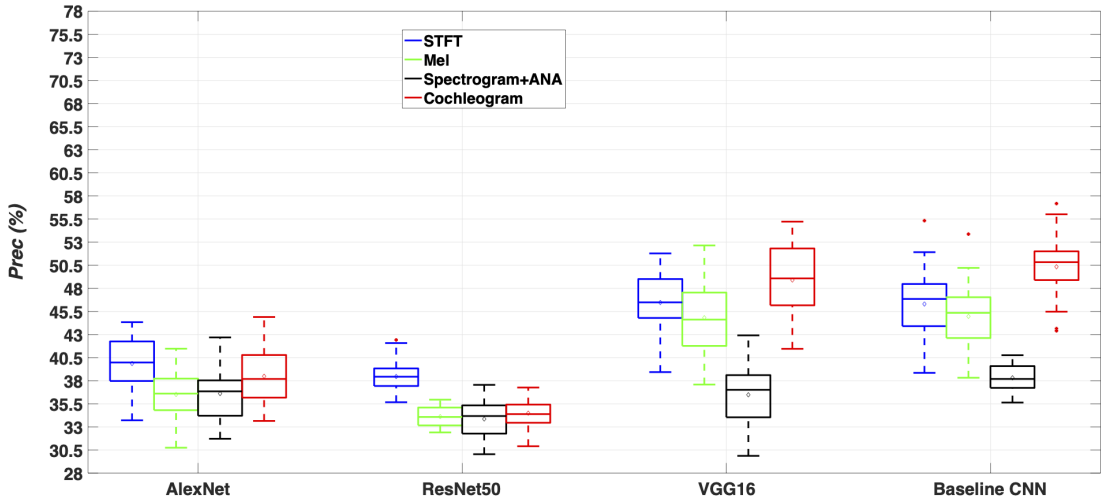


Fig. 5.7 Performance results of different CNN networks (AlexNet, ResNet50, VGG16 and the CNN implemented in [10]) of the STFT spectrogram, Mel-scaled spectrogram, the Spectrogram+ANA features [11] and Cochleogram evaluating four-classes scenario (normal vs. wheezes vs. crackles vs. wheezes+crackles) in the ICBHI database in terms of Precision (Pre).

- Precision (Figure 5.7): Precision gauges the accuracy of the system in classifying adventitious sounds correctly while minimizing false alarms. It calculates the ratio of true positive predictions to the total number of positive predictions. In this case, precision evaluates how well the system identifies ARS while avoiding erroneous classification of normal events.

- Score (Figure 5.6): The score metric represents the harmonic mean of sensitivity and specificity. It provides a balanced assessment of the method's performance in both detecting ARS (sensitivity) and accurately classifying normal sounds (specificity).

The graphical representations of these metrics offer a comprehensive view of the classification behavior of the evaluated TF representations and deep learning architectures in different scenarios, including both binary and four-class classification. Notably, higher specificity values (Figure 5.5) suggest that all evaluated TF representations and CNN architectures perform better in correctly classifying normal sounds, indicating their proficiency in identifying healthy respiratory events. On the other hand, the precision (Figure 5.7) and sensitivity values (Figure 5.4) reveal the methods' effectiveness in accurately classifying ARS and distinguishing them from normal events. In general, the methods exhibit slightly better results in terms of sensitivity, suggesting their heightened reactivity in predicting adventitious sounds over erroneously classifying ARS as normal events.

Finally, the score (Figure 5.6) provides a balanced evaluation, emphasizing that the classification behavior of the assessed TF representations remains consistent across var-

Comparison	Mann-Whitney U Test	Wilcoxon signed-rank test	Significantly better
AlexNet	**(p-value)**	**(p-value)**	
Cochleogram vs. STFT	0.018873	0.002397	yes
Cochleogram vs. Mel	0.004352	0.000374	yes
Cochleogram vs. Spectrogram+ANA	0.020690	0.026728	yes
ResNet50			
Cochleogram vs. STFT	$1.94e-05$	0.0001595	yes
Cochleogram vs. Mel	$4.69e-07$	$1.74e-06$	yes
Cochleogram vs. Spectrogram+ANA	$1.13e-07$	$6.16e-07$	yes
VGG16			
Cochleogram vs. STFT	0.006779	0.002555	yes
Cochleogram vs. Mel	$1.19e-12$	$2.66e-09$	yes
Cochleogram vs. Spectrogram+ANA	$6.94e-18$	$7.55e-10$	yes
Baseline CNN			
Cochleogram vs. STFT	$6.94e-05$	$3.37e-05$	yes
Cochleogram vs. Mel	$5.39e-14$	$1.41e-08$	yes
Cochleogram vs. Spectrogram+ANA	$8.84e-18$	$7.55e-10$	yes

Table 5.2 Man-Whitney U Test and Wilcoxon signed-rank test Results for the sets of results obtained in the Figure 5.3, using a significance level α = 0.05.

ious scenarios and deep learning architectures. This uniformity in performance is noteworthy and underscores the reliability of the methods in detecting respiratory abnormalities. These metrics offer a comprehensive understanding of the methods' performance, revealing their robustness and effectiveness in the classification of respiratory sounds, regardless of the specific scenario or architecture used. This comprehensive analysis is invaluable for advancing the development of diagnostic tools in the field of respiratory healthcare.

In previous subsection, we explained the significance of employing the Mann-Whitney U Test and the Wilcoxon signed-rank test to establish the statistical significance of the Cochleogram as an input feature for four different CNN architectures in comparison to using other spectrogram-based representations, including STFT and Mel-scaled spectrograms, as well as an ANA-based method [11]. The significance level, denoted as α, was set at $\alpha = 0.05$, which is a common threshold in statistical testing to determine the presence of statistically significant differences.

The results of these statistical tests are presented in Table 5.2, and they indicate that the Cochleogram significantly outperforms the other studied TF representations in combination with each of the four deep learning architectures. In other words, the statistical analysis confirms that our proposal, which leverages the Cochleogram as the input feature, leads to a significant improvement in the learning process of CNN architectures when it comes to the classification of respiratory sounds, in comparison to the other standard TF representations that were evaluated.

This statistical validation underscores the robustness and effectiveness of the Cochleogram as a representation for CNN-based classification tasks in the domain of respiratory sound analysis. It demonstrates that the Cochleogram is not only a promising but also a statistically superior choice for enhancing the performance of deep learning models

when dealing with the classification of respiratory sounds. This finding has significant implications for the development of advanced diagnostic tools and systems in the field of respiratory healthcare.

5.3.4 Related works in the field

The classification of adventitious respiratory sounds (ARS) has become a prominent area of research in recent years, largely due to the promising results achieved through machine learning techniques. However, most existing methods have primarily relied on two main time-frequency representations: the Short-Time Fourier Transform (STFT) and the Mel spectrogram when preparing data for use in neural network architectures. To address these limitations, this study conducted a comprehensive investigation into the potential of various time-frequency representations, ultimately advocating for the Cochleogram as a more effective alternative. This unique representation is particularly well-suited for modeling the distinctive temporal and spectral characteristics often exhibited by the majority of adventitious respiratory sounds.

In addition to this, our research systematically examined the influence of these different time-frequency representations when used in conjunction with several state-of-the-art Convolutional Neural Network (CNN)-based architectures.

Based on our extensive evaluation using the ICBHI database, we have found strong evidence that the Cochleogram significantly outperforms other spectrograms in the context of ARS classification. Whether applied to binary classification tasks (distinguishing normal from ARS) or more complex multi-class problems (categorizing normal sounds, crackles, wheezes, and combined crackles and wheezes), the Cochleogram consistently emerged as the superior choice.

Table 3.4 from Chapter 3 provides a valuable comparison with recent state-of-the-art methods documented in the literature, all of which pertain to the classification of adventitious respiratory sounds, specifically across the four classes defined in the ICBHI database. Notably, most of these methods involve the use of STFT in their data preprocessing phase to compute the time-frequency representation, followed by CNN-based approaches for classification. It is essential to highlight the variations in performance among these methods, primarily due to the absence of standard evaluation practices. For instance, some studies utilize only a subset of the ICBHI database instead of the complete dataset employed in our study. Furthermore, different metrics were often used for evaluation, further complicating direct comparisons between these methods. In acordance with this table, we added the work developed in this Thesis acording the same parameters as it can be seen in Table 5.3.

It is important to acknowledge that the results obtained across various methods exhibit a wide range of performance values. This variability can be attributed to the lack

Authors	Time-frequency representation		RC (s)	Technique	Train/Test	Results (%)			
	Type	Parameters				Sen	Spe	Sco	Acc
[211]	STFT	30 ms	-	HMM	60/40	–	–	39.6	–
[212]	STFT	500 ms	-	RNN	- (5-fold)	58.4	73.0	65.7	–
[213]	STFT	512 ms	-	HMM SVM	60/40	20.81	78.5	49.65	49.43
[224]	Mel	250 ms	–	RNN	80/20	**64.0**	**84.0**	**74.0**	–
[214]	STFT, Wavelet	20 ms, $D_2 - D_7, A_7$	–	bi-ResNet	– (10-fold)	31.1	69.2	50.2	52.8
[215]	STFT, Scalogram	40 ms	–	CNN	60/40	28.0	81.0	54.0	–
[216]	STFT	$64 - 128 - 524$ ms	–	CNN SVM	– (10-fold)	–	–	–	65.5
[217]	STFT	20 ms	–	ResNet NL	60/40	41.3	63.2	52.3	–
[225]	Mel	60 ms	–	CNN RNN	80/20	–	58.01	–	–
[218]	STFT	100 ms	2.5	ResNet SE SA	70/30	17.8	81.3	49.6	–
[226]	Mel	–	–	CNN	60/40	–	–	–	**80.4**
[219]	STFT	40 ms	–	CNN bi-LSTM	– (5-fold)	63.0	83.0	73.0	–
[312]	Wavelet	30 ms	–	DAG HMM	–	–	–	–	50.1
[227]	Mel	–	7	CNN	60/40	40.1	72.3	56.2	–
[10]*	STFT	32 ms 64 filters	6	CNN	80/20 (10-fold)	51.61	65.45	58.53	60.61
	Mel			CNN		47.83	63.33	55.58	57.56
	STFT + Mel			CNN		46.97	63.97	55.47	57.33
[228]	STFT, Log-mel	32 ms, 50 bins	8	ResNet	60/40	37.2	79.3	58.3	–
[11]*	Spectrogram + ANA	8 ms 64 filters	6	CNN (AlexNet)	80/20 (10-fold)	44.77	58.37	51.57	53.49
				CNN (ResNet50)		40.42	56.03	48.23	50.43
				CNN (VGG16)		43.08	58.61	50.85	53.24
				CNN (Baseline)		45.88	61.08	53.48	55.71
This work	Cochleagram	84 ms 64 filters	6	CNN (AlexNet)	80/20 (10-fold)	45.12	59.75	52.43	54.48
				CNN (ResNet50)		41.78	57.78	49.78	52.31
				CNN (VGG16)		53.45	68.71	61.08	62.94
				CNN (Baseline)		52.71	68.84	60.78	62.93

Table 5.3 Comparison between the results developed in this Thesis and the state-of-the art methods evaluating the four-classes (normal vs. wheezes vs. crackles vs. crackles+wheezes) classification performance in the ICBHI database. Respiratory cycle (RC) represents the temporal length (in seconds) including zero padding to create respiratory cycles of fixed duration. bi-ResNet: bilinear ResNet, NL: non-local, SE: Squeeze-and-Excitation, SA: Spatial Attention, bi-LSTM: bi-directional LSTM, DAG: Directed Acyclic Graph. The rest of the acronyms have been previously mentioned. The references followed by $*$ means that the method has been implemented in this Thesis following the authors description. The results for other methods have been directly extracted from the corresponding works. In bold letter is indicated the maximum value for each metric.

of uniformity in the evaluation process. Specifically, some of the methods opt to use only a subset of the ICBHI database, as seen in the study by Chanane et al. [226], in contrast to our approach, where we utilized the entire ICBHI database. Additionally, the metrics employed for the evaluation of these different methods differ significantly from one another. This discrepancy in evaluation metrics makes it challenging to directly compare the outcomes across these diverse approaches, which makes it difficult to compare them all [211, 225, 216, 226, 312].

In the context of these differences and challenges, it is noteworthy that the highest-performing method, as reported by Chanane et al. [226], achieved an accuracy (Acc) of 80.4%. While this result is impressive, it's essential to note that some of the best-performing approaches employ different evaluation metrics specifically tailored to the ICBHI database. For instance, the Mel+RNN method proposed by Perna et al. [224] achieved a sensitivity (Sen) of 64.0%, specificity (Spe) of 84.0%, and an F1 score (Sco) of 74.0%. These metrics offer a more detailed evaluation of classification performance within the specific context of the ICBHI database.

This wide range of performance values, variations in evaluation methodologies, and the notable differences in evaluation metrics collectively suggest that there is still ample room for improvement in the field of biomedical signal processing and machine learning. As the research progresses and a more standardized evaluation framework is established, further advancements can be expected in the classification of adventitious respiratory sounds, ultimately leading to more accurate and clinically relevant results.

Notably, the highest-performing method, Chanane et al. [226], achieved an accuracy (Acc) of 80.4%. In terms of the standard ICBHI database metrics, the Mel+RNN approach [224] emerged as the top performer, boasting a sensitivity (Sen) of 64.0%, specificity (Spe) of 84.0%, and an F1 score (Sco) of 74.0%. These findings underscore the potential for further advancements in the field of biomedical signal processing and machine learning.

An intriguing observation is that specificity (Spe) consistently outperformed sensitivity (Sen) across different methods. This suggests that neural networks tend to be more proficient at generalizing features characterizing normal respiratory sounds. The prevalence of normal sounds in the database, constituting 53% of the entire ICBHI dataset, provides a more substantial foundation for reliable feature modeling when compared to the relatively scarce occurrences of adventitious sounds. Nevertheless, this study's primary goal was to highlight that the Cochleogram greatly enhances the learning process within deep learning architectures when classifying adventitious respiratory sounds, regardless of the inherent class imbalance.

CHAPTER 6

Classification of Adventitious Sounds combining Cochleogram and Vision Transformers

6.1 Abstract

Early identification of respiratory irregularities is critical for improving lung health and reducing global mortality rates. The analysis of respiratory sounds plays a significant role in characterizing the respiratory system's condition and identifying abnormalities. The main contribution of this study is to investigate the performance when the input data, represented by cochleogram, is used to feed the Vision Transformer (ViT) architecture since this input-classifier combination is the first time it is applied to adventitious sound classification to our knowledge. Although ViT has shown promising results in audio classification tasks by applying self-attention to spectrogram patches, we extend this approach by applying the cochleogram, which captures specific spectro-temporal features of adventitious sounds. The proposed methodology is evaluated on the ICBHI dataset. We compare the classification performance of ViT with other state-of-the-art CNN approaches using spectrogram, Mel frequency cepstral coefficients, constant-Q transform, and cochleogram as input data. Our results confirm the superior classification performance combining cochleogram and ViT, highlighting the potential of ViT for reliable respiratory sound classification. This study contributes to the ongoing efforts in developing automatic intelligent techniques with the aim to significantly augment the speed and effectiveness of respiratory disease detection, thereby addressing a critical need in the medical field.

6.2 Contribution

The study conducted by Rocha et al. [10] addresses the use of Convolutional Neural Networks (CNNs) in the field of respiratory sound classification. Despite the widespread adoption of CNNs in various research areas, the authors suggest that there is room for improvement in the classification of respiratory sounds. They propose two main aspects for enhancement: the exploration of alternative time-frequency representations and the utilization of novel deep learning architectures.

Time-Frequency Representations:

In the first aspect, the authors focused on the selection of appropriate time-frequency representations. Time-frequency representations are essential for capturing both temporal and spectral characteristics of audio signals, making them suitable for classifying respiratory sounds. The study conducted an extensive analysis of various representations and suggested the use of the cochleogram. The cochleogram is a representation that effectively captures the unique features of most adventitious respiratory sounds.

The results of this work showed that the cochleogram could be effectively used as an input to conventional CNN-based architectures. It was employed in the classification of respiratory sounds into binary categories (normal vs. abnormal respiratory sounds) as well as four distinct classes of abnormal respiratory sounds, including normal respiratory sounds, crackles, wheezes, and a combination of crackles and wheezes.

Deep Learning Architectures (Vision Transformer - ViT):

In the second aspect, the study aimed to investigate the potential of the Vision Transformer (ViT) architecture for respiratory sound classification. The ViT architecture, initially designed for image data, has already been applied to audio classification tasks [332, 333, 334]. The key idea behind ViT is to treat audio signals as a series of spectrogram patches, which are then processed by the model using self-attention mechanisms.

The self-attention mechanism allows ViT to identify important patterns and connections within the audio signal, thereby extracting relevant features for classification purposes. While this approach is still in its early stages, it holds promise for improving the accuracy and efficiency of audio classification tasks.

What makes this study novel is that it extends the ViT architecture to the specific domain of respiratory sound classification. Instead of using conventional time-frequency representations, as in previous work (referred to as [335]), this study leverages the cochleogram as the input data representation. This choice aims to harness the strengths of ViT in capturing complex audio features and patterns, potentially leading to improved classification accuracy for respiratory sounds.

In our investigation, a diverse set of metrics was incorporated, and the 10-fold cross-validation method was employed to categorize patients into training, testing, and validation sets. This involved dividing the entire patient dataset into 10 equivalent segments or 'folds,' with each fold serving as the testing set in rotation while the remaining 9 folds were utilized for training. This process was iterated 10 times, ensuring that each dataset segment was employed precisely once as the testing set.

The adoption of 10-fold cross-validation guarantees a balanced and thorough training and validation process for the model. Breaking down the data into 10 parts minimizes the potential for biases or anomalies that might arise from simpler splits, such as a 70-30 or 80-20 division.

The same training methodology has been employed for the compared architectures

Architectures	(Conv. Layers)	(Pool Layers)	(Activation)	(Parameters)
BaselineCNN	$2(5x5, 3x3)$	$2(2x2)$	$LeakyReLU$	$8M$
AlexNet	$5(11x11, 5x5, 3x3)$	$3(3x3, 2x2)$	$ReLU$	$160M$
VGG16	$13(3x3)$	$5(2x2)$	$ReLU$	$138M$
ResNet50	$50(7x7, 3x3, 1x1)$	$1(3x3)$	$ReLU$	$25.5M$

Table 6.1 A comprehensive overview of several conventional CNN architectures used in this work.

in Table 6.1. The training process involved a total of 30 epochs, utilizing a batch size of 16, a learning rate set at 0.001, and the adaptive data momentum (ADAM) optimization algorithm. The final value for the employed metrics (i.e., accuracy, sensitivity, specificity, etc.) is computed by averaging the individual values for each 10-fold iteration.

The research experiments were conducted utilizing Tensorflow and Keras, which were installed on a computer equipped with an Intel(R) Core(TM) 12th Gen i9-12900, a NVIDIA GeForce RTX3090 GPU, and 128 GB of RAM.

In order to assess the computational implications of the research, Table 6.2 is provided to elucidate the time breakdown, measured in minutes per epoch, during the training of individual models. The computational cost across various neural network architectures can differ significantly owing to variations in model architecture, depth, and design choices. While deeper architectures generally enhance representation learning, they concurrently escalate computational requirements. In this particular task, the heightened depth of AlexNet results in the highest computational cost. Conversely, VGG exhibits increased computational demands due to its uniform architecture and the incorporation of smaller convolutional filters (3x3) across multiple layers. ResNet, characterized by its depth, also incurs a higher computational cost, albeit mitigated by the use of skip connections (residual blocks) addressing the vanishing gradient problem. The BaselineCNN, being relatively shallow, bears a lower computational cost compared to deeper counterparts like AlexNet, ResNet, and VGG, rendering it suitable for simpler tasks or scenarios with restricted computational resources. The computational cost associated with Vision Transformers is variable. While they may necessitate fewer parameters than traditional CNNs for specific tasks, the introduction of the self-attention mechanism introduces additional computational complexity. The selection of an architecture is often contingent on the task's complexity. For more intricate tasks, deeper architectures like Vision Transformers may prove advantageous, exploring novel paradigms that potentially offer competitive performance with reduced computational overhead.

In summary, the study by Rocha et al. emphasizes the importance of exploring alternative data representations and novel deep learning architectures for improving respiratory sound classification. By applying the Vision Transformer architecture with

Comparison	time (min)/epoc
AlexNet	53.2868
ResNet50	20.6706
VGG16	39.5655
BaselineCNN	0.4488
ViT	4.9948

Table 6.2 Comparison of the computational time per epoc of the different architectures.

cochleograms as input, the authors aim to advance the state-of-the-art in this field and potentially enhance the accuracy and efficiency of diagnosing respiratory conditions based on sound analysis.

6.3 Experimental results

In this section, we assess the performance of the Vision Transformer (ViT) in the context of adventitious sound classification, comparing it to other state-of-the-art methods. To evaluate the effectiveness of our proposed approach, we conduct experiments using the ICBHI dataset in two distinct scenarios.

In the first scenario, we perform a binary classification task, where the objective is to determine the presence of two specific respiratory conditions: wheezes and crackles, during each respiratory cycle. This binary classification allows us to identify whether these abnormal sounds are present or absent.

In the second scenario, we tackle a more complex classification task, involving four distinct classes. Here, we aim to categorize respiratory sounds into one of four classes: ”healthy” for normal respiratory sounds, ”wheezing” for instances where wheezes are present, ”crackles” for cases with crackles, and ”both” when both wheezing and crackles are detected in the respiratory cycle. This multiclass classification provides a more comprehensive analysis of respiratory sound patterns and the ability to distinguish between different pathological conditions and normalcy.

Through these two scenarios, we aim to thoroughly evaluate the ViT’s performance in identifying and classifying adventitious respiratory sounds, demonstrating its capabilities in both binary and multiclass classification settings. Additionally, we compare the ViT’s performance to other state-of-the-art methods to assess its effectiveness and potential as an innovative approach for respiratory sound analysis.

6.3.1 2-class (binary) classification results

In our study, we have conducted a comprehensive assessment of the Vision Transformer (ViT) in comparison to other cutting-edge architectural approaches. The primary focus of this evaluation is on the task of detecting the presence of two distinct respiratory

anomalies, namely "crackles" and "wheezes," in respiratory sound signals. Importantly, these signals are associated with individual respiratory cycles.

This evaluation is essential as it allows us to measure and compare the performance of the ViT with respect to well-established state-of-the-art architectures. The goal is to determine whether the ViT offers a competitive advantage in accurately identifying these pathological respiratory sounds within the context of individual breath cycles. This assessment provides insights into the ViT's suitability and effectiveness for applications related to respiratory sound analysis and diagnostics.

By comparing ViT's performance with other state-of-the-art architectures, we gain a better understanding of its capabilities and its potential as a valuable tool in the domain of healthcare and medical diagnostics, particularly in the detection of respiratory abnormalities such as crackles and wheezes.

Figure 6.1 presents the accuracy results of the evaluated models using the studied time-frequency representations (STFT, MFCC, CQT, and the cochleogram) as inputs for distinguishing wheezes from other sounds and crackles from other sounds. The compared neural network architectures include BaselineCNN, AlexNet, VGG16 and ResNet50, and the ViT (Vision Transformer). Results indicate that the proposed method, based on the ViT model using the cochleogram, achieved the best performance for both crackle and wheezing classification, with an average accuracy Acc=85.9% for wheezes and Acc=75.5% for crackles detection. In fact, employing transformers to capture bi-directional dependencies in COPD audio signals holds potential in predicting adventitious sounds, even in the presence of sparse sound events. It is worth noting that the STFT spectrogram also provided competitive performance, with an accuracy Acc=82.1% for wheezes and Acc=72.2% for crackles. These results may be due to the fact that the effective low-pass filtering of the frequencies of interest and the use of an appropriate window length and hop size, resulting in the accurate detection of adventitious sound events even when a linear frequency scale is used. Interestingly, the log-scale frequency transforms (MFCC and CQT) clearly underperform, showing Acc=79.9% for wheezes and Acc=70.1% for crackles detection in the case of MFCC and Acc=78.8% for wheezes and Acc=68.8% using the CQT spectrogram. Although MFCC and CQT are effective for modeling speech and music signals, both do not seem to be the most appropiate TF representation for capturing the most predominant content in the context of adventitious sounds. Comparable behavior is observed among the state-of-the-art neural network architectures when applied to the task of crackle detection, with the exception of the AlexNet model. In the case of the AlexNet model detecting crackles, the MFCC yield the highest accuracy result, specifically Acc=67.9%, among the compared TF input representations. However, this accuracy is diminished by approximately 8% when compared AlexNet to the peak performance, Acc=75.5%, achieved by the proposed method that employs the Vision Transformer (ViT) architec-

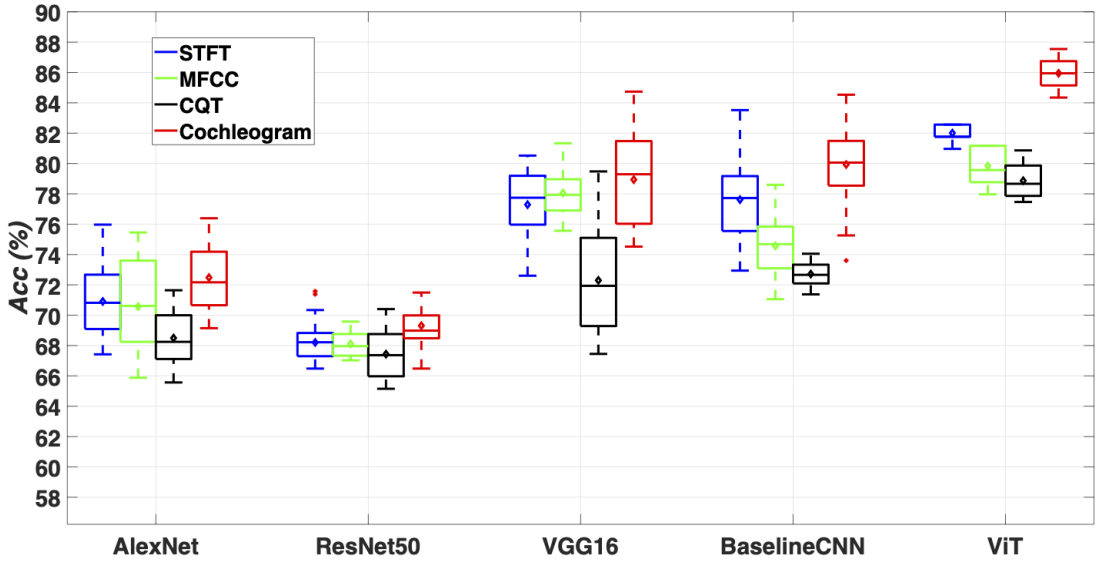


Fig. 6.1 Accuracy results for the evaluated deep learning architectures, with data feeding from feature extraction based on TF representations, in the task of 2-classes scenario wheezes (yes/no) in the ICBHI database.

ture fed from the cochleogram input. It is also interesting to highly the narrower dispersion of the results obtained by the ViT architecture which demonstrate the robustness of this method w.r.t the different acoustic conditions of the input respiratory cycle. Finally, BaselineCNN and VGG16 outperform the AlexNet and ResNet50 architectures.

To assess the statistical significance of the results presented in Figure **??**, two widely referenced and robust non-parametric tests have been used, specifically, the Man-Whitney U Test and the Wilcoxon signed-rank test [**? ?**]. In particular, Tables 6.3 and 6.4 display the results of these tests, which compared the classification performance of the ViT, VGG16, BaselineCNN, AlexNet, and ResNet50 models using cochleogram representation for both crackles and wheezes. The null hypothesis H_o for these tests assumes that there is no significant difference between the two distributions being compared, while the alternative hypothesis H_1 postulates that a significant difference exists. The p-value, which indicates the probability of obtaining results as extreme as those observed assuming that H_o is true, determines whether we reject or accept H_o. Our analysis, using a significance level of $\alpha = 0.05$, reveals that the p-value for all cases did not exceed the significance level, allowing us to reject H_o and conclude that the ViT performs significantly better than VGG16, BaselineCNN, AlexNet, and ResNet50 for both crackles and wheezes classification.

In this work, the accuracy (Acc) has been used as the main metric to provide a general measure of the classification performance, taking into account both successful adventitious events (TP and TN) as well as false adventitious events (FP) and undetected adventitious events (FN). In order to compare the proposed method with other state-of-the-art algorithms, Table 6.5 shows other metrics, such as Sensitivity (Sen),

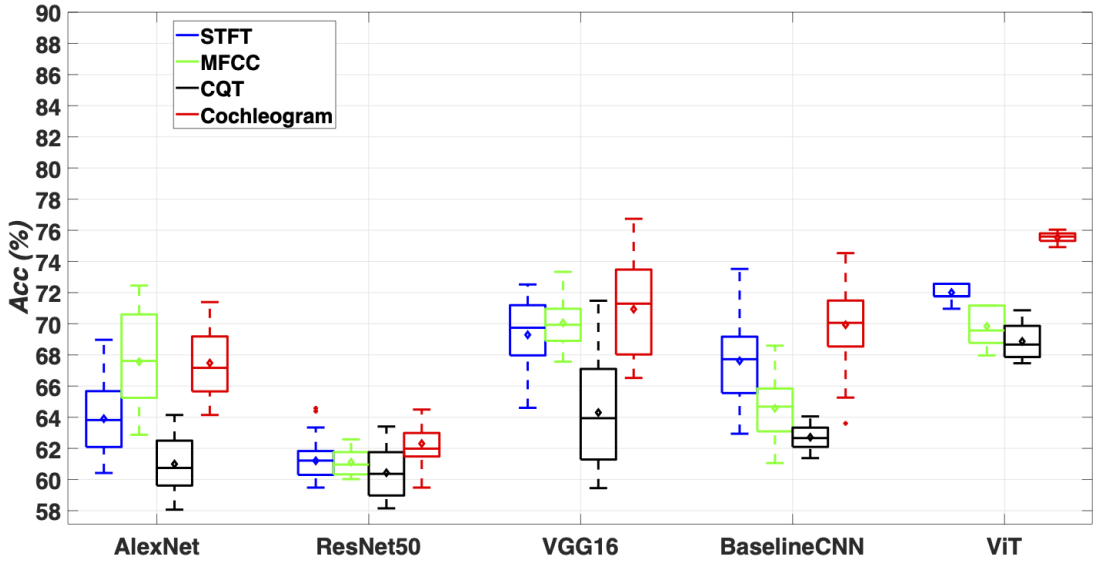


Fig. 6.2 Accuracy results for the evaluated deep learning architectures, with data feeding from feature extraction based on TF representations, in the task of 2-classes scenario crackles (yes/no) in the ICBHI database.

Comparison	Mann-Whitney U Test	Wilcoxon signed-rank test	Significantly better
Cochleogram	**(p-value)**	**(p-value)**	**(yes/no)**
ViT vs. VGG16	$4.99e-18$	$6.13e-13$	yes
ViT vs. BaselineCNN	$5.34e-18$	$8.61e-15$	yes
ViT vs. AlexNet	$6.91e-18$	$5.68e-15$	yes
ViT vs. ResNet50	$8.16e-18$	$6.84e-15$	yes

Table 6.3 The Man-Whitney U Test and Wilcoxon signed-rank test were performed on the data sets shown in Figure 6.1 with a significance level of $\alpha = 0.05$.

Comparison	Mann-Whitney U Test	Wilcoxon signed-rank test	Significantly better
Cochleogram	**(p-value)**	**(p-value)**	**(yes/no)**
ViT vs. VGG16	$9.28e-18$	$1.77e-15$	yes
ViT vs. BaselineCNN	$2.74e-17$	$1.64e-14$	yes
ViT vs. AlexNet	$6.35e-18$	$1.43e-13$	yes
ViT vs. ResNet50	$5.10e-18$	$1.14e-11$	yes

Table 6.4 The Man-Whitney U Test and Wilcoxon signed-rank test were performed on the data sets shown in Figure 6.2 with a significance level of $\alpha = 0.05$.

		Sensibility (Sen)		Specificity (Spe)		Score (Sco)		Precision (Pre)	
Model	**TF**	**Wheezes**	**Crackles**	**Wheezes**	**Crackles**	**Wheezes**	**Crackles**	**Wheezes**	**Crackles**
AlexNet	STFT	65.1	55.1	70.1	60.1	67.6	57.6	44.8	44.8
	MFCC	62.8	52.8	68.9	59.9	65.8	56.3	39.5	39.5
	CQT	61.7	51.7	68.3	55.3	65.0	53.5	37.6	37.6
	Cochleogram	66.3	55.1	72.1	62.7	69.2	58.9	44.4	44.4
ResNet50	STFT	61.7	51.7	72.1	62.1	66.9	56.9	38.4	38.4
	MFCC	59.0	49.0	69.1	59.0	64.0	54.0	38.1	38.1
	CQT	58.4	48.4	69.0	59.0	64.7	53.7	36.8	36.8
	Cochleogram	62.2	51.7	71.7	61.7	66.9	56.7	39.4	39.4
VGG16	STFT	69.4	59.4	81.9	71.9	75.6	65.6	46.4	46.4
	MFCC	62.7	52.7	76.5	66.5	69.6	59.6	44.8	44.8
	CQT	59.0	49.0	72.6	62.6	65.8	66.8	36.4	36.4
	Cochleogram	71.6	59.4	82.7	72.7	77.1	66.0	48.9	48.9
BaselineCNN	STFT	66.7	61.7	82.4	72.4	74.5	67.0	46.3	46.3
	MFCC	62.8	56.8	80.3	70.3	71.58	63.5	44.9	44.9
	CQT	60.8	53.8	78.0	68.0	69.4	60.9	38.3	38.3
	Cochleogram	67.7	62.8	85.8	75.8	76.7	65.3	50.3	50.3
ViT	STFT	71.9	62.9	85.0	75.0	78.5	69.0	52.4	52.4
	MFCC	67.9	59.9	82.9	72.9	75.4	66.4	50.3	50.3
	CQT	65.9	57.9	80.5	70.5	73.2	64.2	47.7	47.7
	Cochleogram	**76.0**	**65.2**	**91.0**	**80.2**	**83.5**	**71.7**	**57.6**	**57.6**

Table 6.5 Sensibility Sen, specificity Spe, score Sco and precision Pre results for the proposed method and the other evaluated neural network architectures applying different TF representations for the task of binary 2-class scenario crackles (yes/no) in the ICBHI database. The maximum value for each metric is highlighted in bold.

Specificity (Spe), Score (Sco), and Precision (Pre), which have also been proposed in the literature to assess the performance associated to the adventitious sound classification. The results show that using the ViT architecture with cochleogram as input gives the best classification performance for each type of adventitious sound. Compared to using STFT, using cochleogram improves wheezes' classification by about 4.1% and crackles' classification by about 2.3%, on average. STFT is ranked in second place in terms of performance, followed by MFCC and CQT, which rank last. Specifically, STFT outperforms MFCC by at least 2.1% on average for both wheezes and crackles. Furthermore, MFCC outperforms CQT with a minimum average improvement of 2.0% for both wheezes and crackles. Focusing on the behavior of the compared systems based on each metric, the highest values are obtained in terms of Specificity (Spe), reporting that the architectures are capable of accurately predicting when patients are healthy. However, the underperformance is shown in terms of Sensitivity (Sen) and Precision (Pre) since the lowest values are related to the Precision (Pre) for all evaluated neural network architectures. This fact reveals that the number of false positives (healthy patients classified as sick) exceeds the number of false negatives (sick patients classified as healthy). Nevertheless, this outcome can be considered highly advantageous from a medical standpoint since it provides assurance that patients who exhibit even the slightest doubt or uncertainty concerning the presence of a respiratory disease will receive immediate attention and the necessary medical care they require.

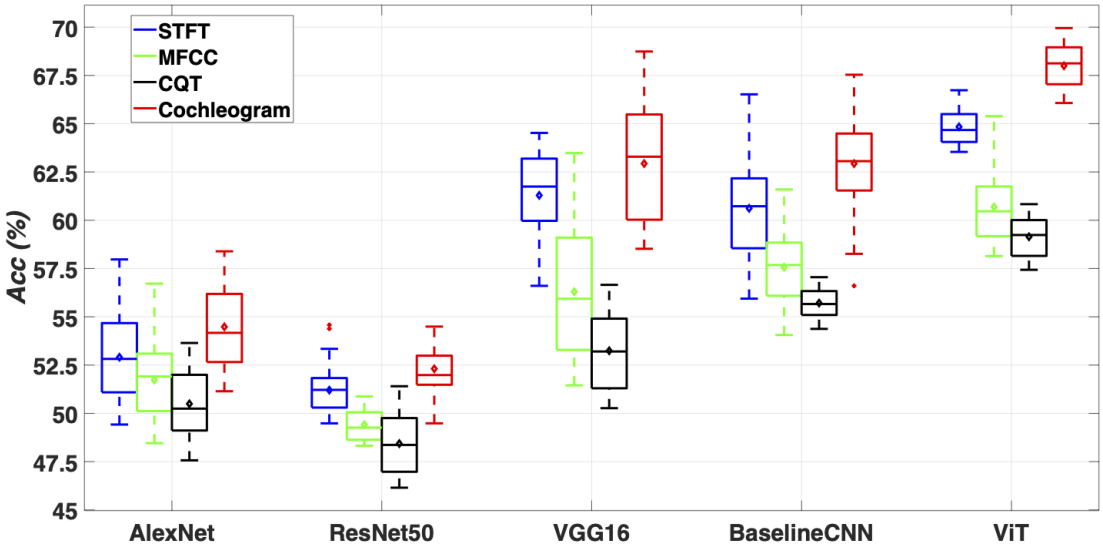


Fig. 6.3 Accuracy results for the evaluated deep learning architectures, with data feeding from feature extraction based on TF representations, in the task of 4-classes scenario (normal, wheezes, crackles, whezees+crackles) in the ICBHI database.

6.3.2 4-class classification results

We have evaluated the performance of the ViT w.r.t the other state-of-the-art architectures in a multiclass classification scenario. The objective is to identify between normal respiratory sounds (healthy) and respiratory sounds with any type of the following adventitious sounds, such as crackles, wheezes or crackles+wheezes from an input breathing cycle.

Figure 6.3 displays the obtained results in terms of accuracy (Acc). As can be seen, the ViT architecture using the cochleogram input TF representation outperforms all the compared methods ($Acc = 67.9\%$). Similar to the 2-class scenario, using the STFT provides better results than the other log-scale transforms (MFCC and CQT). Identical behaviour can be observed for all the compared TF representations, independently of the evaluated architecture. VGG16 and BaselineCNN using the cochleogram obtained competitive results ($Acc = 63.9\%$ for wheezes and 62.8% for crackles), and clearly outperform the results using the AlexNet and ResNet50 architectures.

To demonstrate the statistical significance of using the ViT architecture performance w.r.t the rest of the compared architectures, we used the same procedure as described in the 2-classes scenario. In particular, Table 6.6 shows the results of these tests, indicating that the ViT architecture significantly improves the 4-class classification of respiratory sounds compared to the other evaluated neural network architectures.

Although in this paper we have selected accuracy (Acc) as the main metric, the metrics of sensitivity, specificity, score and precision have been included to provide a more comprehensive analysis of the performance of the proposed method enabling comparison with other state-of-the-art methods as shown in Table 6.7. Similarly to the 2-class

Comparison	Mann-Whitney U Test	Wilcoxon signed-rank test	Significantly better
Cochleogram	**(p-value)**	**(p-value)**	**(yes/no)**
ViT vs. VGG16	$5.46e-15$	$3.67e-13$	yes
ViT vs. BaselineCNN	$5.61e-16$	$3.55e-15$	yes
ViT vs. AlexNet	$6.91e-18$	$1.77e-15$	yes
ViT vs. ResNet50	$6.59e-18$	$2.77e-15$	yes

Table 6.6 The Man-Whitney U Test and Wilcoxon signed-rank test were performed on the data sets shown in Figure 6.3 with a significance level of α = 0.05.

Model	TF	Sensibility (Sen)	Specificity (Spe)	Score (Sco)	Precision (Pre)
AlexNet	STFT	45.7	57.1	51.4	39.8
	MFCC	42.8	57.0	49.9	35.1
	CQT	42.8	57.0	49.4	34.3
	Cochleogram	45.12	59.74	52.43	38.48
ResNet	STFT	40.4	58.1	49.2	38.4
	MFCC	39.0	55.0	47.0	32.4
	CQT	38.4	54.0	46.2	31.8
	Cochleogram	41.7	57.7	49.7	34.4
VGG16	STFT	49.7	67.9	58.84	46.4
	MFCC	46.7	62.5	54.6	44.8
	CQT	43.0	58.6	50.8	36.4
	Cochleogram	53.4	68.7	61.0	48.9
BaselineCNN	STFT	51.6	65.4	58.5	46.3
	MFCC	47.8	63.3	55.5	44.9
	CQT	45.8	61.0	53.4	38.3
	Cochleogram	52.7	68.8	60.7	50.3
ViT	STFT	52.9	68.0	60.5	45.4
	MFCC	49.9	64.9	57.4	42.3
	CQT	47.9	63.5	55.7	40.7
	Cochleogram	**56.6**	**71.3**	**64.0**	**50.2**

Table 6.7 Sensibility, specificity, score and precision results for the proposed method and the other evaluated neural network architectures applying different TF representations for the task of 4-class scenario in the ICBHI database. The maximum value for each metric is highlighted in bold.

scenario, the best results are obtained using the ViT+cochleogram, independently of the compared architecture or input TF representation. In general, it can be observed that the classification performance is better in terms of Spe so, the architectures seem to characterize better healthy sounds than adventitious sounds. Moreover, as in the 2-class scenario, the false positive (healthy patients classified as sick) remains higher than the false negative (sick patients classified as healthy) and, consequently, Sen values are higher than Pre values. Moreover, it can be observed that the results in the 4-class scenario are worse than in the 2-class binary scenario. In fact, all previous metrics may provide lower values when considering the joint occurrence of crackles and wheezes as an independent class. That is, individual detection of wheezes or crackles when both are present is reported as a prediction error. However, to allow a fair comparison with other methods in the literature, we have used the same metric definitions than in the ICBHI challenge.

CHAPTER 7

Conclusions and Future work

7.1 Conclusions

In conclusion, obstructive lung diseases represent a pressing global health challenge characterized by their widespread prevalence, substantial morbidity and mortality rates, and significant socioeconomic burdens. The current reliance on the auscultation process by pulmonologists for respiratory system assessment is driven by its non-invasive, cost-effective, and user-friendly attributes, ensuring patient safety. However, the inherent subjectivity in auscultation-based diagnoses, influenced by the individual skills, experience, and training of physicians, leads to a notable number of misdiagnoses, compromising patient well-being and amplifying healthcare costs. Notably, the identification and analysis of adventitious sounds, such as wheezing and crackles, during auscultation are critical tasks, signalling potential obstructive lung diseases like asthma, bronchiolitis, bronchiectasis, or COPD. Addressing these challenges is imperative for advancing respiratory health diagnostics and mitigating the associated global health concerns. This study focuses on meeting the crucial demand for early detection of respiratory diseases, a paramount concern in global health. Auscultation, being a widely employed, non-invasive, and cost-effective technique, depends on the proficiency of physicians in recognizing abnormal respiratory sounds, such as crackles. Utilizing an electronic stethoscope, medical professionals can capture the sounds heard through a traditional stethoscope, enabling the generation of viable audio data for analysis through signal processing techniques. Our study proposes a comprehensive three-phase approach to enhance the process of respiratory disease detection using these audio from the electronica stethoscopes. Firstly, we introduce a novel method that combines autoregression-based spectral features and a Support Vector Machine (SVM) classifier for the accurate identification of crackle events in respiratory sound signals. This approach, featuring a robust preprocessing stage and short-term signal analysis, achieves competitive results with an accuracy ranging from 80% to 100% across various signal-to-noise ratios. The combination of AR and SVM proves effective, improving the precision of detection on simulated signals based on their mathematical formula. Secondly, addressing the oversight in time-frequency representations for Convolutional Neural Network (CNN)

models, we present the cochleogram as a superior representation for classifying respiratory adventitious sounds. The cochleogram exhibits exceptional performance, with an average accuracy of 85.1% in wheezes and 73.8% in crackles, highlighting its accuracy and robustness in CNN-based classification. Finally, we explore the application of the Vision Transformer (ViT) architecture in respiratory sound classification, emphasizing innovative input data representation. Our integration of the cochleogram with ViT showcases promising results, leveraging self-attention in audio classification. Despite the challenge of standardized databases in biomedical audio signal processing research, our methodology, evaluated on the ICBHI dataset, demonstrates the effectiveness of the cochleogram and the potential of ViT for reliable respiratory sound classification. This research significantly contributes to ongoing initiatives in advancing signal processing and artificial intelligence techniques, aiming to substantially improve the speed and efficacy of respiratory disease detection, addressing a critical imperative in the medical field.

7.2 Future work

In the context of signal processing applied to sound signals from auscultation, the scientific community continues to show a growing interest in the development of new methods and algorithms to enhance the reliability of early detection diagnoses for respiratory pathologies. This thesis has demonstrated that research focused on adventitious sounds, is a challenge with much ground to cover, both in terms of detecting and classifying adventitious sounds and improving the performance and computational cost of current methods. This section presents the main lines of future work that, while not addressed in the development of this doctoral thesis, could make significant contributions to the diagnosis of obstructive respiratory diseases.

- Development of a adventitious sounds tele-monitoring system for intelligent houses devices (for example, alexa). The aim is to create a reliable, non-invasive, low-cost, and individual monitoring tool that can be performed from the subject's home, making it easily accessible to the general population, especially to more vulnerable individuals such as the elderly and young children. This monitoring can be conducted as frequently as the subject deems necessary. It will help acoustically capture sounds that occur during typical respiratory crises, often happening in the early morning or during the night. Additionally, this approach supports a sustainable healthcare system in terms of both human and material resources.

- Quantum computing is a revolutionary paradigm of computation that leverages the principles of quantum mechanics to perform certain types of computations much more efficiently than classical computers. Classical computers, which in-

clude the computers we use every day, process information using bits, which can exist in one of two states, 0 or 1. Quantum computers, on the other hand, use quantum bits or qubits. It's important to note that quantum computing is still in its early stages, facing significant challenges such as error correction, maintaining qubit coherence, and scalability. Researchers and companies are actively working on developing practical quantum computers that can tackle real-world problems. The formulation of an algorithm for the identification and categorization of adventitious sounds through the application of these technologies would ameliorate the computational overhead associated with contemporary state-of-the-art machine learning algorithms.

- AutoML is a machine learning algorithm that automates the process of training and tuning machine learning models. There's no doubt that AutoML has been making waves in the world of machine learning lately! Originally developed by Google, AutoML has since proven itself as an invaluable tool for businesses of all sizes – from small startups looking for ways to speed up their development processes, right up through larger organizations who need automated methods for dealing with large amounts of data or complex modelling problems. As the previous approach, the AutoML could be an alternative to the standard machine learning algorithms.

REFERENCES

[1] Jitong Chen, Yuxuan Wang, and DeLiang Wang. 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, 2014.

[2] Bruno M Rocha, Dimitris Filos, Luís Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P Kahya, Nikša Jakovljevic, Tatjana L Turukalo, Ioannis M Vogiatzis, Eleni Perantoni, et al. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological measurement*, 40(3):035001, 2019.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[4] T Shanthi and RS Sabeenian. Modified alexnet architecture for classification of diabetic retinopathy images. *Computers and Electrical Engineering*, 76:56–64, 2019.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Sheldon Mascarenhas and Mukul Agarwal. A comparison between vgg16, vgg19 and resnet50 architecture frameworks for image classification. In *2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON)*, volume 1, pages 96–99. IEEE, 2021.

[7] Ravi Pal and Anna Barney. A dataset for systematic testing of crackle separation techniques. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4690–4693. IEEE, 2019.

[8] G Dorantes-Mendez, S Charleston-Villalobos, R Gonzalez-Camarena, G Chi-Lem, JG Carrillo, and T Aljama-Corrales. Crackles detection using a time-variant autoregressive model. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1894–1897. IEEE, 2008.

[9] Ravi Pal and Anna Barney. Iterative envelope mean fractal dimension filter for the separation of crackles from normal breath sounds. *Biomedical Signal Processing and Control*, 66:102454, 2021.

[10] Bruno Machado Rocha, Diogo Pessoa, Alda Marques, Paulo Carvalho, and Rui Pedro Paiva. Automatic classification of adventitious respiratory sounds: A (un) solved problem? *Sensors*, 21(1):57, 2020.

[11] Rizwana Zulfiqar, Fiaz Majeed, Rizwana Irfan, Hafiz Tayyab Rauf, Elhadj Benkhelifa, and Abdelkader Nasreddine Belkacem. Abnormal respiratory sounds classification using deep cnn through artificial noise addition. *Frontiers in medicine*, 8:714811, 2021.

[12] Juan P Garcia-Mendez, Amos Lal, Svetlana Herasevich, Aysun Tekin, Yuliya Pinevich, Kirill Lipatov, Hsin-Yi Wang, Shahraz Qamar, Ivan N Ayala, Ivan Khapov, et al. Machine learning for automated classification of abnormal lung sounds obtained from public databases: A systematic review. *Bioengineering*, 10(10):1155, 2023.

[13] World health organization. `https://www.who.int/es`, 2022.

[14] Technology and health foundation. `https://www.health.org.uk/topics/digital-technology`, 2022.

[15] World health organization. tobacco. `https://www.who.int/health-topics/tobacco#tab=tab_1`, 2022.

[16] Eurostat. respiratory diseases statistics. `https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Respiratory_diseases_statistics&oldid=541149#:~:text=The%20EU%27s%20standardised%20death%20rate,high%20as%20that%20for%20females.`, 2022.

[17] Instituto nacional de estadística. defunciones por año, enfermedades del sistema respiratorio y mes de defunción. `https://www.ine.es/jaxi/Tabla.htm?path=/COVID/t15/&file=02003.px&L=0`, 2022.

[18] World health organization. asthma. https://www.who.int/es/news-room/fact-sheets/detail/asthma, 2022.

[19] World health organization. copd. https://www.who.int/es/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd), 2022.

[20] Semergen. cost associated with misdiagnoses. https://semergen.es/gruposTrabajo/gtNoticiaDetalles.php?idGt=112&idN=740&idSub=3, 2022.

[21] Hans Pasterkamp, Steve S Kraman, and George R Wodicka. Respiratory sounds: advances beyond the stethoscope. *American journal of respiratory and critical care medicine*, 156(3):974–987, 1997.

[22] Brenda K Wiederhold, Pietro Cipresso, Daniele Pizzioli, Mark Wiederhold, and Giuseppe Riva. Intervention for physician burnout: a systematic review. *Open Medicine*, 13(1):253–263, 2018.

[23] Morkos Iskander. Burnout, cognitive overload, and metacognition in medicine. *Medical Science Educator*, 29(1):325–328, 2019.

[24] Rajkumar Palaniappan, Kenneth Sundaraj, and Nizam Uddin Ahamed. Machine learning in lung sound analysis: a systematic review. *Biocybernetics and Biomedical Engineering*, 33(3):129–135, 2013.

[25] Shoichi Matsunaga, Katsuya Yamauchi, Masaru Yamashita, and Sueharu Miyahara. Classification between normal and abnormal respiratory sounds based on maximum likelihood approach. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 517–520. IEEE, 2009.

[26] Renard Xaviero Adhi Pramono, Stuart Bowyer, and Esther Rodriguez-Villegas. Automatic adventitious respiratory sound analysis: A systematic review. *PloS one*, 12(5):e0177926, 2017.

[27] Merckmanual.wheezing. https://www.merckmanuals.com/home/lung-and-%20airway-disorders/symptoms-of-lung-disorders/wheezing., 2022.

[28] Kevin E Forkheim, David Scuse, and Hans Pasterkamp. A comparison of neural network models for wheeze detection. In *IEEE WESCANEX 95. Communications, Power, and Computing. Conference Proceedings*, volume 1, pages 214–219. IEEE, 1995.

[29] Steven Le Cam, Akram Belghith, Ch Collet, and Fabien Salzenstein. Wheezing sounds detection using multivariate generalized gaussian distributions. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 541–544. IEEE, 2009.

[30] Bor-Shing Lin, Huey-Dong Wu, Sao-Jie Chen, et al. Automatic wheezing detection based on signal processing of spectrogram and back-propagation neural network. *Journal of healthcare engineering*, 6:649–672, 2015.

[31] Yukio Nagasaka. Lung sounds in bronchial asthma. *Allergology International*, 61(3):353–363, 2012.

[32] MARIO Milicevic, IGOR Mazic, and MIRJANA Bonkovic. Classification accuracy comparison of asthmatic wheezing sounds recorded under ideal and real-world conditions. In *15th International Conference on Artificial Intelligence, Knowledge Engineering and Databases (AIKED 2016), Venice*, 2016.

[33] Renard Xaviero Adhi Pramono, Syed Anas Imtiaz, and Esther Rodriguez-Villegas. Evaluation of features for classification of wheezes and normal respiratory sounds. *PloS one*, 14(3):e0213659, 2019.

[34] T Richard Fenton, Hans Pasterkamp, A Tal, and Victor Chernick. Automated spectral characterization of wheezing in asthmatic children. *IEEE transactions on biomedical engineering*, (1):50–55, 1985.

[35] A Suzuki, C Sumi, K Nakayama, and M Mori. Real-time adaptive cancelling of ambient noise in lung sound measurement. *Medical and Biological Engineering and Computing*, 33:704–708, 1995.

[36] Gwo-Ching Chang and Yung-Fa Lai. Performance evaluation and enhancement of lung sound recognition system in two real noisy environments. *Computer methods and programs in biomedicine*, 97(2):141–150, 2010.

[37] Dimitra Emmanouilidou, Eric D McCollum, Daniel E Park, and Mounya Elhilali. Adaptive noise suppression of pediatric lung auscultations with real applications to noisy clinical settings in developing countries. *IEEE Transactions on Biomedical Engineering*, 62(9):2279–2288, 2015.

[38] Kirill Kochetov, Evgeny Putin, Svyatoslav Azizov, Ilya Skorobogatov, and Andrey Filchenkov. Wheeze detection using convolutional neural networks. In *Progress in Artificial Intelligence: 18th EPIA Conference on Artificial Intelligence, EPIA 2017, Porto, Portugal, September 5-8, 2017, Proceedings 18*, pages 162–173. Springer, 2017.

[39] Dinko Oletic and Vedran Bilas. Asthmatic wheeze detection from compressively sensed respiratory sound spectra. *IEEE journal of biomedical and health informatics*, 22(5):1406–1414, 2017.

[40] Sezer Ulukaya, Gorkem Serbes, and Yasemin P Kahya. Wheeze type classification using non-dyadic wavelet transform based optimal energy ratio technique. *Computers in biology and medicine*, 104:175–182, 2019.

[41] Amjad Hashemi, Hossein Arabalibiek, and Khosrow Agin. Classification of wheeze sounds using wavelets and neural networks. In *International conference on biomedical engineering and technology*, volume 11, pages 127–131. IACSIT Press Singapore, 2011.

[42] Nandini Sengupta, Md Sahidullah, and Goutam Saha. Lung sound classification using cepstral-based statistical features. *Computers in biology and medicine*, 75:118–129, 2016.

[43] E Andrès, R Gass, A Charloux, C Brandt, and A Hentzler. Respiratory sound analysis in the era of evidence-based medicine and the world of medicine 2.0. *Journal of medicine and life*, 11(2):89, 2018.

[44] Paolo Palange and Gernot Rohde. *ERS handbook of respiratory medicine*. European Respiratory Society, 2019.

[45] Kara Rogers et al. *The respiratory system*. Britannica Educational Publishing, 2010.

[46] Clara Mihaela Ionescu and Clara Mihaela Ionescu. The human respiratory system. *The Human Respiratory System: An Analysis of the Interplay between Anatomy, Structure, Breathing and Fractal Dynamics*, pages 13–22, 2013.

[47] Steven E Weinberger, Barbara A Cockrill, and Jess Mandel. *Principles of Pulmonary Medicine E-Book*. Elsevier Health Sciences, 2017.

[48] Cenk Balta and Mustafa Kuzucuoğlu. *Göğüs Cerrahisi Stajyer Kitabı*. Akademisyen Kitabevi, 2020.

[49] Michael G Levitzky. Pulmonary physiology. 2018.

[50] Aurora Hernando, Concepción Guillamas, Enrique Gutiérrez, Gloria Sánchez-Cascado, Luis Tordesillas, and Mª Jesús Méndez. *Técnicas básicas de enfermería. Novedad 2017*. Editex, 2017.

[51] Jiyuan Tu, Kiao Inthavong, Goodarz Ahmadi, Jiyuan Tu, Kiao Inthavong, and Goodarz Ahmadi. The human respiratory system. *Computational fluid and particle dynamics in the human respiratory system*, pages 19–44, 2013.

[52] Anoop Kumar Sinha and Sunit K Singh. Overview on anatomy of human respiratory system. In *Human Respiratory Viral Infections*, pages 3–15. CRC Press, 2014.

[53] Ian Peate. Anatomy and physiology, 10. the respiratory system. *British Journal of Healthcare Assistants*, 12(4):178–181, 2018.

[54] Hee Young Sohn, Sung Kyu Kim, and Ki Ho Kim. Anatomy of the respiratory system. *Tuberculosis and Respiratory Diseases*, 32(1):1–18, 1985.

[55] R Johnson and C Hsia. Anatomy and physiology of the human respiratory system. *Human Respiration: Anatomy and Physiology, Mathematical Modeling, Numerical Simulations and Applications*, pages 1–29, 2006.

[56] Vital signs (body temperature, pulse rate, respiration rate, blood pressure). https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjX0970usuCAxXvd6QEHWmIDM8QFnoECA8QAQ&url=https%3A%2F%2Fwww.hopkinsmedicine.org%2Fhealth%2Fconditions-and-diseases%2Fvital-signs-body-temperature-pulse-rate-respiration-rate-bloo usg=AOvVaw2qwmGbE1qacUpPg5ws8cz8&opi=89978449, 2022.

[57] Vital signs (body temperature, pulse rate, respiration rate, blood pressure). https://www.columbiadoctors.org/treatments-conditions/vital-signs-body-temperature-pulse-rate-respiration-rate-blood- 2022.

[58] World health organization. asthma. https://www.who.int/news-room/fact-sheets/detail/asthma, 2022.

[59] Eric D Bateman, Suzanne S Hurd, Peter J Barnes, Jean Bousquet, Jeffrey M Drazen, Mark FitzGerald, Peter Gibson, Ken Ohta, Paul O'Byrne, Soren Erik Pedersen, et al. Global strategy for asthma management and prevention: Gina executive summary. *European Respiratory Journal*, 31(1):143–178, 2008.

[60] What is copd? https://www.nhlbi.nih.gov/health/copd, 2022.

[61] William M Thurlbeck and NL Müller. Emphysema: definition, imaging, and quantification. *AJR. American journal of roentgenology*, 163(5):1017–1025, 1994.

[62] Thomas V Colby. Bronchiolitis: pathologic considerations. *American journal of clinical pathology*, 109(1):101–109, 1998.

[63] Todd A Florin, Amy C Plint, and Joseph J Zorc. Viral bronchiolitis. *The Lancet*, 389(10065):211–224, 2017.

[64] Talmadge E King. Bronchiolitis obliterans. *Lung*, 167:69–93, 1989.

[65] Jennifer K Quint, Elizabeth RC Millett, Miland Joshi, Vidya Navaratnam, Sara L Thomas, John R Hurst, Liam Smeeth, and Jeremy S Brown. Changes in the incidence, prevalence and mortality of bronchiectasis in the uk from 2004 to 2013: a population-based cohort study. *European Respiratory Journal*, 47(1):186–193, 2016.

[66] Montserrat Vendrell, Javier de Gracia, Casilda Olveira, Miguel Ángel Martínez, Rosa Girón, Luis Máiz, Rafael Cantón, Ramon Coll, Amparo Escribano, and Amparo Solé. Diagnosis and treatment of bronchiectasis. *Archivos de Bronconeumologia ((English Edition))*, 44(11):629–640, 2008.

[67] Pamela B Davis. Cystic fibrosis since 1938. *American journal of respiratory and critical care medicine*, 173(5):475–482, 2006.

[68] Harish Jasti, Eric M Mortensen, David Scott Obrosky, Wishwa N Kapoor, and Michael J Fine. Causes and risk factors for rehospitalization of patients hospitalized with community-acquired pneumonia. *Clinical infectious diseases*, 46(4):550–556, 2008.

[69] Maurice Demedts, AU Wells, JM Anto, U Costabel, R Hubbard, P Cullinan, H Slabbynck, G Rizzato, V Poletti, EK Verbeken, et al. Interstitial lung diseases: an epidemiological overview. *European Respiratory Journal*, 18(32 suppl):2s–16s, 2001.

[70] Norman C Staub. Pulmonary edema. *Physiological reviews*, 54(3):678–811, 1974.

[71] René TH Laennec. *De l'auscultation médiate: ou traité du diagnostic des maladies des poumons et du coeur*, volume 2. 1819.

[72] Littmann stethoscopes: A brief history. `https://www.littmann.in/3M/en_IN/littmann-stethoscopes-in/advantages/why-choose/dr-littmanns-legacy/`, 2022.

[73] Terry Ferns and Susan West. The art of auscultation: evaluating a patient's respiratory pathology. *British Journal of Nursing*, 17(12):772–777, 2008.

[74] Raymond LH Murphy. In defense of the stethoscope. *Respiratory care*, 53(3):355–369, 2008.

[75] Jessica S Coviello. *Auscultation skills: breath and heart sounds*. Lippincott Williams and Wilkins, 2013.

[76] Emmanuel Andrès, Amir Hajjam, and Christian Brandt. Advances and innovations in the field of auscultation, with a special focus on the development of new intelligent communicating stethoscope systems. *Health and Technology*, 2:5–16, 2012.

[77] J Proctor and E Rickards. How to perform chest auscultation and interpret the findings. *Nursing Times*, 116(1):23–26, 2020.

[78] Malay Sarkar, Irappa Madabhavi, Narasimhalu Niranjan, and Megha Dogra. Auscultation of the respiratory system. *Annals of thoracic medicine*, 10(3):158, 2015.

[79] Renata Báez Saldaña, Sergio Monraz Pérez, Patricia Castillo González, Uriel Rumbo Nava, Rogelio García Torrentera, Rebeca Ortiz Siordia, and Teresa I Fortoul Van Der Goes. La exploración del tórax: una guía para descifrar sus mensajes. *Revista de la Facultad de Medicina UNAM*, 59(6):43–57, 2016.

[80] LM Delaunois. Lung auscultation: back to basic medecine. *Swiss Medical Weekly*, 135(3536):511–512, 2005.

[81] Salvatore Mangione and Linda Z Nieman. Pulmonary auscultatory skills during training in internal medicine and family practice. *American journal of respiratory and critical care medicine*, 159(4):1119–1124, 1999.

[82] Steven McGee. *Evidence-based physical diagnosis e-book*. Elsevier Health Sciences, 2021.

[83] D Kumar, P d Carvalho, M Antunes, and J Henriques. Noise detection during heart sound recording. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3119–3123. IEEE, 2009.

[84] Shuang Leng, Ru San Tan, Kevin Tshun Chuan Chai, Chao Wang, Dhanjoo Ghista, and Liang Zhong. The electronic stethoscope. *Biomedical engineering online*, 14(1):1–37, 2015.

[85] Ipek Sen, Murat Saraclar, and Yasemin P Kahya. A comparison of svm and gmm-based classifier configurations for diagnostic classification of pulmonary sounds. *IEEE Transactions on Biomedical Engineering*, 62(7):1768–1776, 2015.

[86] A Torres-Jimenez, S Charleston-Villalobos, R Gonzalez-Camarena, G Chi-Lem, and T Aljama-Corrales. Asymmetry in lung sound intensities detected by respiratory acoustic thoracic imaging (rathi) and clinical pulmonary auscultation. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4797–4800. IEEE, 2008.

[87] Alison Marshall and Said Boussakta. Signal analysis of medical acoustic sounds with applications to chest medicine. *Journal of the Franklin Institute*, 344(3-4):230–242, 2007.

[88] Jeffrey J Goldberger and Jason Ng. *Practical signal and image processing in clinical cardiology*. Springer, 2010.

[89] Thinkslab stethoscopes. `https://www.thinklabs.com.`, 2022.

[90] Littmann stethoscopes. `https://www.littmann.in/3M/en_IN/littmann-stethoscopes-in/?utm_medium=redirect&utm_source=vanity-url&utm_campaign=www.littmann.in`, 2022.

[91] Eko stethoscopes. `https://www.ekohealth.com`, 2022.

[92] Ekuore stethoscopes. `https://ekuore.com/human-health/ekuore-pro-electronic-stethoscope/`, 2022.

[93] G John Gibson, Robert Loddenkemper, Yves Sibille, and Bo Lundbäck. *European lung white book*. European Respiratory Society, 2013.

[94] Manuel Lozano, José Antonio Fiz, and Raimon Jané. Automatic differentiation of normal and continuous adventitious respiratory sounds using ensemble empirical mode decomposition and instantaneous frequency. *IEEE journal of biomedical and health informatics*, 20(2):486–497, 2015.

[95] Antonio José Salazar, Catalina Alvarado, and Fernando Enrique Lozano. System of heart and lung sounds separation for store-and-forward telemedicine applications. *Revista Facultad de Ingeniería Universidad de Antioquia*, (64):175–181, 2012.

[96] Adam Rao, Emily Huynh, Thomas J Royston, Aaron Kornblith, and Shuvo Roy. Acoustic methods for pulmonary diagnosis. *IEEE reviews in biomedical engineering*, 12:221–239, 2018.

[97] ARA Sovijarvi, F Dalmasso, J Vanderschoot, LP Malmberg, G Righini, and SAT Stoneman. Definition of terms for applications of respiratory sounds. *European Respiratory Review*, 10(77):597–610, 2000.

[98] Volker Gross, Anke Dittmar, Thomas Penzel, Frank Schuttler, and Peter Von Wichert. The relationship between normal lung sounds, age, and gender. *American journal of respiratory and critical care medicine*, 162(3):905–909, 2000.

[99] A Sovijärvi, J Vanderschoot, and J Earis. Standardization of computerized respiratory sound analysis. *Crit Care Med*, 156:974–987, 1997.

[100] FJ Canadas-Quesada, N Ruiz-Reyes, J Carabias-Orti, P Vera-Candeas, and J Fuertes-Garcia. A non-negative matrix factorization approach based on spectro-temporal clustering to extract heart sounds. *Applied Acoustics*, 125:7–19, 2017.

[101] Sonia Charleston-Villalobos, Luis Felipe Domínguez-Robert, Ramón Gonzalez-Camarena, and AT Aljama-Corrales. Heart sounds interference cancellation in lung sounds. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1694–1697. IEEE, 2006.

[102] S Debbal and F Bereksi-Reguig. Spectral analysis of the pcg signals. *Int. J. Bioeng*, 2, 2007.

[103] ChingShun Lin and Erwin Hasting. Blind source separation of heart and lung sounds based on nonnegative matrix factorization. In *2013 International Symposium on Intelligent Signal Processing and Communication Systems*, pages 731–736. IEEE, 2013.

[104] Hans Pasterkamp, Richard Fenton, Asher Tal, and Victor Chernick. Interference of cardiovascular sounds with phonopneumography in children. *American Review of Respiratory Disease*, 131(1):61–64, 1985.

[105] Azadeh Yadollahi and Zahra MK Moussavi. A robust method for heart sounds localization using lung sounds entropy. *IEEE transactions on biomedical engineering*, 53(3):497–502, 2006.

[106] Plamen Bokov, Bruno Mahut, Patrice Flaud, and Christophe Delclaux. Wheezing recognition algorithm using recordings of respiratory sounds at the mouth in a pediatric population. *Computers in biology and medicine*, 70:40–50, 2016.

[107] F Dalmay, MT Antonini, P Marquet, and R Menier. Acoustic properties of the normal chest. *European Respiratory Journal*, 8(10):1761–1769, 1995.

[108] S McGee. Auscultation of the lungs. *Evidence-based physical diagnosis (3 rd ed.). Elsevier Saunders, Philadelphia*, 2012.

[109] Sandra Reichert, Raymond Gass, Christian Brandt, and Emmanuel Andrès. Analysis of respiratory sounds: state of the art. *Clinical medicine. Circulatory, respiratory and pulmonary medicine*, 2:CCRPM–S530, 2008.

[110] Paul Forgacs. Lung sounds. *British journal of diseases of the chest*, 63(1):1–12, 1969.

[111] Harvey Fletcher and Wilden A Munson. Loudness, its definition, measurement and calculation. *Bell System Technical Journal*, 12(4):377–430, 1933.

[112] RJ Riella, P Nohama, and JM Maia. Method for automatic detection of wheezing in lung sounds. *Brazilian Journal of Medical and Biological Research*, 42:674–684, 2009.

[113] S Rietveld, Mireille Oud, and Edo Hans Dooijes. Classification of asthmatic breath sounds: preliminary results of the classifying capacity of human examiners versus artificial neural networks. *Computers and Biomedical Research*, 32(5):440–448, 1999.

[114] Lemuel R Waitman, Kevin P Clarkson, John A Barwise, and Paul H King. Representation and classification of breath sounds recorded in an intensive care setting using neural networks. *Journal of clinical monitoring and computing*, 16:95–105, 2000.

[115] Noam Gavriely, YORAM Palti, and Gideon Alroy. Spectral characteristics of normal breath sounds. *Journal of applied physiology*, 50(2):307–314, 1981.

[116] Abraham Bohadana, Gabriel Izbicki, and Steve S Kraman. Fundamentals of lung auscultation. *New England Journal of Medicine*, 370(8):744–751, 2014.

[117] Hiroshi Nakano, Makito Hayashi, Etsuko Ohshima, Nahoko Nishikata, and Toshimitsu Shinohara. Validation of a new system of tracheal sound analysis for the diagnosis of sleep apnea-hypopnea syndrome. *Sleep*, 27(5):951–957, 2004.

[118] Paul Forgacs, AR Nathoo, and HD Richardson. Breath sounds. *Thorax*, 26(3):288–295, 1971.

[119] SOVIJARVI ARA. Characteristics of breath sounds and adventitious respiratory sounds. *Eur Respir Rev*, 10:591–596, 2000.

[120] ARA Sovijärvi, J Vanderschoot, and JE Earis. *Computerized respiratory sound analysis (CORSA): recommended standards for terms and techniques: ERS task force report*. Munksgaard, 2000.

[121] Paul Forgacs. The functional basis of pulmonary sounds. *Chest*, 73(3):399–405, 1978.

[122] Robert Loudon and Raymond LH Murphy Jr. Lung sounds. *American Review of Respiratory Disease*, 130(4):663–673, 1984.

[123] American Thoracic Society et al. Updated nomenclature for membership reaction. *ATS NEWS*, 3:5–6, 1977.

[124] Paul Forgacs. Crackles and wheezes. *The Lancet*, 290(7508):203–205, 1967.

[125] JAMES B Grotberg and NOAM Gavriely. Flutter in collapsible tubes: a theoretical model of wheezes. *Journal of Applied Physiology*, 66(5):2262–2273, 1989.

[126] Abhishek Jain and Jithendra Vepa. Lung sound analysis for wheeze episode detection. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2582–2585. IEEE, 2008.

[127] Jeannette Hoevers and Robert G Loudon. Measuring crackles. *Chest*, 98(5):1240–1243, 1990.

[128] Sibghatullah I Khan and Vasif Ahmed. Classification of pulmonary crackles and pleural friction rubs using mfcc statistical parameters. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2437–2440. IEEE, 2016.

[129] Raymond LH Murphy Jr, Stephen K Holford, and William C Knowler. Visual lung-sound characterization by time-expanded wave-form analysis. *New England Journal of Medicine*, 296(17):968–971, 1977.

[130] M Yeginer and YP Kahya. Modeling of pulmonary crackles using wavelet networks. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 7560–7563. IEEE, 2006.

[131] Laura Vannuccini, Marcello Rossi, and Gabriele Pasquali. A new method to detect crackles in respiratory sounds. *Technology and Health Care*, 6(1):75–79, 1998.

[132] JE Earis, K Marsh, MG Pearson, and CM Ogilvie. The inspiratory" squawk" in extrinsic allergic alveolitis and other pulmonary fibroses. *Thorax*, 37(12):923–926, 1982.

[133] DUNCAN M GEDDES, BRYAN CORRIN, DA Brewerton, RJ Davies, and MARGARET TURNER-WARWICK. Progressive airway obliteration in adults and its association with rheumatoid disease. *QJM: An International Journal of Medicine*, 46(4):427–444, 1977.

[134] R Paciej, A Vyshedskiy, D Bana, and R Murphy. Squawks in pneumonia. *Thorax*, 59(2):177–178, 2004.

[135] Jerome H Saltzer and Michael D Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975.

[136] David L Donoho and Iain M Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association*, 90(432):1200–1224, 1995.

[137] Zhaoxue Chen, Xiwen Sun, and Shengdong Nie. An efficient method of automatic pulmonary parenchyma segmentation in ct images. In *2007 29th annual international conference of the IEEE Engineering in Medicine and Biology Society*, pages 5540–5542. IEEE, 2007.

[138] CS Poornimadevi and Helen Sulochana. Automatic detection of pulmonary tuberculosis using image processing techniques. In *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 798–802. IEEE, 2016.

[139] Jibi John and MG Mini. Multilevel thresholding based segmentation and feature extraction for pulmonary nodule detection. *Procedia Technology*, 24:957–963, 2016.

[140] Archana B Kanwade, Mohini P Sardey, Sarika A Panwar, Milind P Gajare, Monali N Chaudhari, and Kamal Upreti. Combined weighted feature extraction and deep learning approach for chronic obstructive pulmonary disease classification using electromyography. *International Journal of Information Technology*, pages 1–10, 2023.

[141] Neeraj Baghel, Vivek Nangia, and Malay Kishore Dutta. Alsd-net: Automatic lung sounds diagnosis network from pulmonary signals. *Neural Computing and Applications*, 33:17103–17118, 2021.

[142] Yoshitaka Kimori. A morphological image processing method to improve the visibility of pulmonary nodules on chest radiographic images. *Biomedical Signal Processing and Control*, 57:101744, 2020.

[143] Sezer Ulukaya, Gorkem Serbes, and Yasemin P Kahya. Performance comparison of wavelet based denoising methods on discontinuous adventitious lung sounds. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2928–2931. IEEE, 2017.

[144] Mrityunjay Kumar and Sarat Dass. A total variation-based algorithm for pixel-level image fusion. *IEEE Transactions on Image Processing*, 18(9):2137–2143, 2009.

[145] Dimitris G Manolakis, Vinay K Ingle, and Stephen M Kogon. *Spectral Estimation, Signal Modeling, Adaptive Filtering, and Array Processing*. ARTECH HOUSE: Boston, MA, USA, 2005.

[146] Béla Suki, Adriano M Alencar, URS Frey, Plamen Ch Ivanov, Sergey V Buldyrev, Arnab Majumdar, H Eugene Stanley, Christopher A Dawson, Gary S Krenz, and Michiaki Mishima. Fluctuations, noise and scaling in the cardiopulmonary system. *Fluctuation and Noise Letters*, 3(01):R1–R25, 2003.

[147] A Cortney Henderson, G Kim Prisk, David L Levin, Susan R Hopkins, and Richard B Buxton. Characterizing pulmonary blood flow distribution measured using arterial spin labeling. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In vivo*, 22(10):1025–1035, 2009.

[148] Rajkumar Palaniappan, Kenneth Sundaraj, Sebastian Sundaraj, N Huliraj, and SS Revadi. A telemedicine tool to detect pulmonary pathology using computerized pulmonary acoustic signal analysis. *Applied Soft Computing*, 37:952–959, 2015.

[149] Martin J Tobin, Gilbert Jenouri, Bonnie Lind, Herman Watson, Anne Schneider, and Marvin A Sackner. Validation of respiratory inductive plethysmography in patients with pulmonary disease. *Chest*, 83(4):615–620, 1983.

[150] Gorkem Serbes, C Okan Sakar, Yasemin P Kahya, and Nizamettin Aydin. Feature extraction using time-frequency/scale analysis and ensemble of feature sets for crackle detection. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3314–3317. IEEE, 2011.

[151] Zij er de Stichting. Geneeskundige stichting koningin elisabeth.

[152] Lawrence R Rabiner and Biing-Hwang Juang. Speech recognition: Statistical methods. *Encyclopedia of language and linguistics*, pages 1–18, 2006.

[153] Shengjun Zhou, Yuanzhi Cheng, and Shinichi Tamura. Automated lung segmentation and smoothing techniques for inclusion of juxtapleural nodules and pulmonary vessels on chest ct images. *Biomedical Signal Processing and Control*, 13:62–70, 2014.

[154] Omid Talakoub, Javad Alirezaie, and Paul Babyn. Lung segmentation in pulmonary ct images using wavelet transform. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 1, pages I–453. IEEE, 2007.

[155] Yousef Ebrahimdoost, Salah D Qanadli, Alireza Nikravanshalmani, Tim J Ellis, Zahra Falah Shojaee, and Jamshid Dehmeshki. Automatic segmentation of pulmonary artery (pa) in 3d pulmonary cta images. In *2011 17th International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2011.

[156] Eva M Van Rikxoort and Bram Van Ginneken. Automated segmentation of pulmonary structures in thoracic computed tomography scans: a review. *Physics in Medicine and Biology*, 58(17):R187, 2013.

[157] Felix JS Bragman, Jamie R McClelland, Joseph Jacob, John R Hurst, and David J Hawkes. Pulmonary lobe segmentation with probabilistic segmentation of the fissures and a groupwise fissure prior. *IEEE transactions on medical imaging*, 36(8):1650–1663, 2017.

[158] Gorkem Serbes, C Okan Sakar, Yasemin P Kahya, and Nizamettin Aydin. Pulmonary crackle detection using time–frequency and time–scale analysis. *Digital Signal Processing*, 23(3):1012–1021, 2013.

[159] Michael J Swain and Dana H Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991.

[160] Richard Lippmann. An introduction to computing with neural nets. *IEEE Assp magazine*, 4(2):4–22, 1987.

[161] Tessa A Sundaram, Brian B Avants, and James C Gee. Towards a dynamic model of pulmonary parenchymal deformation: Evaluation of methods for temporal reparameterization of lung data. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005: 8th International Conference, Palm Springs, CA, USA, October 26-29, 2005, Proceedings, Part II 8*, pages 328–335. Springer, 2005.

[162] Åsa M Wheelock and Craig E Wheelock. Trials and tribulations of 'omics data analysis: assessing quality of simca-based multivariate models using examples from pulmonary medicine. *Molecular BioSystems*, 9(11):2589–2596, 2013.

[163] Sacha P Salzberg, Mario L Lachat, Kai von Harbou, Gregor Zünd, and Marko I Turina. Normalization of high pulmonary vascular resistance with lvad support in heart transplantation candidates. *European journal of cardio-thoracic surgery*, 27(2):222–225, 2005.

[164] Jeffrey D Hosenpud, Thomas A Stibolt, Kamaljit Atwal, and David Shelley. Abnormal pulmonary function specifically related to congestive heart failure: comparison of patients before and after cardiac transplantation. *The American journal of medicine*, 88(5):493–496, 1990.

[165] Renfu Yin, Furong Tian, Birgit Frankenberger, Martin Hrabé de Angelis, and Tobias Stoeger. Selection and evaluation of stable housekeeping genes for gene expression normalization in carbon nanoparticle-induced acute pulmonary inflammation in mice. *Biochemical and Biophysical Research Communications*, 399(4):531–536, 2010.

[166] Jade B Lau Young, Geoffrey W Rodgers, Geoffrey M Shaw, and J Geoffrey Chase. Preliminary studies into acoustic sensing of lung recruitment during mechanical ventilation. *IFAC-PapersOnLine*, 48(20):141–146, 2015.

[167] Lijie Liu, Trac Tran, and Pankaj Topiwala. A new resampling approach for optimal reconstruction. In *Applications of Digital Image Processing XXXII*, volume 7443, pages 295–304. SPIE, 2009.

[168] Gary L Grunkemeier and YingXing Wu. Bootstrap resampling methods: something for nothing? *The Annals of thoracic surgery*, 77(4):1142–1144, 2004.

[169] Yaoyao Zhuo, Jie Shen, Yi Zhan, Ye Tian, Mingfeng Yu, Shuyi Yang, Peiyan Ye, Li Fan, Zhiyong Zhang, and Fei Shan. Optimization and validation of voxel size-related radiomics variability by combatting batch effect harmonization in pulmonary nodules: a phantom and clinical study. *Quantitative Imaging in Medicine and Surgery*, 13(9):6139, 2023.

[170] Ryan C Au, Wan C Tan, Jean Bourbeau, James C Hogg, and Miranda Kirby. Impact of image pre-processing methods on computed tomography radiomics features in chronic obstructive pulmonary disease. *Physics in Medicine and Biology*, 66(24):245015, 2021.

[171] Alessandro Brunelli, Marco Monteverde, Alessandro Borri, Michele Salati, Rita D Marasco, and Aroldo Fianchini. Predictors of prolonged air leak after pulmonary lobectomy. *The Annals of thoracic surgery*, 77(4):1205–1210, 2004.

[172] Pedro Faustino, Jorge Oliveira, and Miguel Coimbra. Crackle and wheeze detection in lung sound signals using convolutional neural networks. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 345–348. IEEE, 2021.

[173] Icbhi 2017 challenge. `https://bhichallenge.med.auth.gr.`, 2022.

[174] Jin Li and Andrew D Heap. A review of spatial interpolation methods for environmental scientists. 2008.

[175] Werner Stöber and Roger O McClellan. Pulmonary retention and clearance of inhaled biopersistent aerosol particles: Data-reducing interpolation models and models of physiologically based systems:-a review of recent progress and remaining problems. *Critical reviews in toxicology*, 27(6):539–598, 1997.

[176] Xiaochen Fan, Xin Xu, Jianxing Feng, Haixia Huang, Xiang Zuo, Guohou Xu, Guanghui Ma, Bin Chen, Jianbin Wu, Yinhua Huang, et al. Learnable interpolation and extrapolation network for fuzzy pulmonary lobe segmentation. *IET Image Processing*, 17(11):3258–3270, 2023.

[177] Seraina A Dual, Constance Verdonk, Myriam Amsallem, Jonathan Pham, Courtney Obasohan, Patrick Nataf, Doff B McElhinney, Alisa Arunamata, Tatiana Kuznetsova, Roham Zamanian, et al. Elucidating tricuspid doppler signal interpolation and its implication for assessing pulmonary hypertension. *Pulmonary Circulation*, 12(3):e12125, 2022.

[178] Rohit Kumar, Subrata Bhattacharya, and Govind Murmu. Exploring optimality of piecewise polynomial interpolation functions for lung field modeling in 2d chest x-ray images. *Frontiers in Physics*, 9:770752, 2021.

[179] Harpreet Kaur, Nam Pham, and Sergey Fomel. Seismic data interpolation using cyclegan. In *SEG technical program expanded abstracts 2019*, pages 2202–2206. Society of Exploration Geophysicists, 2019.

[180] Shixuan He, Wei Zhang, Lijuan Liu, Yu Huang, Jiming He, Wanyi Xie, Peng Wu, and Chunlei Du. Baseline correction for raman spectra using an improved asymmetric least squares method. *Analytical Methods*, 6(12):4402–4407, 2014.

[181] Alexander Chernobelsky, Oleg Shubayev, Cindy R Comeau, and Steven D Wolff. Baseline correction of phase contrast images improves quantification of blood

flow in the great vessels. *Journal of Cardiovascular Magnetic Resonance*, 9(4):681–685, 2007.

[182] Brett Ley, Williamson Z Bradford, Derek Weycker, Eric Vittinghoff, Roland M du Bois, and Harold R Collard. Unified baseline and longitudinal mortality prediction in idiopathic pulmonary fibrosis. *European Respiratory Journal*, 45(5):1374–1381, 2015.

[183] Stephen M Humphries, Kunihiro Yagihashi, Jason Huckleberry, Byung-Hak Rho, Joyce D Schroeder, Matthew Strand, Marvin I Schwarz, Kevin R Flaherty, Ella A Kazerooni, Edwin JR van Beek, et al. Idiopathic pulmonary fibrosis: data-driven textural analysis of extent of fibrosis at baseline and 15-month follow-up. *Radiology*, 285(1):270–278, 2017.

[184] Robert E Kelly Jr, Robert C Shamberger, Robert B Mellins, Karen K Mitchell, M Louise Lawson, Keith Oldham, Richard G Azizkhan, Andre V Hebra, Donald Nuss, Michael J Goretsky, et al. Prospective multicenter study of surgical correction of pectus excavatum: design, perioperative complications, pain, and baseline pulmonary function facilitated by internet-based data collection. *Journal of the American College of Surgeons*, 205(2):205–216, 2007.

[185] M Fraiwan, L Fraiwan, M Alkhodari, and O Hassanin. Recognition of pulmonary diseases from lung sounds using convolutional neural networks and long short-term memory. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–13, 2021.

[186] Cohen Leon. Time-frequency analysis: theory and applications. *USA: Pnentice Hall*, 1995.

[187] Jonathan Allen. Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(3):235–238, 1977.

[188] François Auger and Patrick Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on signal processing*, 43(5):1068–1089, 1995.

[189] Leon Cohen. Time-frequency distributions-a review. *Proceedings of the IEEE*, 77(7):941–981, 1989.

[190] Boualem Boashash. Estimating and interpreting the instantaneous frequency of a signal. i. fundamentals. *Proceedings of the IEEE*, 80(4):520–538, 1992.

[191] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

[192] Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, 41(7):909–996, 1988.

[193] Christopher Torrence and Gilbert P Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological society*, 79(1):61–78, 1998.

[194] Donald B Percival and Andrew T Walden. *Wavelet methods for time series analysis*, volume 4. Cambridge university press, 2000.

[195] Paul S Addison. *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance*. CRC press, 2017.

[196] Norden E Huang, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, 454(1971):903–995, 1998.

[197] Franz Hlawatsch and Gloria Faye Boudreaux-Bartels. Linear and quadratic time-frequency signal representations. *IEEE signal processing magazine*, 9(2):21–67, 1992.

[198] Dennis Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-part III: radio and communication engineering*, 93(26):429–441, 1946.

[199] Gilbert Strang and Truong Nguyen. *Wavelets and filter banks*. SIAM, 1996.

[200] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[201] Michael Unser. Texture classification and segmentation using wavelet frames. *IEEE Transactions on image processing*, 4(11):1549–1560, 1995.

[202] James L Flanagan and Roger M Golden. Phase vocoder. *Bell System Technical Journal*, 45(9):1493–1509, 1966.

[203] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[204] Gellért Sárosi, Mihály Mozsáry, Péter Mihajlik, and Tibor Fegyó. Comparison of feature extraction methods for speech recognition in noise-free and in traffic noise environment. In *2011 6th conference on speech technology and human-computer dialogue (SpeD)*, pages 1–8. IEEE, 2011.

[205] Vibha Tiwari. Mfcc and its applications in speaker recognition. *International journal on emerging technologies*, 1(1):19–22, 2010.

[206] Judith C Brown. Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

[207] Christian Schörkhuber and Anssi Klapuri. Constant-q transform toolbox for music processing. In *7th sound and music computing conference, Barcelona, Spain*, pages 3–64, 2010.

[208] Nicki Holighaus, Monika Dörfler, Gino Angelo Velasco, and Thomas Grill. A framework for invertible, real-time constant-q transforms. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(4):775–785, 2012.

[209] Zhicun Xu et al. Audio event classification using deep learning methods. 2018.

[210] Selina Chu, Shrikanth Narayanan, and C-C Jay Kuo. Environmental sound recognition with time–frequency audio features. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1142–1158, 2009.

[211] Nikša Jakovljević and Tatjana Lončar-Turukalo. Hidden markov model based respiratory sound classification. In *Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017*, pages 39–43. Springer, 2018.

[212] Kirill Kochetov, Evgeny Putin, Maksim Balashov, Andrey Filchenkov, and Anatoly Shalyto. Noise masking recurrent neural network for respiratory sound classification. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 208–217. Springer, 2018.

[213] Gaëtan Chambres, Pierre Hanna, and Myriam Desainte-Catherine. Automatic detection of patient with respiratory diseases using lung sound analysis. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2018.

[214] Yi Ma, Xinzi Xu, Qing Yu, Yuhang Zhang, Yongfu Li, Jian Zhao, and Guoxing Wang. Lungbrn: A smart digital stethoscope for detecting respiratory disease

using bi-resnet deep learning algorithm. In *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–4. IEEE, 2019.

[215] Koki Minami, Huimin Lu, Hyoungseop Kim, Shingo Mabu, Yasushi Hirano, and Shoji Kido. Automatic classification of large-scale respiratory sound dataset based on convolutional neural network. In *2019 19th International Conference on Control, Automation and Systems (ICCAS)*, pages 804–807. IEEE, 2019.

[216] Fatih Demir, Aras Masood Ismael, and Abdulkadir Sengur. Classification of lung sounds with cnn model using parallel pooling structure. *IEEE Access*, 8:105376–105383, 2020.

[217] Yi Ma, Xinzi Xu, and Yongfu Li. Lungrn+ nl: An improved adventitious lung sound classification using non-local block resnet neural network with mixup data augmentation. In *Interspeech*, pages 2902–2906, 2020.

[218] Zijiang Yang, Shuo Liu, Meishu Song, Emilia Parada-Cabaleiro, and Björn W Schuller. Adventitious respiratory classification using attentive residual neural networks. 2020.

[219] Naoki Asatani, Tohru Kamiya, Shingo Mabu, and Shoji Kido. Classification of respiratory sounds using improved convolutional recurrent neural network. *Computers and Electrical Engineering*, 94:107367, 2021.

[220] Sangita Das, Saurabh Pal, and Madhuchhanda Mitra. Acoustic feature based unsupervised approach of heart sound event detection. *Computers in Biology and Medicine*, 126:103990, 2020.

[221] Fatih Demir, Abdulkadir Sengur, and Varun Bajaj. Convolutional neural networks based efficient approach for classification of lung diseases. *Health information science and systems*, 8:1–8, 2020.

[222] Jiali Xie, Xavier Aubert, Xi Long, Johannes van Dijk, Bruno Arsenali, Pedro Fonseca, and Sebastiaan Overeem. Audio-based snore detection using deep neural networks. *Computer Methods and Programs in Biomedicine*, 200:105917, 2021.

[223] Elmar Messner, Melanie Fediuk, Paul Swatek, Stefan Scheidl, Freyja-Maria Smolle-Juttner, Horst Olschewski, and Franz Pernkopf. Crackle and breathing phase detection in lung sounds with deep bidirectional gated recurrent neural networks. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 356–359. IEEE, 2018.

[224] Diego Perna and Andrea Tagarelli. Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 50–55. IEEE, 2019.

[225] Jyotibdha Acharya and Arindam Basu. Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning. *IEEE transactions on biomedical circuits and systems*, 14(3):535–544, 2020.

[226] Hassen Chanane and Mohammed Bahoura. Convolutional neural network-based model for lung sounds classification. In *2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 555–558. IEEE, 2021.

[227] Siddhartha Gairola, Francis Tom, Nipun Kwatra, and Mohit Jain. Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 527–530. IEEE, 2021.

[228] Truc Nguyen and Franz Pernkopf. Lung sound classification using co-tuning and stochastic normalization. *IEEE Transactions on Biomedical Engineering*, 69(9):2872–2882, 2022.

[229] Krishna Mridha, Shakil Sarkar, and Dinesh Kumar. Respiratory disease classification by cnn using mfcc. In *2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA)*, pages 517–523. IEEE, 2021.

[230] Shing-Yun Jung, Chia-Hung Liao, Yu-Sheng Wu, Shyan-Ming Yuan, and Chuen-Tsai Sun. Efficiently classifying lung sounds through depthwise separable cnn models with fused stft and mfcc features. *Diagnostics*, 11(4):732, 2021.

[231] ASK Sreeram, Udhaya Ravishankar, Narayana Rao Sripada, and Baswaraj Mamidgi. Investigating the potential of mfcc features in classifying respiratory diseases. In *2020 7th International Conference on Internet of Things: Systems, Management and Security (IOTSMS)*, pages 1–7. IEEE, 2020.

[232] Murat Akcay. The effect of moderate altitude on tp-e interval, tp-e/qt, qt, cqt and p-wave dispersion. *Journal of Electrocardiology*, 51(6):929–933, 2018.

[233] S Jayalakshmy, Gnanou Florence Sudha, K Banupriya, and N Kavya. Comparison of various time-frequency analysis methods and classification of respiratory sounds using pre-trained googlenet classifier. In *Soft Computing and Signal Processing: Proceedings of 3rd ICSCSP 2020, Volume 1*, pages 455–465. Springer, 2021.

[234] Arka Roy and Udit Satija. A novel melspectrogram snippet representation learning framework for severity detection of chronic obstructive pulmonary diseases. *IEEE Transactions on Instrumentation and Measurement*, 72:1–11, 2023.

[235] Bin Gao, Wai Lok Woo, and LC Khor. Cochleagram-based audio pattern separation using two-dimensional non-negative matrix factorization with automatic sparsity adaptation. *The Journal of the Acoustical Society of America*, 135(3):1171–1185, 2014.

[236] JULIUS LEMPERT, ERNEST GLEN WEVER, and MERLE LAWRENCE. The cochleogram and its clinical application: a preliminary report. *Archives of otolaryngology*, 45(1):61–67, 1947.

[237] A Lev and H_ Sohmer. Sources of averaged neural responses recorded in animal and human subjects during cochlear audiometry (electro-cochleogram). *Archiv für klinische und experimentelle Ohren-, Nasen-und Kehlkopfheilkunde*, 201:79–90, 1972.

[238] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

[239] Barry De Ville. Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(6):448–455, 2013.

[240] Alexey Ya Chervonenkis. Early history of support vector machines. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 13–20, 2013.

[241] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[242] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[243] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[244] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[245] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[246] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

[247] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49, 1999.

[248] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3:127–163, 2000.

[249] David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer, 1998.

[250] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001.

[251] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[252] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[253] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[254] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[255] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam,*

The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pages 21–37. Springer, 2016.

[256] Daniel Jurafsky and James H Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

[257] Christopher D Manning, Prabhakar Raghavan, and Hinriche Schütze. Xml retrieval. *Introduction to Information Retrieval*, 2008.

[258] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[259] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[260] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[261] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410, 2016.

[262] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[263] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[264] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[265] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[266] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[267] Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer, 2010.

[268] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295, 2001.

[269] Ruey-Shiang Guh. Integrating artificial intelligence into on-line statistical process control. *Quality and reliability engineering international*, 19(1):1–20, 2003.

[270] Michael Pecht. Prognostics and health management of electronics. *Encyclopedia of structural health monitoring*, 2009.

[271] Alan Agresti. *Categorical data analysis*, volume 792. John Wiley and Sons, 2012.

[272] Patrick Royston, Gareth Ambler, and Willi Sauerbrei. The use of fractional polynomials to model continuous risk variables in epidemiology. *International journal of epidemiology*, 28(5):964–974, 1999.

[273] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[274] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[275] D Richard Cutler, Thomas C Edwards Jr, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson, and Joshua J Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.

[276] Anantha M Prasad, Louis R Iverson, and Andy Liaw. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9:181–199, 2006.

[277] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[278] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

[279] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[280] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[281] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[282] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[283] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[284] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[285] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[286] Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*, 2014.

[287] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016.

[288] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[289] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 196–202. Springer, 1992.

[290] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[291] The rale respository. `http://www.rale.ca`, 2022.

[292] Mohammad Fraiwan, Luay Fraiwan, Basheer Khassawneh, and Ali Ibnian. A dataset of lung sounds recorded from the chest wall using an electronic stethoscope. *Data in Brief*, 35:106913, 2021.

[293] Respiratorydatabase@ tr. `https://data.mendeley.com/datasets/p9z4h98s6j/1`, 2022.

[294] Thinklabs one lung sounds library. `https://www.thinklabs.com/sound-library`, 2022.

[295] East tennessee state university. pulmonary breath sounds. `https://faculty.etsu.edu/arnall/www/public_html/heartlung/breathsounds/contents.html`, 2022.

[296] J Racineux. *L'auscultation à L'écoute du Poumon ASTRA*. 1994.

[297] J.S Coviello. *Auscultation Skills: Breath Heart Sounds*. Lippincott Williams Wilkins, 2013.

[298] R. Wilkins, J. Hodgkin, and B. Lopez. *Fundamentals of Lung and Heart Sounds*. CV Mosby: Maryland Heights, 2004.

[299] S Lehrer. *Understanding Lung Sounds*. Saunders: Philadelphia, 2018.

[300] Mohammed Bahoura. *Analyse des signaux acoustiques respiratoires: contribution à la detection automatique des sibilants par paquets d'ondelettes*. PhD thesis, Rouen, 1999.

[301] Chiu-Han Hsiao, Ting-Wei Lin, Chii-Wann Lin, Fu-Shun Hsu, Frank Yeong-Sung Lin, Chung-Wei Chen, and Chi-Ming Chung. Breathing sound segmentation and detection using transfer learning techniques on an attention-based encoder-decoder architecture. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 754–759. IEEE, 2020.

[302] AA Grinchenko, VT Makarenkov, and AP Makarenkova. Kompjuternaya auskultaciya-novij metod objektivizacii haraktierictik zvykov dihaniya [computer auscultation is a new method of objectifying the lung sounds characteristics]. *Klin. Inform. i telemeditsina*, 6(7):31–36, 2010.

[303] Elmar Messner, Melanie Fediuk, Paul Swatek, Stefan Scheidl, Freyja-Maria Smolle-Jüttner, Horst Olschewski, and Franz Pernkopf. Multi-channel lung sound classification with convolutional recurrent neural networks. *Computers in Biology and Medicine*, 122:103831, 2020.

[304] Georgios Petmezas, Grigorios-Aris Cheimariotis, Leandros Stefanopoulos, Bruno Rocha, Rui Pedro Paiva, Aggelos K Katsaggelos, and Nicos Maglaveras. Automated lung sound classification using a hybrid cnn-lstm network and focal loss function. *Sensors*, 22(3):1232, 2022.

[305] Conor Wall, Li Zhang, Yonghong Yu, Akshi Kumar, and Rong Gao. A deep ensemble neural network with attention mechanisms for lung abnormality classification using audio inputs. *Sensors*, 22(15):5566, 2022.

[306] Ali Mohammad Alqudah, Shoroq Qazan, and Yusra M Obeidat. Deep learning models for detecting respiratory pathologies from raw lung auscultation sounds. *Soft Computing*, 26(24):13405–13429, 2022.

[307] Murat Aykanat, Özkan Kılıç, Bahar Kurt, and Sevgi Saryal. Classification of lung sounds using convolutional neural networks. *EURASIP Journal on Image and Video Processing*, 2017(1):1–9, 2017.

[308] Dalal Bardou, Kun Zhang, and Sayed Mohammad Ahmad. Lung sounds classification using convolutional neural networks. *Artificial intelligence in medicine*, 88:58–69, 2018.

[309] Renyu Liu, Shengsheng Cai, Kexin Zhang, and Nan Hu. Detection of adventitious respiratory sounds based on convolutional neural network. In *2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, pages 298–303. IEEE, 2019.

[310] Dat Ngo, Lam Pham, Anh Nguyen, Ben Phan, Khoa Tran, and Truong Nguyen. Deep learning framework applied for predicting anomaly of respiratory sounds. In *2021 International Symposium on Electrical and Electronics Engineering (ISEE)*, pages 42–47. IEEE, 2021.

[311] Truc Nguyen and Franz Pernkopf. Lung sound classification using snapshot ensemble of convolutional neural networks. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 760–763. IEEE, 2020.

[312] Stavros Ntalampiras and Ilyas Potamitis. Automatic acoustic identification of respiratory diseases. *Evolving Systems*, 12:69–77, 2021.

[313] Abdelkader Nasreddine Belkacem, Sofia Ouhbi, Abderrahmane Lakas, Elhadj Benkhelifa, and Chao Chen. End-to-end ai-based point-of-care diagnosis system for classifying respiratory illnesses and early detection of covid-19: a theoretical framework. *Frontiers in Medicine*, 8:585578, 2021.

[314] Yoonjoo Kim, YunKyong Hyon, Sung Soo Jung, Sunju Lee, Geon Yoo, Chaeuk Chung, and Taeyoung Ha. Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning. *Scientific reports*, 11(1):17186, 2021.

[315] Wenjie Song, Jiqing Han, and Hongwei Song. Contrastive embeddind learning method for respiratory sound classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1275–1279. IEEE, 2021.

[316] Arpan Srivastava, Sonakshi Jain, Ryan Miranda, Shruti Patil, Sharnil Pandya, and Ketan Kotecha. Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease. *PeerJ Computer Science*, 7:e369, 2021.

[317] Zeenat Tariq, Sayed Khushal Shah, and Yugyung Lee. Feature-based fusion using cnn for lung and heart sound classification. *Sensors*, 22(4):1521, 2022.

[318] Youngjin Choi, Hoeryeon Choi, Hwayoung Lee, Sookyoung Lee, and Hongchul Lee. Lightweight skip connections with efficient feature stacking for respiratory sound classification. *Ieee Access*, 10:53027–53042, 2022.

[319] Ziping Zhao, Zhen Gong, Mingyue Niu, Jiali Ma, Haishuai Wang, Zixing Zhang, and Ya Li. Automatic respiratory sound classification via multi-branch temporal convolutional network. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9102–9106. IEEE, 2022.

[320] Jane Saldanha, Shaunak Chakraborty, Shruti Patil, Ketan Kotecha, Satish Kumar, and Anand Nayyar. Data augmentation using variational autoencoders for improvement of respiratory disease classification. *Plos one*, 17(8):e0266467, 2022.

[321] Han Sung Kim and Hong Seong Park. Ensemble learning model for classification of respiratory anomalies. *Journal of Electrical Engineering and Technology*, pages 1–8, 2023.

[322] Rafia Sharmin Alice, Laurent Wendling, and KC Santosh. 2d respiratory sound analysis to detect lung abnormalities. In *International Conference on Recent Trends in Image Processing and Pattern Recognition*, pages 46–58. Springer, 2022.

[323] Vipul Chudasama, Krina Bhikadiya, Sapan H Mankad, Ajaykumar Patel, and Maunil P Mistry. Voice based pathology detection from respiratory sounds using optimized classifiers. *International Journal of Computing and Digital Systems*, 13(1):327–339, 2023.

[324] Funda Cinyol, Uğur Baysal, Deniz Köksal, Elif Babaoğlu, and Sevinç Sarınç Ulaşlı. Incorporating support vector machine to the classification of respiratory sounds by convolutional neural network. *Biomedical Signal Processing and Control*, 79:104093, 2023.

[325] Arnon Cohen. Signal processing methods for upper airway and pulmonary dysfunction diagnosis. *IEEE Engineering in Medicine and Biology Magazine*, 9(1):72–75, 1990.

[326] G Charbonneau. Basic techniques for respiratory sound analysis. *Eur Respir Rev*, 10:625–635, 2000.

[327] Curtis Roads. *The computer music tutorial*. MIT press, 1996.

[328] Sylvia Schulz and Thorsten Herfet. Binaural source separation in non-ideal reverberant environments. In *Proceedings of 10th International Conference on Digital Audio Effects (DAFx-07), Bordeaux, France*, 2007.

[329] Roneel V Sharan and Tom J Moir. Acoustic event recognition using cochleagram image and convolutional neural networks. *Applied Acoustics*, 148:62–66, 2019.

[330] Haiying Wang, Jyotsna Wassan, and Huiru Zheng. Encyclopedia of bioinformatics and computational biology. 2019.

[331] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[332] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[333] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018, 2015.

[334] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

[335] José Neto, Nicksson Arrais, Tiago Vinuto, and João Lucena. Convolution-vision transformer for automatic lung sound classification. In *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, volume 1, pages 97–102. IEEE, 2022.

Appendices

Appendix A

Automatic Robust Crackle Detection and Localization Approach Using AR-Based Spectral Estimation and Support Vector Machine

---

Loredana Daria Mang, Julio José Carabias-Orti, Francisco Jesús Canadas-Quesada, Juan de la Torre-Cruz, Antonio Muñoz-Montoro, Pablo Revuelta-Sanz, Eilas Fernandez Combarro, Automatic Robust Crackle Detection and Localization Approach Using AR-Based Spectral Estimation and Support Vector Machine, in Applied Sciences, Appl. Sci. 2023, 13(19), 10683; https://doi.org/10.3390/app131910683.

- Status: Published
- Journal: Applied Sciences
- Special Issue: Pattern Recognition and Artificial Intelligence in Biomedical Signal Processing
- EISSN: 2076-3417
- Impact Factor: 2.7
- Quartile: Q2 - 42/90

Appendix B

Cochleogram-based adventitious sounds classification using convolutional neural networks

---

Loredana Daria Mang, Francisco Jesús Canadas-Quesada, Julio José Carabias-Orti, Elias Combarro, José Ranilla, Cochleogram-based adventitious sounds classification using convolutional neural networks in Biomedical Signal Processing and Control, https://doi.org/10.1016/j.bspc.2022.104555.

- Status: Published
- Journal: Biomedical Signal Processing and Control
- EISSN: 1746-8108
- Impact Factor: 5.1
- Quartile: Q1 - 28/117

Appendix C

Classification of Adventitious Sounds combining Cochleogram and Vision Transformers

---

Loredana Daria Mang, Francisco David Gonzalez Martínez, Damián Martinez Muñoz, Sebastián García Galán and Raquel Cortina, Classification of Adventitious Sounds combining Cochleogram and Vision Transformers in Sensors.

- Status: Under Review
- Journal: Biomedical Sensors
- Special Issue: Advanced Machine Intelligence for Biomedical Signal Processing
- EISSN: 1424-8220
- Impact Factor: 3.9
- Quartile: Q2 - 109/352