UNIVERSIDADE D
COIMBRA

Tiago Ferreira Fernandes

Deep Learning for Automatic Segmentation and Classification of Adventitious Respiratory Sounds

Dissertation in the context of the Master in Data Science and Engineering, advised by Professor Rui Pedro Pinto de Carvalho e Paiva and Prof. Paulo Fernando Pereira de Carvalho, presented to the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra.

September of 2022

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D
COIMBRA

DEPARTMENT OF INFORMATICS ENGINEERING

Tiago Ferreira Fernandes

Deep Learning for Automatic Segmentation and Classification of Adventitious Respiratory Sounds

Dissertation in the context of the Master in Data Science and Engineering, advised by Prof. Rui Pedro Pinto de Carvalho e Paiva and Prof. Paulo Fernando Pereira de Carvalho and presented to the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra.

September 2022

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D
COIMBRA

DEPARTAMENTO DE ENGENHARIA INFORMÁTICA

Tiago Ferreira Fernandes

Deep Learning para Segmentação e Classificação Automática de Sons Respiratórios Adventícios

Dissertação no âmbito do Mestrado em Engenharia e Ciência dos Dados, orientada pelo Professor Doutor Rui Pedro Pinto de Carvalho e Paiva e pelo Professor Doutor Paulo Fernando Pereira de Carvalho apresentada ao Departamento de Engenharia Informática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Setembro 2022

Acknowledgements

Firstly, I would like to thank my advisers Professor Rui Pedro Pinto de Carvalho e Paiva and Professor Paulo Fernando Pereira de Carvalho for presenting me with this project, availability and full support given throughout this work, with their positive feedback and constructive opinions that will help me to become a better Data Scientist and Engineer.

To the WELMO project (Wearable Electronics for Effective Lung Monitoring, H2020-825572), which offered the context for this work and funded my work with a research scholarship.

To the Centre for Informatics and Systems of the University of Coimbra (CISUC), in particular the Adaptive Computation (AC) research group, which provided logistic support for the conduction of this research work.

To my research team colleagues, Bruno Rocha and Diogo Pessoa, who help me develop this work successfully, and helped me solve all my problems, with constant support and availability.

To my friends and colleagues from University, whom I have learnt a lot with, personally and academically.

To my family, who gave me constant support and always trusted in me to conclude this degree with success.

And, lastly, to everyone else who helped me achieve this goal.

Abstract

Respiratory diseases are among the deadliest in the world. These pathologies are characterised by Adventitious Respiratory Sounds, such as wheezes and crackles, throughout the respiratory cycle.

In this thesis, a study was conducted regarding the applicability of Deep Learning (DL) with the aim of automatically classifying and segmenting these adventitious and normal respiratory sounds present in patients' respiratory breathing cycles, mostly wheezes and crackles. Since DL models require large and diverse datasets, three datasets were used: Respiratory Sound Database (RSD), a variation of the RSD not publicly available (RSD New Annotations), and the HF_Lung_V1 dataset. Several DL architectures such as Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (BiLSTM), and a combination of both (CNN-BiLSTM) are evaluated, as well as classical Machine Learning (ML) such as Linear Discriminant Analysis (LDA), Support Vector Machine with radial basis function (SVMrbf), and Random Undersampling Boosted Trees (RUSBoost), in order to evaluate and compare different ML approaches.

In the classification phase, the classical ML approaches described above served as baseline models to compare against the DL models (CNNs) and to start the adaptation of this kind of data. These models were replicated from previous work by our team with the three datasets mentioned above (F1-Score macro of 79.1% in the RSD, F1-Score macro of 68.8% in the RSD New Annotations, and F1-Score macro of 65.2% in the HF_Lung_V1), as well as a crossing between them to better understand the capability of these models to generalise, which proved not to be very successful given their differences in the annotations of the datasets (F1-Score macro of 39.5% trained with RSD and tested with HF_Lung_V1, and F1-Score macro of 38.8% trained with the HF_Lung_V1 and tested with RSD - 3-class problem). Also, stratification of the RSD using the same models was performed, in order to better understand which demographic category and recording device achieved better results (F1-Score macro of 81.8% with the AKGC417L microphone, F1-Score macro of 78.9% in Adults, F1-Score macro of 79.6% in Male subjects, F1-Score macro of 83.3% subjects with Normal body-mass index, and F1-Score macro of 85.3% in subjects with Non-Chronic diagnosis).

As for the segmentation phase, two approaches were developed: the first consisted on two CNNs to classify individual frames and, it was used as a baseline to compare with the second approach; and the second approach was a replication of one of the models from a cited article, the CNN-BiLSTM, which achieved better results than the first approach in RSD (F1-Score of 26.8% vs. F1-Score of 22.3% in crackles vs. normal sounds, and F1-Score of 26.5% vs. F1-Score of 41.0% in wheezes vs. normal sounds) and HF_Lung_V1 (F1-Score of 35.9% vs. F1-Score of 41.5% in crackles vs. normal sounds, and F1-Score of 26.0% vs. F1-Score of 42.1% in wheezes vs. normal sounds). A crossing between both datasets was performed to check the capability of these models to generalise, which proved to be not very successful given their differences in the annotations of the datasets. Also, a small stratification of the RSD with this last model was performed only using the recordings of the AKGC417L microphone (F1-Score of 14.7% in crackles

vs. normal sounds, and F1-Score of 41.6% in wheezes vs. normal sounds).

The proposed solutions for classification and segmentation allowed to advance to state-of-the-art on this problem, specially using the RSD, although the current approaches still need to be improved to permit its accurate use on real-world scenarios.

Keywords

Deep Learning, Machine Learning, Adventitious Respiratory Sounds, Classification, Segmentation

Resumo

As patologias do foro respiratório são das mais mortíferas causas de morte em todo o mundo. Estas patologias são caracterizadas pela existência de Sons Respiratórios Adventícios, como as sibilâncias e fervores, ao longo do ciclo respiratório.

Nesta tese, é realizado um estudo sobre a aplicabilidade de abordagens de Aprendizagem Profunda para classificar e segmentar estes Sons Respiratórios Normais e Adventícios presentes nos ciclos respiratórios dos pacientes, especialmente as sibilâncias e os fervores. Como os modelos de Aprendizagem Profunda necessitam de bases de dados grandes e variadas, três bases de dados foram usadas: *Respiratory Sound Database* (*RSD*), uma variação da *RSD* que não é pública (*RSD New Annotations*) e a base de dados *HF_Lung_V1*. Vários modelos de Aprendizagem Profunda como as *Convolutional Neural Network* (*CNN*), *Bidirectional Long Short-Term Memory* (*BiLSTM*) e ainda uma combinação de ambos (*CNN-BiLSTM*) foram desenvolvidos, assim como modelos de Aprendizagem Computacional clássicos como *Linear Discriminant Analysis* (*LDA*), *Support Vector Machine with radial basis function* (*SVMrbf*) e *Random Undersampling Boosted Trees* (*RUSBoost*), para avaliar e comparar diferentes modelos de Aprendizagem Computacional.

Durante a fase da classificação, os modelos de Aprendizagem Computacional clássicos descritos acima apenas foram testados e desenvolvidos para poder comparar com os modelos de Aprendizagem Profunda (*CNNs*) e para iniciar a adaptação a trabalhar com este tipo de dados. Estes modelos foram replicados de um trabalho prévio desenvolvido pela equipa com as três bases de dados acima mencionados (F1-Score macro de 79.1% na *RSD*, F1-Score macro de 68.8% na *RSD New Annotations*, e F1-Score macro de 65.2% na base de dados HF_Lung_V1), assim como o cruzamentos entre elas para compreender a capacidade destes modelos de generalizar, concluindo-se que não são bons nessa tarefa, dada a diferença nas anotações dos eventos nas base de dados (F1-Score macro de 39.5% treinado na *RSD* e testado em *HF_Lung_V1*, and F1-Score macro de 38.8% treinado em *HF_Lung_V1* e testado na *RSD* - problema a 3 classes). Para além disso, uma estratificação da *RSD* usando os mesmos modelos foi feita, para uma melhor compreensão de qual categoria demográfica e equipamento usado para gravar consegue melhores resultados (F1-Score macro de 81.8% com o microfone *AKGC417L*, F1-Score macro de 78.9% nos adultos, F1-Score macro de 79.6% nos homens, F1-Score macro de 83.3% nos pacientes com índice de massa corporal (*IMC*) normal, e F1-Score macro de 85.3% nos pacientes com doenças não crónicas).

Já durante a fase de segmentação, duas abordagens foram desenvolvidas: a primeira foram duas *CNN* para classificar *frames* individualmente e foi usado como base para poder comparar com a segunda abordagem; e a segunda abordagem foi a replicação de um dos modelos de um artigo citado, a *CNN-BiLSTM*, que conseguiu melhores resultados que a primeira abordagem na *RSD* (F1-Score de 26.8% vs. F1-Score de 22.3% nos fervores vs. sons normais, e F1-Score de 26.5% vs. F1-Score de 41.0% nas sibilâncias vs. sons normais) e na *HF_Lung_V1* (F1-Score de 35.9% vs. F1-Score de 41.5% nos fervores vs. sons normais, e F1-Score de 26.0% vs. F1-Score de 42.1% nas sibilâncias vs. sons normais). O cruzamento entre estas bases de dados também foi feito para compreender a capacidade de generalizar

destes modelos, concluindo-se que não são bons nessa tarefa, dada a diferença nas anotações dos eventos nas base de dados. Para concluir, uma pequena estratificação da *RSD* apenas com os ficheiros que usaram um microfone *AKGC417L* também foi feita usando este último modelo (F1-Score de 14.7% nos fervores vs. sons normais, e F1-Score de 41.6% nas sibilâncias vs. sons normais).

As soluções propostas para a classificação e segmentação permitiram avançar relativamente ao estado-de-arte sobre este problema, especialmente utilizando a *RSD*, embora as abordagens atuais ainda necessitem de ser melhoradas para permitir a sua utilização em cenários reais.

Palavras-Chave

Aprendizagem Profunda, Aprendizagem Computacional, Sons Respiratórios Adventícios, Classificação, Segmentação

Contents

Abbreviations

AI Artificial Intelligence.

ARS Adventitious Respiratory Sounds.

AUC Area Under the ROC Curve.

BHI International Conference on Biomedical and Health Informatics.

BiGRNN Bidirectional Gated RNN.

BiGRU Bidirectional Gated Recurrent Unit.

BiLSTM Bidirectional Long Short-Term Memory.

BMI Body-Mass Index.

CAS Continuous Adventitious Sound.

CNN Convolutional Neural Network.

COPD Chronic Obstructive Pulmonary Disease.

DAS Discontinuous Adventitious Sound.

DCT Discrete Cosine Transform.

DEI Department of Informatics Engineering.

DL Deep Learning.

FN False Negative.

FP False Positive.

GRU Gated Recurrent Unit.

JI Jaccard Index.

LDA Linear Discriminant Analysis.

LRTI Lower Respiratory Tract Infection.

LSTM Long Short-Term Memory.

MCC Matthews Correlation Coefficient.

MFCC Mel-Frequency Cepstral Coefficients.

ML Machine Learning.

MRMR Minimum Redundancy Maximum Relevance.

NLP Natural Language Processing.

OC Overlap Coefficient.

RD Respiratory Diseases.

RNN Recurrent Neural Network.

ROC Receiver Operating Characteristic.

RS Respiratory Sounds.

RUSBoost Random Undersampling Boosted Trees.

STFT Short-Time Fourier Transform.

SVM Support Vector Machine.

SVMrbf Support Vector Machine with radial basis function.

TN True Negative.

TP True Positive.

TT Train-Test.

URTI Upper Respiratory Tract Infection.

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, the goal is to give the context and motivation to this Master's thesis, as well as present its objectives, a small introduction of the approaches and methods investigated, and some results and contributions, and the outline of the document.

1.1 Context and Motivation

The number of deaths caused by respiratory diseases RD such as Chronic Obstructive Pulmonary Disease (COPD), Lower Respiratory Tract Infection (LRTI) and Trachea, Bronchus, and Lung Cancer is increasing every year since these are the third, fourth, and sixth biggest causes of death worldwide, respectively [1], as illustrated in Figure 1.1. The deaths caused by RD in 2019 in Portugal were 10.9% of all deaths [2], having a huge impact on the healthcare systems, along with the COVID-19 disease. Early diagnosis and routine monitoring of patients with respiratory problems are very important to prevent the development of more serious illnesses.

At the moment, since it is the cheapest option and the less intrusive method, physicians use stethoscopes to auscultate the patients and try to identify any respiratory disorder. The results obtained are not always very accurate, since they depend on the level of hearing ability and expertise of the physician, background noise, and the patients' movements. Along with that, to correctly assess a patient, continuous monitoring is necessary, as hearing for a few seconds might not be enough to conclude whether a patient produces specific abnormal sounds, since such a small sample might not be sufficient to detect those sounds. Besides, with the COVID-19 crisis, the stethoscopes can be a source of transmitting the virus [3]. This lack of accuracy regarding the auscultation can be surpassed by automating the processing of analysing the respiratory sounds, using Artificial Intelligence (AI).

In the past years, some devices that continuously record respiratory sounds have been developed to attempt to improve the quality of the assessments performed by physicians, even though that is a time-consuming and hard task. Some Ma-

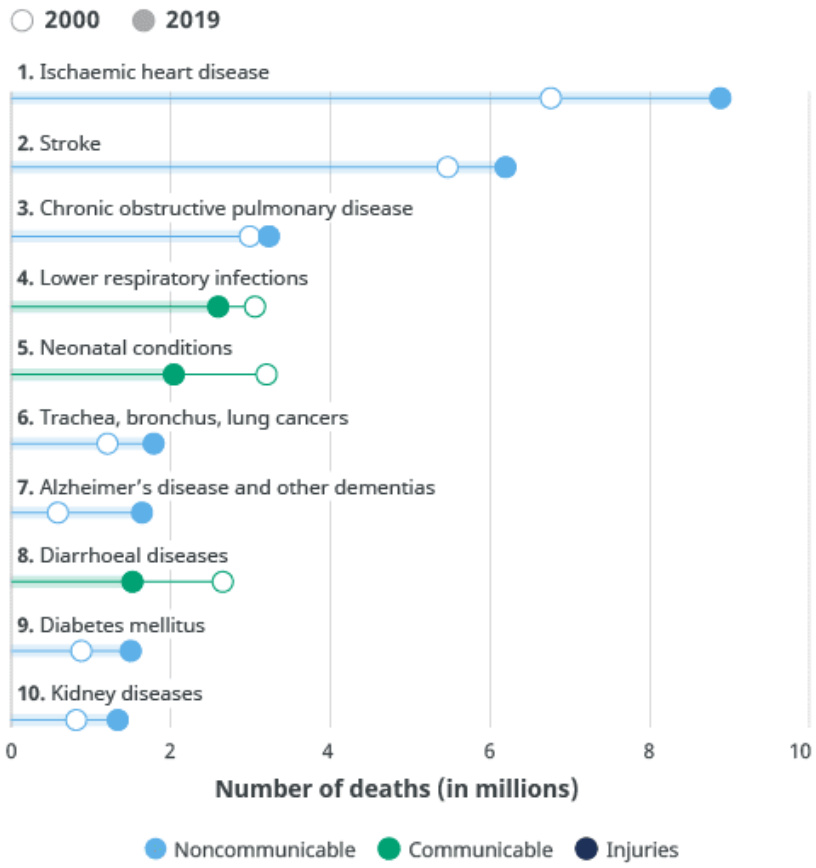


Figure 1.1: Top 10 deaths by disease worldwide [1]

chine Learning (ML) models have been developed to help in this process with some success. The growth of Deep Learning (DL) methods in the past years has proved to achieve better results in some tasks. Given the potential of DL, it may be possible to improve the quality of the classification and segmentation and also help health professionals.

There are two main motivations for this work. Firstly, the need to prevent the development of respiratory diseases: the sooner these respiratory pathologies are detected, the sooner the patients can start their treatment, which leads to a decrease in exacerbations, hospitalisations and mortality worldwide. Equally important is the relevance of the results obtained, by automatic sound classification and segmentation algorithms. If good results are obtained, it is a good starting point to create other DL models to segment and classify any type of sound.

1.2 Objectives and Approaches

This dissertation is involved in a European Project called Wearable Electronics for Effective Lung Monitoring (WELMO) [1], whose main objective is to provide continuous patient monitoring, for which automatic segmentation and classification of adventitious respiratory sounds ARS is a key requirement. WELMO is

[1]https://cordis.europa.eu/project/id/825572 (Accessed 2022-02-08)

divided into 2 parts: the creation of a wearable vest that can record respiratory sounds from various locations and the creation of models to segment and classify those sounds.

The overall plan for this WELMO Project, in a perfect scenario, is to help physicians to understand the possible respiratory diseases that their patients have as soon as possible and also allow the patients to start their treatments to give them a better life.

The main focus of this dissertation is the usage of DL approaches to classify and segment ARS of the real world (and not unrealistic sounds, e.g., overly simplistic sounds used in classes to teach students) and achieve the best results possible. The overall plan for this Master's Thesis is as follows:

- Critical analysis of the state of the art regarding the segmentation and classification of ARS
- Critical analysis of the state of the art regarding the usage of DL for segmentation and classification of sounds in general
- Research and development of DL approaches for segmentation and classification of ARS
- Validation of the models created

To address all the points mentioned above, the workload was divided into various phases. Firstly, it was necessary to research the main concepts like Respiratory Sounds (RS), ARS and its characteristics, and types of features and DL models used in this type of problem. All of this is in Chapter 2 with the explanation of some concepts necessary to better understand this thesis and Chapter 3 with a literature review to get the idea of what has been done in this field of study, and the approaches and models used.

To attain our objectives, one of the most important aspect to take into consideration is the dataset used: it needs to have a great amount of data and be as diverse as possible. With this in mind, there are two possible datasets that are going to be used: the Respiratory Sound Database (RSD), introduced in [4], [5], a re-annotated version of the same database (not available publicly and hereafter denoted RSD New Annotations) and HF_Lung_V1 ([6]). Both have some advantages and disadvantages, which are discussed in more detail in the beginning of Chapter 3, Section 3.1.

Regarding the classification task of this thesis, the main goal was to get used to working with this type of data and be able to fully understand what has been done by the research team. The reproduction of the models created in the paper [7] was the main work conducted. Those models had been trained and tested on the RSD. Hence, the sames models were trained and tested on the RSD New Annotations database. Also, the same models were tested with the HF_Lung_V1 dataset. In order to evaluate the quality of the models to generalise for any situation, the datasets were crossed, i.e., trained with one dataset and tested on

the other. The last experience concerning the classification task was a stratified analysis using the RSD, to try to understand if different population strata or recording devices alter the results significantly (e.g., the usage of a specific stethoscope/microphone to record the sounds). All of these methods, approaches and results are explained in detail in Chapter 4.

Concerning the segmentation task of this thesis, two approaches were performed. The first one was using individual frames to segment a file, while the second approach was a reproduction of one of the models proposed by the authors of the HF_Lung_V1 in [6]. In the first approach, the models used were developed by me, with some inspiration from what has been done previously when reproducing the DL models of [7]. Regarding the second approach, one of the models presented in that paper is reproduced. Also, a small stratification of the RSD with this last model was performed (only using the recordings of one of the microphones). All of these methods, experiments and results are explained in detail in Chapter 5.

1.3 Results and Contributions

With this thesis, some contributions were done to the scientific community, especially for the ones working in this field.

Regarding the classification of ARS, the stratification of the RSD was the most novel contribution, with an article accepted to be presented in the International Conference on Biomedical and Health Informatics (BHI) 2022[2], which had an acceptance rate of 32.7% in 2021. In this experiment, the models of [7] achieved higher results using specific recording devices (AKGC417L microphone is better overall), or a specific demographic characteristic can performed better (e.g., Male subjects).

Concerning the segmentation of ARS, the results obtained in the performed analysis on the RSD were superior to the state-of-the-art, using the complete dataset (best result in crackles vs. normal sounds with an F1-Score of 59.8%, and best result in wheezes vs. normal sounds with an F1-Score of 45.8%) and a portion of it (best result in crackles vs. normal sounds with an F1-Score of 66.4%, and best result in wheezes vs. normal sounds with an F1-Score of 58.6%).

1.4 Resources and Planning

All the experiments of this thesis were performed in a server based in Department of Informatics Engineering, accessed with a Virtual Private Network (VPN). The hardware of the server is the following:

- Intel Xeon(R) Silver 4214 CPU @ 2.20GHz, 48 cores

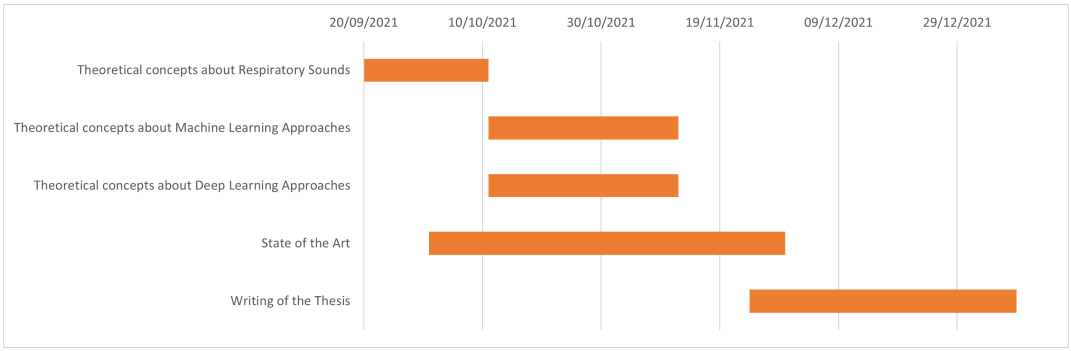
[2]https://bhi-bsn-2022.org/ (Accessed 2022-08-12)

- 256GB of RAM
- 3 NVIDIA Quadro P5000 GPUs, with 16GB of dedicated memory
- 3 NVIDIA RTX A5000 GPUs, with 24GB of dedicated memory (added later on)

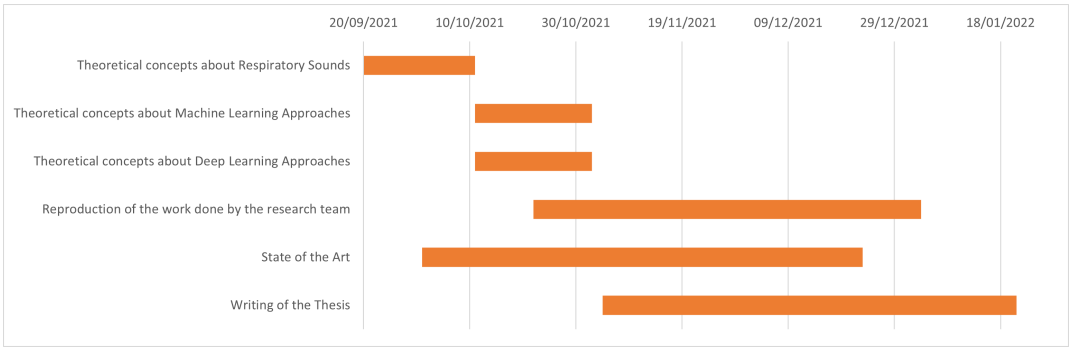
This server has 12 active users, all working on their Master's/PhD thesis, which led to a complicated resource management, meaning some experiments could have be improved (such as the last model reproduction).

Prior to the start of this work, it was previously planned by the advisers. This planning can be seen in Figure 1.2a

Figure 1.2b displays the work done in the 1st semester. It suffered some changes since it was the beginning of the work and then, it was not known its complexity and the workload of other courses. As so, some of the tasks were developed in the 2nd semester.



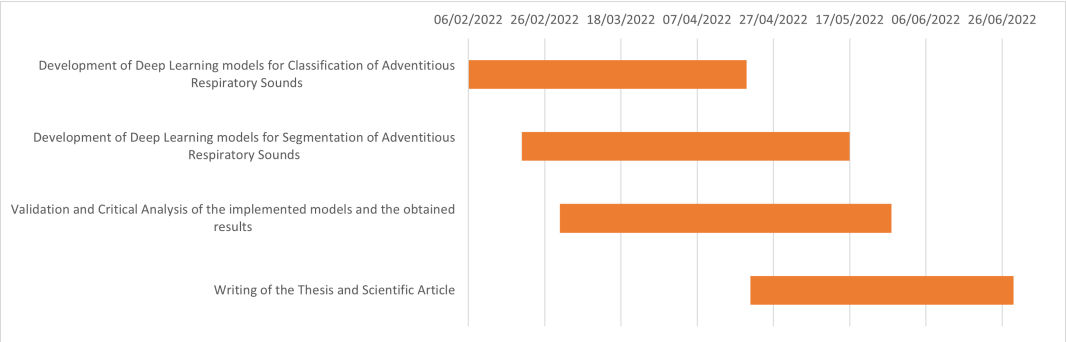
(a) Expected work for the 1st semester



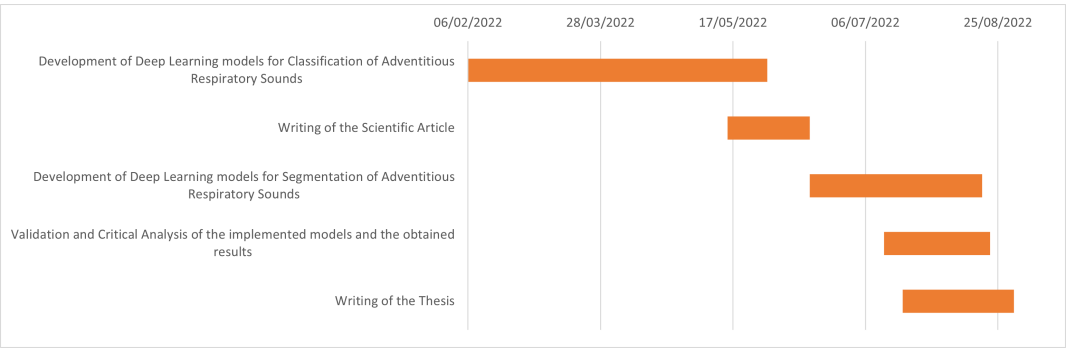
(b) Real work for the 1st semester

Figure 1.2: Working plan in the regarding the 1st semester

Figure 1.3a shows the expected work for the 2nd semester, after the first deadline. As we can see in that Figure, this thesis was planned to be finished until July 4th. Since an article was written and was not expected, as well as the frequent server overload, the final deadline was changed to September 5th, as we can see in Figure 1.3b.



(a) Expected work for the 2nd semester



(b) Real work for the 2nd semester

Figure 1.3: Working plan in the regarding the 2nd semester

1.5 Outline of the thesis

This dissertation is divided into 6 chapters:

- **Chapter 1** - This chapter consists of an introduction to this dissertation, as well as its motivations, objectives, approaches and workflow.
- **Chapter 2** - Some background concepts that are necessary to fully understand this work are presented in this chapter
- **Chapter 3** - This chapter presents a critical and comparative analysis of other works in this area (classification and segmentation), as well as a more detailed analysis of the most common datasets used in this field
- **Chapter 4** - The methods and experiments performed in regard to the classification task are explained in this chapter
- **Chapter 5** - This chapter explains the methods and experiments conducted in regard to the segmentation task
- **Chapter 6** - The conclusion of this thesis and some final comments about possible future work are addressed in the final chapter of the document

Chapter 2

Background Concepts

This chapter presents the principal and fundamental concepts to understand this work. Some concepts that are introduced here are regarding Respiratory Sounds, Machine Learning, and Deep Learning.

2.1 Respiratory Sounds

Respiratory Sounds are defined as being every sound related to breathing, including physiological sounds (normal), adventitious, coughing, etc. (Figure 2.1). Every RS is produced by air-flow in the respiratory tract, during the inspiration and expiration phases, and can be recorded in the thorax, trachea or mouth. Normal respiratory sounds are characterised as non-musical sounds with low-frequency that are provided by breathing and can be heard over the trachea and chest wall. Adventitious Respiratory Sounds are abnormal respiratory sounds that are superimposed on breathing sounds. These can be of 2 different types: continuous ARS (e.g., wheezes) or discontinuous ARS (e.g., crackles) [8]. Depending on their duration, intensity and location on the respiratory cycle, these can be related with some kind of respiratory problem. These two specific cases of continuous and discontinuous ARS are going to be the main focus of this work.

2.1.1 Discontinuous Adventitious Sounds

Crackles are caused by the sudden opening and closing of abnormally closed airways. They are explosive, non-musical and discontinuous. The frequency range of crackles is bounded by 60 Hz and 2 kHz, but most of their energy is concentrated between 60 Hz and 1.2 kHz. Usually, they last less than 20ms and can be classified as either fine[1] or coarse[2], depending on their duration and frequency range (short duration and high frequency vs. longer duration and low frequency, respectively). Depending on the characteristics of the crackles, they can be used

[1]https://www.youtube.com/watch?v=LHqqvrm2j6g (Accessed 2022-02-08)
[2]https://www.youtube.com/watch?v=aSor2XBc9K8 (Accessed 2022-02-08)

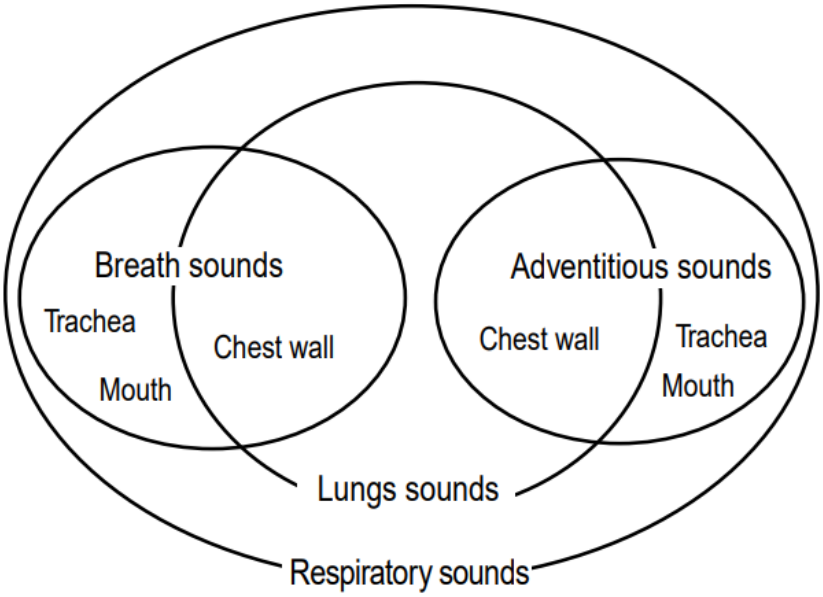


Figure 2.1: Relationship between the terms breath sounds, adventitious sounds, lung sounds and respiratory sounds [8]

to diagnose various types of lung diseases such as bronchiectasis or pneumonia [9].

2.1.2 Continuous Adventitious Sounds

Wheezes[3] are caused by an interaction between the airway wall and the gas moving through the airway, causing its oscillation. They are continuous and musical. The frequency range of wheezes is bounded by 100 Hz and 1000 Hz, or even higher if measured inside the airways. Usually, they last longer than 80-100ms. Wheezes can help diagnose various respiratory conditions such as COPD (in adults), and bronchiolitis (in children) [9]. There are other types of CAS like stridors and rhonchus. Figure 2.2 represents the spectrogram of normal breath sounds, wheezes and crackles.

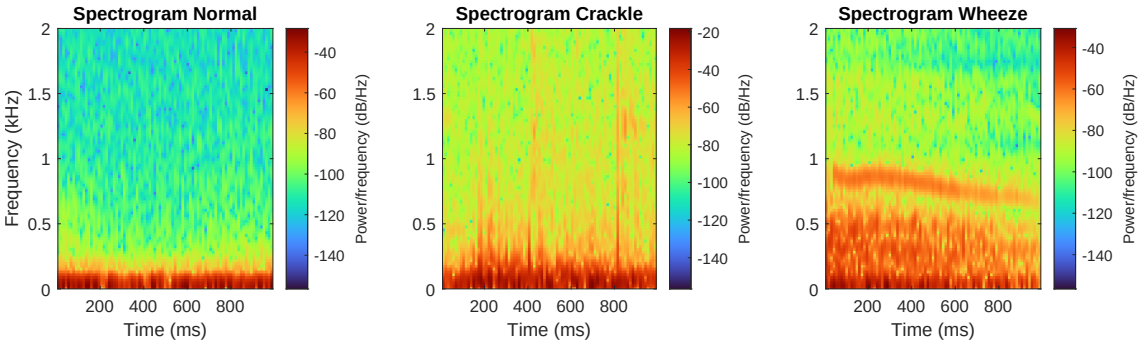


Figure 2.2: Spectrogram representation of the normal breath sounds, crackles and wheezes

[3]https://www.youtube.com/watch?v=T4qNgi4Vrvo (Accessed 2022-02-08)

2.2 Features

Machine Learning cannot use the raw data, since there usually is too much information or not useful information, so it is necessary to extract relevant features that are able to represent the underlying concepts to capture. A good feature set needs to have features with a strong correlation with the target and low correlation between them (i.e., not have redundant information). Since the feature engineering process is not the main focus of the thesis, but it is important to give some contexts about them, in the following sections, it is presented the features used on the paper [7], which were replicated in this thesis.

2.2.1 Spectral Features

Spectral features are based on the frequency of the signal, and are obtained by converting the time-based signal, using the Fourier Transform, such as spectral centroid, spread, skewness, kurtosis, entropy, zero-crossing rate, brightness, etc. [10]

2.2.2 Mel-Frequency Cepstral Coefficients Features

MFCC describe the spectral shape of the sound. These are calculated by converting the logarithm of the magnitude spectrum to the mel scale (which approximates the human auditory system's response more closely than the linearly-spaced frequency bands) and then computing the Discrete Cosine Transform (DCT). Since most of the signal information tends to be concentrated in a few low-frequency components of the DCT, it is typical to extract the first 13 components. [11]

2.2.3 Melodic Features

Melodic features are related to the pitch and describe the melody of the sounds. These can be computed from the pitch curve that caracterizes the frequencies of the signal. Pitch (fundamental frequency), voicing and inharmonicity are some examples of melodic features.

2.3 Deep Learning

Deep Learning is a subset of Machine Learning (ML) that imitates the way humans gain certain types of knowledge. Whilst classical ML leverage structured, labelled data to make predictions, i.e., require some type of pre-processing to organise it into a structured format, DL eliminates some of that data pre-processing, even though it can also have some data pre-processing. Also, DL algorithms are stacked in a hierarchy of increasing complexity and abstraction, and ML models

are easier to understand. It is what is behind various things that are present in our daily life, such as driverless cars (e.g., Tesla) or voice assistants (e.g., Alexa by Amazon) and they are trying to solve even more complex problems.

These models can achieve results above the state-of-the-art, sometimes surpassing human-level performance [12]. For that, they require large amounts of data (training examples). As these type of approaches are the main focus of the thesis, in the following subsections, specific DL models are will be explained in more detail.

2.3.1 Convolutional Neural Network

Convolutional Neural Network is the state-of-the-art approach for Computer Vision (CV) (recognition or classification), but it is also used for tasks involving one-dimension structures, like text (Natural Language Processing), time series analysis and, as in the case of this work, audio processing.

This model has 3 different types of layers:

- **Convolution layer** - its function is similar to a filter to detect features, since there is the dot-product between two tensors. Afterwards, it is usually applied an activation function (e.g., sigmoid, hyperbolic tangent, rectified linear unit - ReLU)

- **Pooling layer** - reduces the size of the input image to reduce the computational load, memory usage, the number of parameters and also reduces the possibility of overfit the model (because it reduces drastically the size of the input)

- **Fully Connected layer** - a set of neuron applies a linear transformation to the input vector (the output of the final layer but all its values are in a single vector) through a weight matrix (also known as feed-forward neural network)

Figure 2.3 shows an example of a CNN architecture.

2.3.2 Recurrent Neural Network

Recurrent Neural Network is a type of network that attempts to model time or sequence-dependent behaviour by feeding back the output of a neural network layer at a given instant in time to the input of the same network layer at the following instant. Figure 2.4 shows the structure of this model. It has many usages such as NLP in sentiment analysis or language translation, or time series analysis or even music generation.

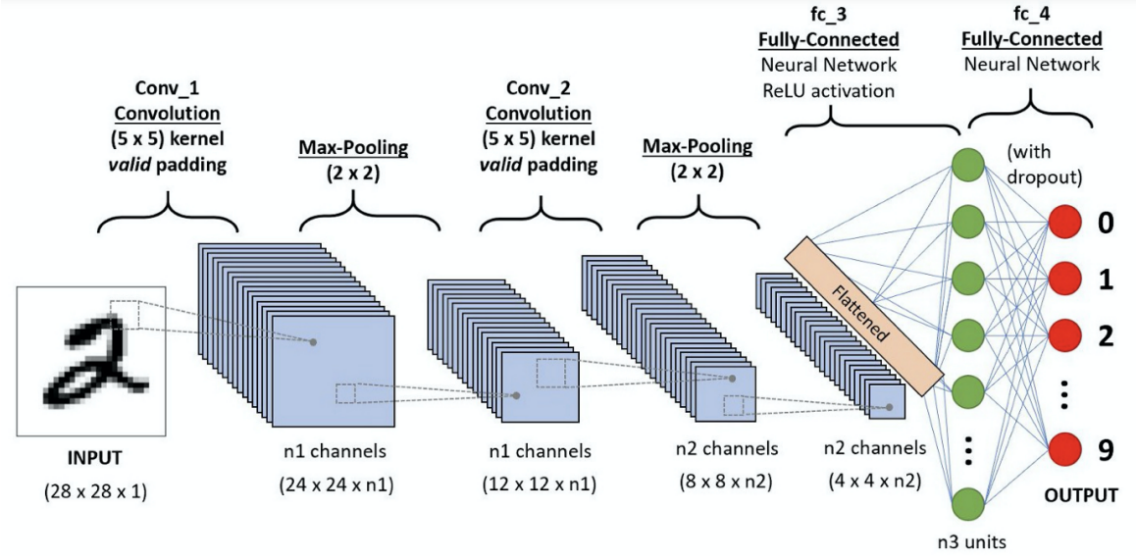


Figure 2.3: Example of CNN

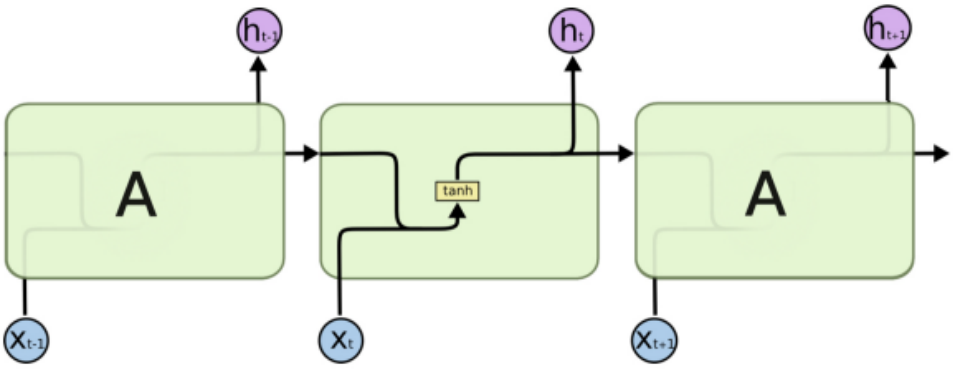


Figure 2.4: Example of RNN

2.3.3 Gated Recurrent Unit

GRU is a variant of the RNN which also takes the past into account, but it has the possibility/ability to neglect or not the past (Reset Gate) and identify/decide how important that past information is (Update Gate). In Figure 2.5, we can see in more detail the Recurrent Cell of the GRU model. There is also a variant of this model that does the same, but takes into account the past and the future. It is called BiGRU.

2.3.4 Long Short-Term Memory

This model is a variant of the RNN since it can retain "memory" for longer periods [13]. The core idea is that it is changed slowly, with only minor linear interactions. The main difference between LSTM and RNN is that the hidden layer has neurons with memory (Figure 2.6). There is also a variant of this model that takes into account the past and the future called Bidirectional Long Short-Term Memory (BiLSTM).

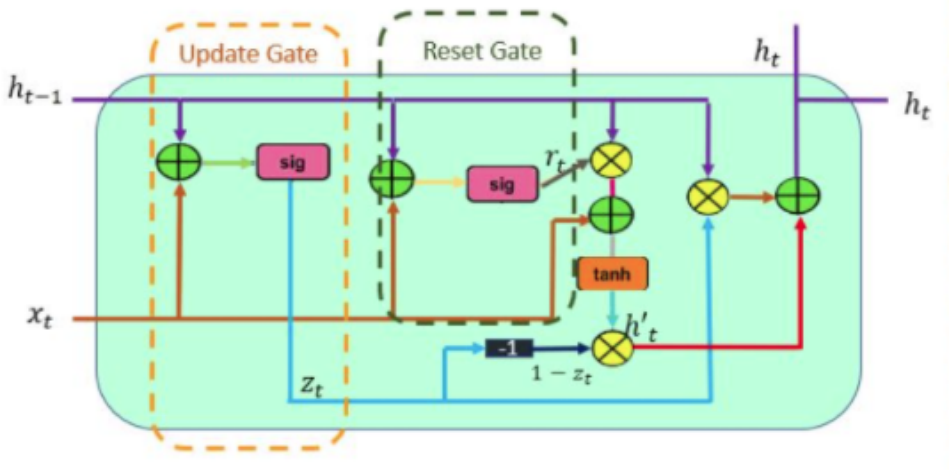


Figure 2.5: GRU Recurrent Cell

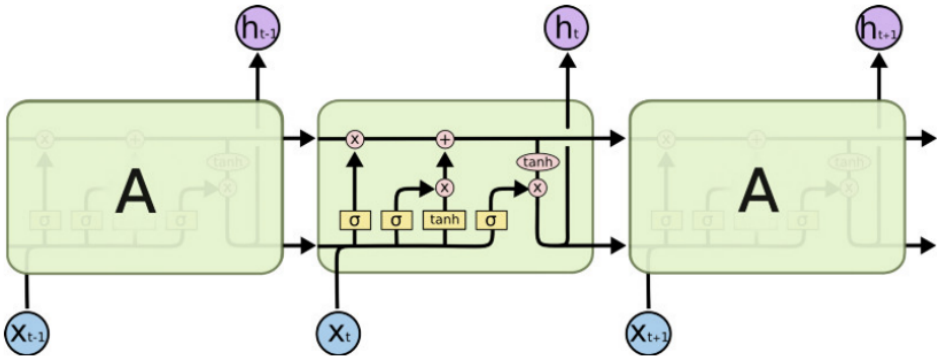


Figure 2.6: Example of LSTM

2.4 Evaluation Metrics

A confusion matrix is a table that allows us to evaluate the performance of the models by classes, in a simplified way. Table 2.1 is an example of a 2-class model. In this table, we can understand how many instances were predicted correctly or incorrectly. To do that, we need to define which class is the positive (the one which we want to take conclusions from) and which is the negative (the rest of the classes).

		True Class	
		Positive Class	*Negative Class*
Predicted Class	*Positive Class*	TP	FP
	Negative Class	FN	TN

Table 2.1: Confusion Matrix

Each sample has 4 possible types:

- **True Positive (TP)** - the sample was correctly predicted as belonging to the positive class

- **False Positive (FP)** - the sample is from the negative class and was incorrectly classified as being from the positive class
- **True Negative (TN)** - the sample was correctly predicted as belonging to the negative class
- **False Negative (FN)** - the sample is from the positive class and was incorrectly classified as being from the negative class

From this table, we can calculate other metrics that can assess better the performance of the model. The metrics chosen for evaluating the models are important since some are preferable in some cases than other (e.g., when the dataset is unbalanced). In the following subsections, the metrics that were used to assess the models are explained.

2.4.1 Accuracy

This metric is good to understand how many instances were correctly classified (positive and negative classes). To notice this metric is not that good for unbalanced dataset: if it has 900 samples of normal respiratory sounds and 100 samples of ARS and the model classifies every samples as being a normal respiratory sound, its accuracy is 90%, but its performance is not good.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

As this metric is not that good for unbalanced datasets, an alternative version of this metric was created, the Balanced Accuracy.

$$BalancedAccuracy = \frac{\sum Accuracyofeachclass}{Numberofclasses}$$

2.4.2 Precision

This metric helps us understand how many instances were predicted as being from the positive class were correctly classified

$$Precision = \frac{TP}{TP+FP}$$

2.4.3 Recall or Sensitivity

This metric assesses the number of instances from the positive class that were correctly classified.

$$RecallorSensitivity = \frac{TP}{TP+FN}$$

2.4.4 Specificity

This metric assesses the number of instances from the negative class that were correctly classified.

$$Specificity = \frac{TN}{TN+FP}$$

2.4.5 F1 Score

It is the harmonic mean between the Precision and Recall and because of that, it is one of the most used metrics. The higher this value is, the better the capability of the model to predict correctly.

$$F1Score = \frac{TP}{TP+\frac{FP+FN}{2}} = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$$

There is also the F1 Score Macro which calculates the average F1 Score of all classes in study.

$$F1ScoreMacro = \frac{\sum F1Score of each class}{Number of classes}$$

2.4.6 Matthews Correlation Coefficient

MCC takes into account all four values of the confusion matrix, and a high value (close to 1) means that both classes are predicted well, even if the dataset is unbalanced.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Similar to F1 Score Macro, there is also the MCC Macro, that averages the MCC of all classes in study

$$MCCMacro = \frac{\sum MCC of each class}{Number of classes}$$

2.4.7 Receiver Operating Characteristic Curve

This graph (Figure 2.7) shows the performance of a given model at all classification thresholds. This curve plots 2 parameters:

- True Positive Rate (TPR) (also known as Recall)

$$TPR = \frac{TP}{TP+FN}$$

- False Positive Rate (FPR)

$$FPR = \frac{FP}{FP+TN}$$

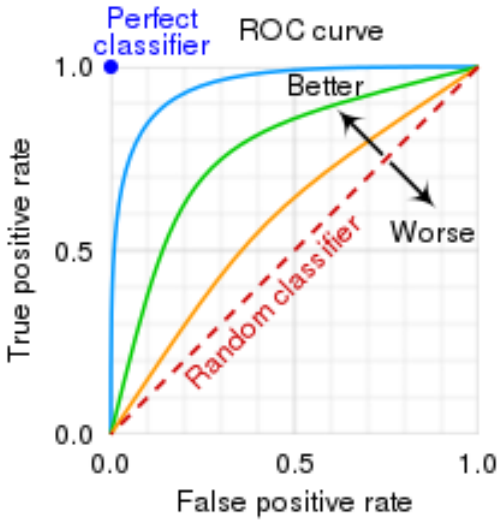


Figure 2.7: ROC Curve

2.4.8 Area Under the ROC Curve

This metric measures the area underneath the entire ROC curve and represents the degree of separability. The closer to one this value is, the better (Figure 2.8).

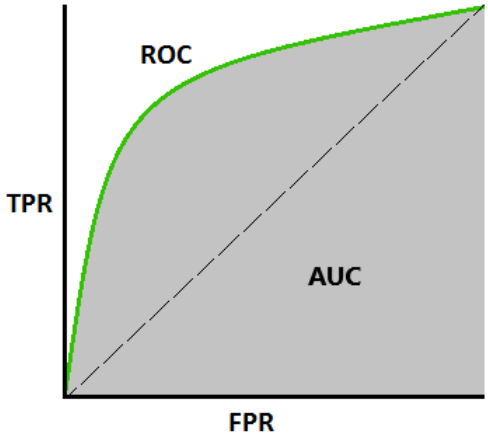


Figure 2.8: AUC

2.4.9 Jaccard Index

This metric helps us to understand how similar and diverse two samples sets are. Usually, it is used for segmentation problems.

$$JaccardIndex = \frac{|A \bigcap B|}{|A \bigcup B|}$$

2.4.10 Overlap Coefficient

This metric, as well as the Jaccard Index, is usually used in segmentation problems and it measures the overlap between 2 finite sets.

$$OverlapCoefficient = \frac{|A \bigcap B|}{min(|A|,|B|)}$$

Chapter 3

State of the Art

There are a lot of studies regarding the classification of RS and not so many regarding the segmentation of RS. In this State of the Art, some papers are described according to their results, approaches, and datasets used. All these articles were published in conferences and/or trusted sources. This chapter is divided in 4 sections: datasets, classification of RS using ML and DL approaches, segmentation of RS using ML and DL approaches, and their limitations.

3.1 Datasets

3.1.1 Respiratory Sound Database

In the articles [4] and [5], the authors explain how they created a dataset that can be adopted as the benchmark to this type of problems.
The dataset was created by two independent teams from two countries (Portugal and Greece). It contains 5.5 hours of respiratory sounds (920 annotated audio samples) of 126 patients, i.e., 6898 respiratory cycles, where 1864 have crackles, 886 have wheezes and 506 have both (more demographic information on Table 3.1). The authors created a specific was of splitting the data in Train-Test, in order to standardise the comparison between results among other studies that use this dataset. It was divided in Train-Test 60/40, with 539/381 audio files per set (mode information regarding that splitting on Table 3.2).

The team from School of Health Sciences, University of Aveiro, Portugal (ESSUA) obtained their sounds at Lab3R (Respiratory Research and Rehabilitation Laboratory) and at Hospital Infante D.Pedro. The sounds were collected from 7 chest locations: trachea, left and right anterior, posterior, and lateral. These respiratory sounds were collected from patients of all ages (infants, adults, and elderly patients) and with various complications, such as Upper Respiratory Tract Infection (URTI) and Lower Respiratory Tract Infection (LRTI), Chronic Obstructive Pulmonary Disease (COPD), asthma, pneumonia, bronchiectasis, and bronchiolitis. These sounds were recorded with a digital stethoscope (Welch Allyn Master Elite Plus Stethoscope Model 5079-400), or seven stethoscopes with a micro-

Number of recordings	920
Sampling frequency (number of recordings)	4 kHz (90); 10 kHz (6); 44.1 kHz (824)
Bits per sample	16
Average recording duration	21.5 s
Number of participants	126: 77 adults, 49 children
Sex	79 male, 46 female (NA: 1)
Age (mean ± standard deviation)	43.0 ± 32.2 years (NA: 1)
Age of adult participants	67.6 ± 11.6 years (NA: 1)
Age of child participants	4.8 ± 4.6 years
BMI of adult participants	227.2 ± 5.4 kg/m^2 (NA: 2)
Weight of child participants	21.4 ± 17.2 kg (NA: 5)
Height of child participants	104.7 ± 30.8 cm (NA: 7)

Table 3.1: Demographic Information of RSD (NA: Not Available)

Database	**Training Set**			**Testing Set**		
	ESSUA	AUTH	All	ESSUA	AUTH	All
Number of Patients	72	7	79	38	11	49
Number of Recordings	507	32	539	317	64	381
Number of Wheezes	459	42	501	588	61	649
Number of Crackles	1104	111	1215	273	112	385
Number of Crackles + Wheezes	335	28	363	106	37	143
Number of Normal	1740	323	2063	1216	363	1579

Table 3.2: Summary of the Training and Testing sets

phone in the main tube (3 M Littmann Classic II SE) or even seven air-coupled electret microphones (C 417 PP, AKG Acoustics) located into capsules made of Teflon. Two respiratory physiotherapists and one medical doctor with experience in visual-auditory crackle/wheeze recognition annotated the sounds by the presence/absence of adventitious sounds and identification of the breathing phases, but the time-consuming task of annotating the sounds was done by one respiratory physiotherapist. All the annotations performed by this team were conducted in the Computerised Lung Auscultation – Sound System (CLASS) proprietary tool.

The respiratory sounds collected by the team from the Aristotle University of Thessaloniki (AUTH) were acquired at the Papanikolaou General Hospital in Thessaloniki and at the General Hospital of Imathia (Health Unit of Naousa), Greece. The sounds were collected from 6 specific locations (Figure 3.1), from adult and elderly patients with COPD with comorbidities (heart failure, diabetes, and hypertension). The annotation process was performed by 3 experienced physicians (2 pulmonologists and one cardiologist) in Audacity 2.0.6. The annotations were the following: normal (respiratory sound), fine crackles, coarse crackles, wheezing, speech, cough, and artifact.

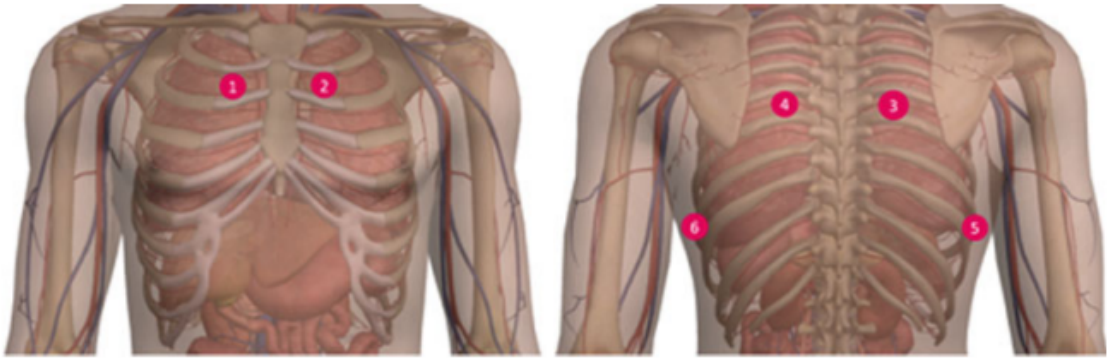


Figure 3.1: Chest Locations for the Recording of RS of RSD

3.1.2 HF_Lung_V1 Database

In [6], the authors explain how they created a dataset that can also be adopted as the benchmark to this type of problems.
The dataset was created using two sources: one was used in a datathon in Taiwan Smart Emergency and Critical Care (TSECC) that has lung sound recordings from 261 patients, while the other source has sound recordings of 18 residents of a Respiratory Care Ward (RCW) or a Respiratory Care Center (RCC) in Northern Taiwan (all under mechanical ventilation). From both sources, the patients are all Taiwanese and aged older than 20 years (more detailed information on Table 3.3).

	Subjects from RCW/RCC	**Subjects from TSECC**
Number	18	261
Gender (M/F)	11/7	NA
Age	67.5 (36.7, 98.3)	NA
Height (cm)	163.6 (147.2, 180.0)	NA
Weight (cm)	62.1 (38.2, 86.1)	NA
BMI (kg/m^2)	23.1 (15.6, 30.7)	NA
Respiratory Diseases		
Acute Respiratory Failure	4 (22.2%)	NA
Chronic Respiratory Failure	8 (44.4%)	NA
Acute Exacerbation of COPD	1 (5.6%)	NA
COPD	2 (11.1%)	NA
Pneumonia	4 (22.2%)	NA
Acute Respiratory Distress Syndrome	1 (5.6%)	NA
Emphysema	1 (5.6%)	NA
Comorbidities		
Chronic Kidney Disease	1 (5.6%)	NA
Acute Kidney Injury	3 (16.7%)	NA
Chronic Heart Failure	2 (11.1%)	NA
Diabetes Mellitus	7 (38.9%)	NA
Hypertension	6 (33.3%)	NA
Malignancy	1 (5.6%)	NA
Arrythmia	1 (5.6%)	NA
Cardiovascular Disease	1 (5.6%)	NA

Table 3.3: Demographic Information of HF_Lung_V1 (NA: Not Available)

For the recordings, they used 2 different devices: a commercial eletronic stetho-

scope (Littmann 3200) (TSECC and RCW/RCC) and a customized multichannel acoustic recording device (HF-Type-1 for short) that supports the connection of 8 eletrect microphones (RCW/RCC)). The 2 devices had a sampling rate of 4000 Hz, 16 bits per sample, recorded in the .wav format, and were collected from 8 specific locations (Figure 3.2).

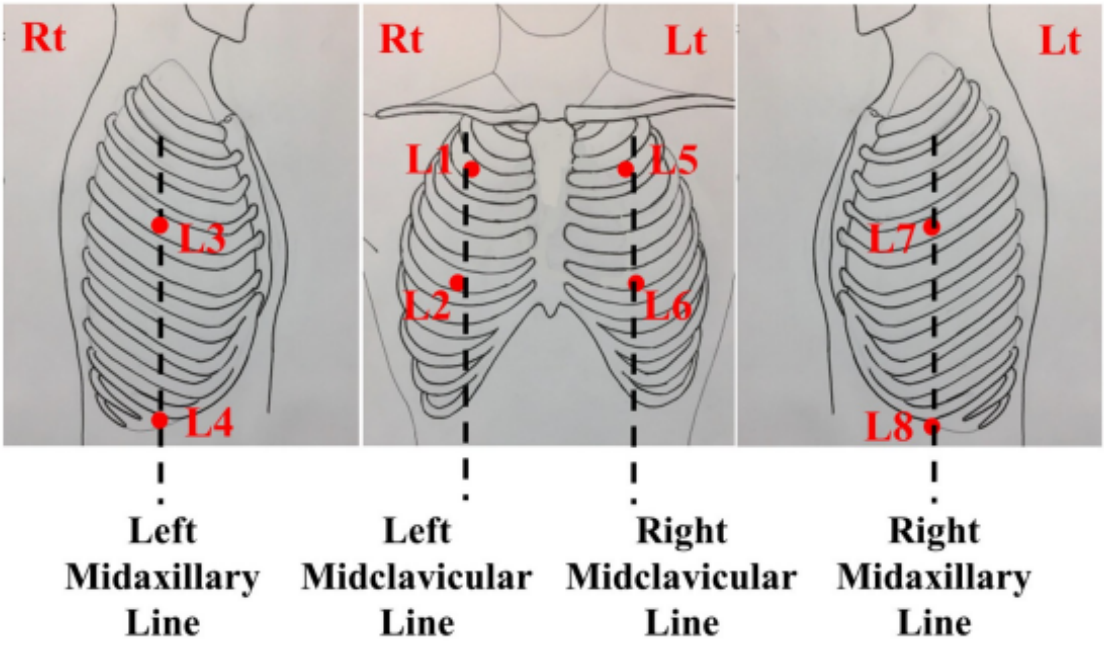


Figure 3.2: Chest Locations for the Recording of RS of HF_Lung_V1

Regarding the labeling process, the data in the TSECC dataset only had labels indicating whether a CAS or DAS existed, so 2 board-certified respiratory therapists and one board-certified nurse were recruited to label the start and end points of inhalation, exhalation, wheeze, stridor, rhonchus, and DAS, using a customized labeling software. Regarding the labeling process of the data in the RCW/RCC dataset, there is no information.

In this dataset, the standard audio duration used for inhalation, exhalation, and adventitious sound detection was 15 s because it contains at least 3 complete breath cycles, which are adequate for a clinician to reach a clinical conclusion. In total, this dataset has 9765 recordings of 15 s (40 hours and 41 minutes), where there are 15606 DAS (crackles) and 13883 CAS (8457 wheezes, 686 stridors, and 4740 rhonchus). Similar to the RSD dataset, the authors also divided the data in Train-Test to standardise the comparison between results among other studies that use this dataset. It was divided in Train-Test 80/20, with 7809/1956 audio files per set (more information regarding that splitting on Table 3.4).

	Training Set	Testing Set	Total
Recordings			
Number of 15s Recordings	7809	1956	9765
Total Duration (min)	1952.25	489	2441.25
Labels			
Number of I	27223	6872	34095
Total Duration of I (min)	422.17	105.97	528.14
Mean Duration of I (s)	0.93	0.93	0.93
Number of E	15601	2748	18349
Total Duration of E (min)	248.05	44.81	292.85
Mean Duration of E (s)	0.95	0.98	0.96
Number of C/W/S/R	11464/7027/657/3780	2419/1430/29/960	13883/8457/686/4740
Total Duration of C/W/S/R (min)	160.16/100.71/9.10/50.53	31.01/19.02/0.36/11.63	191.16/119.73/9.46/61.98
Mean Duration of C/W/S/R (s)	0.84/0.86/0.83/0.80	0.77/0.80/0.74/0.73	0.83/0.85/0.83/0.78
Number of D	13794	1812	15606
Total Duration of D (min)	203.59	27.29	230.87
Mean Duration of D (s)	0.89	0.90	0.89

Table 3.4: Summary of the Training and Testing sets (I: Inhalation, E: Exhalation, W: Wheeze, S: Stridor, R: Rhonchus, C: CAS, D: DAS, S, and R were combined to form C)

3.1.3 Other relevant datasets

Norwegian Database of Health Conditions and Chronic Diseases (Tromsø 7)

In [14], it is referred the usage of a specific database, the Tromsø 7. This study is conducted every 6-7 years in the Tromsø municipality to assess the health conditions of the population of that municipality. In 2015-2016, 21083 participants attended for a first visit to assess their health conditions and from those, 6048 participants (mean age 63.2 and 54.7% female) had their lung sounds recorded. Those sounds were recorded using an electret microphone (MKE 2-eW Gold, Sennheiser electronic GmbH & Co. KG) inserted at the tube of a stethoscope, 10cm away from the chest piece. The files were captured in .wav format, at 44100 Hz and recorded at 6 chest locations (Figure 3.3), during 10-15s.

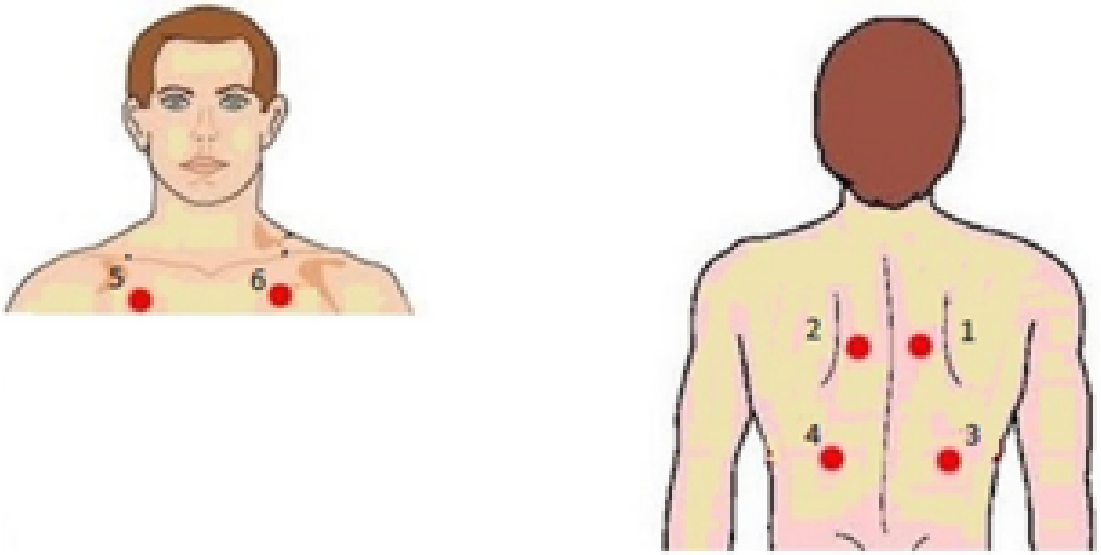


Figure 3.3: Chest Locations for the Recording of RS of Tromsø 7

Regarding the labeling process, in [14], they divided the dataset in 3 subsets, all annotated by different persons (physiotherapist/lung sound researcher, a computer scientist with previous experience in detection of wheezes in lung sounds, and a general practitioner and experienced lung sound researcher). These sounds were annotated by breathing phases (inhalation and exhalation). The first two subsets were used in the training part, while the last one was used in the test part. This dataset will not be used because it is private.

3.2 Classification of Respiratory Sounds

The study conducted on [7] the authors tested 4 machine learning models (LDA, SVMrbf, RUSBoost, and CNN).
The 3 first models were fed with features extracted from the spectrograms and some novel acoustic features (spectral features, mel-features, and melodic features - reaching a total of 2430 features, because for each feature and event, 5 statistical moments were calculated: mean, standard deviation, median, minimum value, and maximum value) and the last one was fed with the spectrogram and mel spectrogram images.
The dataset used in this study was the RSD but some random generated events were added in order to increase the realism of the challenge for the models (with a fixed duration of 50ms and 150ms or variable duration following a Burr distribution - lower than 100ms "otherCrack" or between 100ms and 2s "otherWheez"). They tried 3 approaches on the classification of the ARS: crackles vs. wheezes vs. others (3 classes), crackles vs. others, and wheezes vs. others. To start the tests, the authors tried to understand the performance training with fixed duration events and evaluating the results with a fixed duration and they were great (best model with 96.9% with 3 classes, 99.6% with crackles vs. others, and 98.6%). Since the results were almost perfect, they suspected that the models were learning the duration of the events instead of the characteristics of each type of sound. To test this, they created a script to generate random events (explained above) and trained the models with a fixed duration but the testing part was done with a variable duration. This time, the overall performance decreased (best model with 63.7% with 3 classes, 66.1% with crackles vs. others, and 64.1% with wheezes vs. others). The final experiment was done with training and testing data with variable duration and the results were in between the 2 experiments done before (best model with 81.8% with 3 classes, 87.4% with crackles vs. others, and 73.2% with wheezes vs. others).

In the article "Influence of Event Duration on Automatic Wheeze Classification" [15], the approach was to classify wheezes using Linear Discriminant Analysis (LDA), Linear SVM (SVMlin), Gaussian SVM (SVMrbf), boosted trees (Boost), and a CNN (using the spectrogram images as input).
The dataset used in this article is RSD, but some random generated events were added in order to increase the difficulty of the challenge for the models (with a fixed duration of 150ms or variable duration following a Burr distribution - between 100ms and 2s).
47 different features (usually used in music information research and wheeze

classification such as spectral features and mel features) were used in the models (except CNN) and for each feature, 5 statistical moments (mean, standard deviation, median, minimum value, and maximum value), which means in total were extracted 235 features.
The first 4 models were trained with 10 different seeds, the parameters were optimized on a validation set containing 25% of the training set and then tested on the testing set. After that, they were optimized using Bayes optimization. In the case of the CNN model, they created a based architecture and optimized the parameters using a grid search approach for convolution size, number of convolutional filters, dropout rate, pooling size, and size of the fully connected layer. They were trained with a maximum number of 15 epochs, a mini batch of 128, ADAM optimization (adaptative moment estimation), and 10% of the training set was used for validation during the training phase.
When the events had a fixed duration, the best model had a 96% of accuracy and when the events had a variable duration, the performance decreased and the accuracy of the best model was 67.8%.

In [16], the authors tried to classify between various diseases (normal, asthma, pneumonia, bronchiectasis (BRON), chronic obstructive pulmonary disease (COPD), and heart failure (HF)) using a CNN and bidirectional LSTM network (CNN + BDLSTM).
The dataset used in this article was the combination of two datasets: RSD (but only on 110 instead of all 126 patients) and some recordings gathered by the authors at King Abdullah University Hospital, Jordan University of Science and Technology, Irbid, Jordan. The second part of the dataset has 301 total recordings from 103 patients (62 males and 41 females, mean age 50), out of which 35 patients had no respiratory disease, while 68 of them suffered from one of those diseases mentioned above. The sounds were recorded using a single-channel electronic stethoscope (3M Littmann model 3200) placed on either upper, middle, or lower left/right chest wall locations, and collected and annotated by two professional thoracic clinicians.
For each signal, 3 preprocessing steps were applied, in the following order: a one-dimensional discrete wavelet transform function (maximal overlap discrete wavelet transform - MODWT), a rLOESS displacement removal, and a z-score normalization with mean value of zero and a standard deviation of one.
The CNN model has a Conv1D layer (3 in total), followed by a Batch Normalization layer and a ReLU, except for the last Conv1D layer which is followed first by a max-pooling layer. After the first two ReLU layers, 30% dropout was added to prevent overftting of the model. The BDLSTM model has a total of 200 hidden-units (100 units in each direction) and at the end, a 20% dropout to prevent the overfitting.
In the end, the authors obtained a 99.62% accuracy, 98.43% sensitivity, and 99.69% specificity.

Aykanat et al. [17] created 8 different models (4 SVMs and 4 CNNs) to classify between various ARS, but the relevant one is where they tried to classify between normal, rhonchus, squeak, stridor, wheeze, rales, bronchovesicular, friction rub, bronchial, absent, decreased, aggravation, or long expirium duration (basically, normal respiratory sounds vs. CAS vs. DAS).

The dataset used was created by them and three hospitals agreed to participate in their research (Ankara University, Yıldırım Beyazıt University, and Yıldırım Beyazıt Education and Research Hospital). The stethoscope and the recording software used were both built and developed by them. In the end, they recorded respiratory sounds on 1630 patients, from 11 positions from each patient, totalling 17930 audio clips with 10s each.
The features extracted from this dataset were mel frequency cepstral coefficient (MFCC) for the SVM model and spectrogram using Short-Time Fourier Transform (STFT) for the CNN model. Finally, the results obtained were 76% accuracy, 79% precision, and 74% recall for the CNN and 75% accuracy, 75% precision, and 99% recall for the SVM.

The study done in [18], Acharya et al. proposed a hybrid CNN-RNN model to perform classification of four different breathing sounds (normal, wheeze, crackle, and both wheeze and crackle).
The dataset used in this article is RSD and since it has various sampling frequency, every recording was resampled to 4000Hz. The authors also did some data augmentation techniques such as noise addiction, speed variation, random shifting, pitch shifting, etc.
Regarding the features, they used Mel-frequency spectrogram with a 60ms window size with 50% overlap.
Concerning the developed model, they divided it in 3 stages: the CNN part - it has a batch-normalization layer to start off, then convolutional (always followed by a ReLU activation function) and max-pooling layers; the BiLSTM part - with hyperbolic tangent activation function; and classification part - with a dropout layer of 50% to prevent overfitting, a fully-connected layer with 100 neurons, and a softmax layer. The model is trained with categorical crossentropy loss and Adam optimizer.
In the end, the results obtained were 48.63% and 84.14% for sensitivity and specificity (micro metrics), and 58.47%, 58.01%, and 57.91% for precision, recall and F1 Score, respectively (macro metrics).

The RNN models proposed in [19] are designed to classify four different breathing sounds (normal, wheeze, crackle, and both wheeze and crackle), using the RSD.
The preprocessing of the sound data is divided in 3 steps: the first is the segmentation of the audio recording in windows with variable size and overlaps; the second step is the extraction of Mel-Frequency Cepstral Coefficients (MFCCs) for each window; and the last step is feature normalization (either using Z-score normalization or Min-max normalization - noting that they have tested the best normalization method and Z-score achieved better results).
Regarding the RNN models, the authors tried four different models (LSTM, GRU, BiLSTM, and BiGRU), where the worst model overall was GRU and the best model overall was LSTM. In summary, the best model obtained a 84% specificity and 64% sensitivity.

3.3 Segmentation of Respiratory Sounds

In [20], the idea is to use an attention-based encoder-decoder since it is a better model to segment audio files rather than a deep learning model.
The database used is private and consists of 15 seconds of audio files from 22 patients (12 men and 10 women). Two types of stethoscopes were used: a digital stethoscope (Littmann 3200, 3M corp: sampling frequency 2000Hz) and an anti-noising microphone set (Accursound, Heroic Faith Medical Science Co., Ltd: sampling frequency 4000Hz). All of the audio files were annotated by experienced respiratory therapists or medical doctors to indicate the period of inspiration, expiration, and adventitious sound at the resolution of a sub-millisecond range.
The features extracted from the ARS were the spectrogram and the acoustic features of the encoder.
The encoder uses a ResNet (transfer learning approach with some variations - ResNet50, ResNet101, and ResNet152), which converts the spectrogram into a fixed form. The decoder uses a LSTM for sequence analysis and an attention mechanism for creating a weighted image so that the model can focus on specific parts of the spectrogram at each step of time. In the experiment part, they used a train-test approach (440 recordings for training and 49 recordings for testing) with a 10-fold cross validation.
In the end, the best model achieved 92% of accuracy and 90% and 93% F1-score for inspiration and expiration, respectively.

In the article [21], the authors developed a multi-label classification system with Bidirectional Gated RNN (BiGRNN) to segment (and classify) lung sounds in inspiration vs. expiration phases, and normal sound vs crackle sound.
The dataset used is small and private (10 healthy patients and 5 patients with idiopathic pulmonary fibrosis - IPF), where for each patient, 32 single-channel sounds of 30s were recorded, with 3 to 8 breathing cycles. The process of labelling the recordings was done by annotating the temporal onset and offset positions of the events inspiration, expiration and crackles.
Regarding the feature extraction process, they resampled the data to 16kHz, used a STFT with a window size of 32ms and 12ms overlap (i.e., frame-shift of 20ms). After, they extracted two types of features: Mel Frequency Cepstral Coefficients (MFCCs) (20 static coefficients, 20 delta coefficients, and 20 acceleration coefficients) and Spectrogram (257-bin log magnitude).
The model used was BiGRNN with two hidden layers with 100 neurons for the forward and backward layers, respectively. The output layer is divided in two softmax layers and the activation functions in the hidden layers are rectifier nonlinearities. The model was initialized with orthogonal weights, for optimizing the cross-entropy error, they used the ADAM function, and a droupout of 50% applied to the hidden layers. They used 5-fold cross validation with 100 epochs for the training process with early stop.
In the end, the F1-Score for inspiration, expiration and crackles was 87.0%, 84.6%, and 72.1%, respectively.

The objective of the investigation done on [14] was to segment lung sounds to

detect the breathing phases (inspiration and expiration) using a Faster R-CNN. The dataset used was already explained above in the section 3.1 (Tromsø 7).
The feature used to train the model was the spectrogram with a window size of 4096 samples, with an overlap of 3200 samples, and data points only lower than 2000Hz.
A Faster R-CNN is a model used in object detection and consists of two CNN: a Region Proposal Network (RPN) that is responsible for identifying potential objects and its bounding boxes (it used the convolutional layers form the ResNet101 architecture), and a classification network that takes the input image and classifies each segment (background, inspiration or expiration). Since the model creates bounding boxes in the spectrogram with a certain level of confidence, they pruned away the one with a confidence level below 50%. After, they needed to ensure there was no multiple detection for the same breathing phase (more than 50% overlap between them), so it was necessary to remove the one with the lowest confidence. The final step was to remove the overlaps that still existed, so they shrunk the bounding boxes in equal amounts until they no longer overlapped.
To evaluate the algorithm, they used two approaches: the first one calculates the percentage of agreement between each annotator and the model, using the Jaccard Index (JI), where if it was larger than 0.5 and the boxes were the same class, they had agreement (this means that this method was not concerned with the exact beginning and end of each cycle); and the second one calculates the Cohen's Kappa to understand the level of agreement, taking into account the beginning and end of each cycle. For the first evaluation method, the average agreement was 97% and 87% for inspiration and expiration, respectively. And for the second evaluation method, the average sensitivity was 97% and the average specificity was 84%.

The framework proposed in [6] is modular and designed to segment and predict (inhalation, exhalation and ARS) using DL models: LSTM, GRU, BiLSTM, BiGRU, CNN-LSTM, CNN-GRU, CNN-BiLSTM, and CNN-BiGRU. The framework is divided in 3 parts: preprocessing, DL-based modeling and postprocessing. The dataset used was already explained above in the section 3.1 (HF_Lung_V1 database).
For the preprocessing part, the authors started by resampling all the recordings to 4000 Hz, then applied a high-pass filter (at 80 Hz) to eliminate the heart sound noise and STFT with Hanning window with a size of 256 and 64 hop length, without additional zero-padding. After this process, a 15s sound signal is transformed into the corresponding spectrogram. Following the spectrogram extraction, more features are extracted: MFCCs (20 static coefficients, 20 delta coefficients, and 20 acceleration coefficients) and Energy Summation (four frequency bands). In the end, a 938 x 193 feature matrix combines all of those above. To conclude, a min-max normalization is performed on each feature (values between 0 and 1).
The DL models mentioned above can be divided in 2 categories: the ones that start with CNNs and the ones that do not. The usage of the CNN models can help extract abstract features first and then feed them to the RNN variants models. The output of the models that do not start with the CNN have an output of 938 x 1 and the others have an output of 469 x 1 (the "1" value in these outputs represent if they have inhalation, exhalation or ARS present on the time-segment or not).

Regarding the postprocessing phase, the output vector is analyzed where each index represents a segment. If the interval between x and $x+1$ segments is smaller than 0.5 s, the difference in frequency between their energy peaks is computed and if that difference was below 25 Hz, they were merged into a single event. Also, if the duration of an event was shorter than 0.05s, the event was deleted.
The metrics used in this paper are a bit different because they refer to the detection of segments, meaning they are not evaluated the same way. Firstly, in Figure 3.4, the authors show how they defined the TP, TN, FP, and FN for the segment detection and the usage of the Jaccard Index (JI) for event detection. Then, the metrics used those values explained above and for evaluating the performance of the models (segment detection), the F1 Score is used, while for the evaluation of the event detection, the ROC and AUC metrics are used, along with the F1 Score. For CAS, the best model for segment and event dectections was CNN-BiGRU with an F1 Score of 53.3% and 51.6%, respectively, while for DAS, the best model for segment detection was CNN-BiLSTM with an F1 Score of 71.2% and for event detection was BiGRU with an F1 Score of 71.4%.

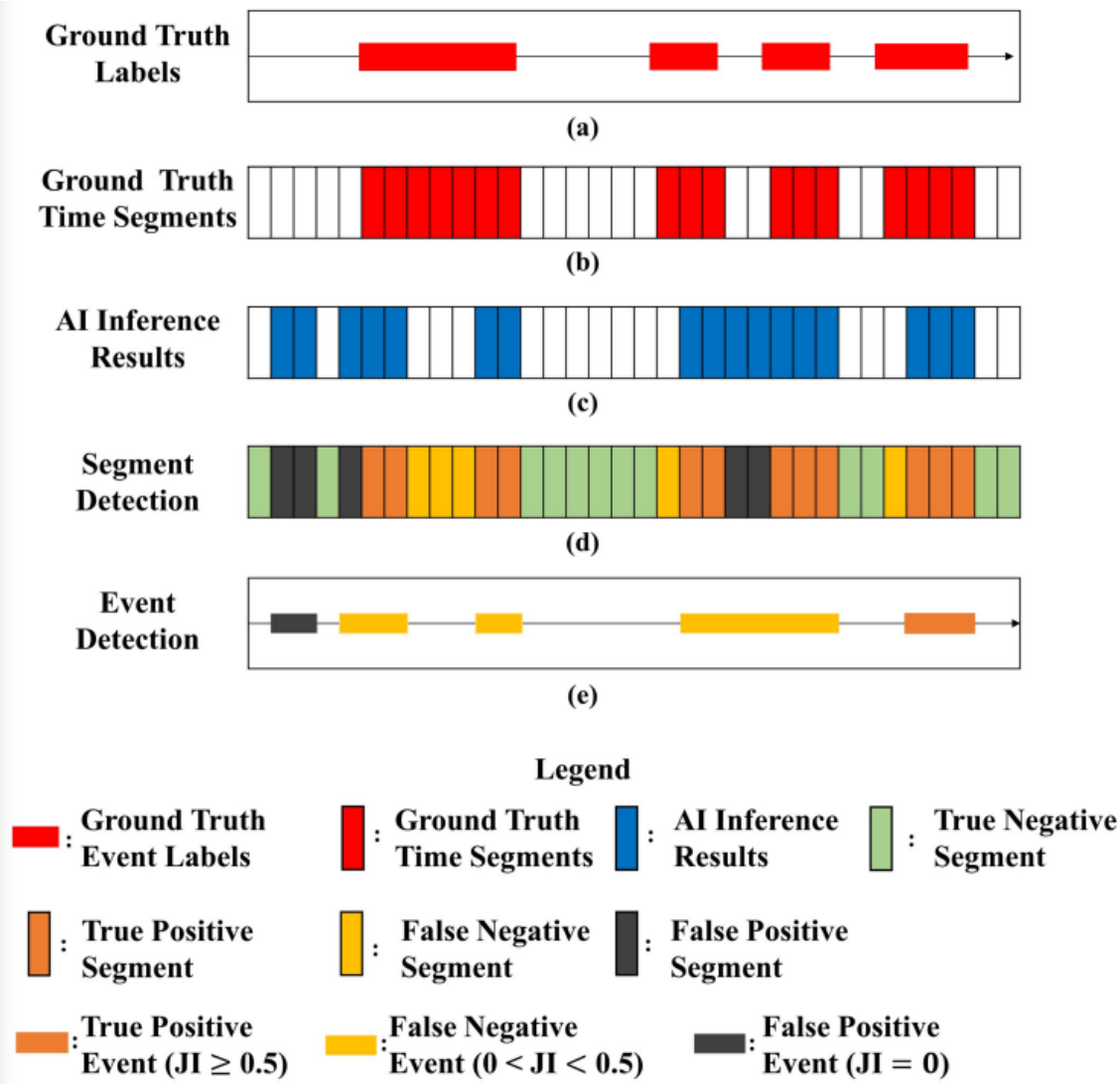


Figure 3.4: Task definition and evaluation metrics (JI: Jaccard Index)

3.4 Limitations of the State of the Art

3.4.1 Datasets

From the datasets presented, only the RSD and HF_Lung_V1 are going to be discussed, since they are the ones publicly available with more data and more information about them (the Tromsø 7 dataset is private, as mentioned above). Regarding these, both have advantages and disadvantages on their side. The RSD has the advantage that the research team is being using it for a long time, since they were the ones who created it. The HF_Lung_V1 has more advantages, since it has more data to work with, has more types of annotations (respiratory cycles - inhalation and exhalation and more type of CAS). One important problem of the HF_Lung_V1 dataset is that a large portion of the dataset does not have any demographic information (TSECC part). Table 3.5 shows the major differences between both datasets.

	RSD	**HF_Lung_V1**
Hours	5h 30min	40h 41min
No. of participants (M/F)	126 (79/46/NA:1)	18 (11/7) + 261 (NA: 261)
No. of recordings	920	9765
Average recording duration	21.5s	15s
Type of ARS annotated	Crackles, Wheezes	Crackles, Wheezes, Stridors, Rhonchus, Respiratory Cycles
No. of annotated DAS	1864	15606
No. of annotated CAS	886	13883
No. of breathing cycles	6898	29295
Diseases of the patients		
Healthy	26	0
Asthma	1	0
COPD	64	1
URTI	14	0
LRTI	2	0
Bronchiectasis	7	0
Bronchilitis	6	0
Pneumonia	6	4
Acute Respiratory Failure	0	4
Chronic Respiratory Failure	0	8
Acute Exacerbation of COPD	0	1
Acute Respiratory Distress Syndrome	0	1
Emphysema	0	1

Table 3.5: Comparison between RSD and HF_Lung_V1 datasets (NA: Not Available)

3.4.2 Classification and Segmentation

There are a lot of studies regarding the classification of ARS, and some achieve very good results but they do not explain everything or they do not do things in the correct way. In some papers, they try to apply DL approaches with a dataset

really small and/or without much variety, which is not possible since those types of models require a large amount of variable data. Other authors achieve "good results" but they do not explain how those models are created, specially DL models, lacking information about their structure. It is also noticeable that there are more papers regarding the classification of ARS than their segmentation, since it is a more complex task that requires the dataset to have annotations with quality (e.g. golden annotations), which is a long time-consuming task. Table 3.6 shows a summary of all the papers mentioned above.

Reference	Dataset	Number of patients	Sounds/Diseases Analysed	Classification or Segmentation	Feature Extraction	Implemented Models	Results
[7]	RSD	126	Normal Wheezes Crackles	Classification	Spectral Features Melodic Features Mel Features Spectrogram Mel Spectrogram	LDA SVMrbf RUSBoost CNN	3 classes: 81.8% Accuracy 2 class Crackle: 87.4% Accuracy 2 class Wheeze: 73.2% Accuracy
[15]	RSD with Random Generated Events	126	Wheezes	Classification	Spectral Features Mel Features Spectrogram	LDA Linear SVM SVMrbf Boosted Trees CNN	Fixed Duration: 96% Accuracy Variable Duration: 67.8% Accuracy
[16]	Partial RSD + Private Dataset	110 + 103	Normal Asthma Pneumonia Bronchiectasis COPD Heart Failure	Classification	1D Wavelet Smoothing Displacement removal	CNN + BiLSTM	99.62% Accuracy 98.43% Sensitivity 99.69% Specificity
[17]	Private Dataset	1630	Normal CAS DAS	Classification	MFCC Spectrogram	SVM CNN	CNN: 76% Accuracy, 79% Precision, 74% Recall SVM: 75% Accuracy, 75% Precision, 99% Recall
[18]	RSD	126	Normal Wheezes Crackles	Classification	Mel Spectrogram	CNN + RNN	Micro: 48.63% Sensitivity, 84.14% Specificity Macro: 58.47% Precision, 58.01% Recall, 57.91 F1 Score
[19]	RSD	126	Normal Wheezes Crackles	Classification	Segmentation of audio file MFCC	LSTM GRU BiLSTM BiGRU	84% Specificity 64% Sensitivity
[20]	Private Dataset	22	Inspiration Expiration	Segmentation	Spectrogram Acoustic Features	ResNet + LSTM	92% Accuracy Inspiration: 90% F1 Score Expiration: 93% F1 Score
[21]	Private Dataset	15	Inspiration Expiration Normal Crackles	Segmentation	MFCC Spectrogram	BiGRNN	Inspiration: 87.0% F1 Score Expiration: 84.6% F1 Score Crackles: 72.1% F1 Score
[14]	Tromsø 7	6048	Inspiration Expiration	Segmentation	Spectrogram	Faster R-CNN	Inspiration: 97% JI, Expiration: 87% JI 97% Sensitivity 84% Specificity
[6]	HF_Lung_V1	279	Inhalation Exhalation ARS	Segmentation (Seg) and Event Detection (ED)	Spectrogram + MFCC + Energy Summation	LSTM GRU BiLSTM BiGRU CNN-LSTM CNN-GRU CNN-BiLSTM CNN-BiGRU	CAS Seg: 53.3% F1 Score CAS ED: 51.6% F1 Score DAS Seg: 71.2% F1 Score DAS ED: 71.4% F1 Score

Table 3.6: Summary of papers presented in State of the Art

Chapter 4

Classification of adventitious events

This thesis was divided into two parts, the classification of adventitious events and the segmentation of adventitious sounds. In this section, a detailed description of the classification part was made.

Since this task was the first one to be developed, to adapt to work to this kind of data and models, the first performed taskwas the replication of the results of the paper [7]. Afterwards, those same models were replicated using the other datasets, to assess if those models were able to generalise and crossed between each other (i.e., models trained with one dataset and tested with another dataset). To conclude this classification task, a stratified analysis of RSD was performed, to better understand which recording devices/demographic quality achieved better results. As said before, with the stratification analysis, an article was published in the International Conference on Biomedical and Health Informatics 2022.

4.1 Dataset

The dataset used is the RSD, explained in Section 3. The data was divided into Train-Test (TT), as already explained before, but while training the models, a random validation set containing 25% of the training set was generated.

The re-annotated data with corrected labels was also used since it had some errors such as labelling, the start and end points were badly annotated and switched labelling files between sound files, and the generation of "other" events is different. From now on, to simplify, this dataset with the corrected labels is going to be called **RSD New Annotations**.

The HF_Lung_V1 Database was also tested, to check if the models developed for [7] were generalised for other datasets. Since the dataset is divided into continuous (wheezes, stridors, and rhonchus) and discontinuous (crackles) sounds, the analysis was done using this division. In this dataset, the events "other" based on the distributions of events were also added, as was done with the RSD.

4.2 Feature Extraction

The features used for the ML are explained with more detail in Chapter 2, Section 2.2. In the article [7], they are extracted by window (3 window methods - Hamming, Blackman-Harris and rectangular; 6 window sizes - 16ms, 32ms, 64ms, 128ms, 256ms, 512ms with 75% overlap), but in these experiments, only the Hamming window method was used (the window sizes used were the same). In total, for each window, there are 81 features (25 spectral, 26 MFCC, and 30 melodic features) and for each feature, five statistics were calculated (mean, standard deviation, median, minimum and maximum values), totalling 2430 features fed to the classifiers.

The features used in the paper for the DL are also extracted by window (3 window methods - Hamming, Blackman-Harris and rectangular; 3 window sizes - 32ms, 64ms and 128ms with 75% overlap), but this time, the spectrogram and the mel-spectrogram are extracted. In these experiments, only the Blackman-Harris window method was used (the window sizes used were the same).

4.3 Feature Selection

Regarding the feature selection process, the Minimum Redundancy Maximum Relevance (MRMR) was the algorithm chosen. This algorithm ranks the features that have maximum relevance regarding the target variable and minimum redundancy pertaining to the features that have been selected. For each experiment, 3 subsets of features were selected: 10 best features, 100 best features and all of the 2430 features.

4.4 Classifiers

The models experimented on paper are Linear Discriminant Analysis (LDA), Random Undersampling Boosted Trees (RUSBoost), Support Vector Machine with radial basis function (SVMrbf) and Convolutional Neural Network (CNN). All these classifiers were trained 10 times with different seeds.

For the ML approaches, the hyperparameters were optimized using the Bayesian Optimization on the validation set explained and the models with the best hyperparameters were then applied to the test set. The hyperparameters optimized were the following:

- Delta for LDA
- Box constraint and kernel scale for SVMrbf
- Number of variables to sample, number of learning cycles, minimum leaf size, and maximum number of splits for RUSBoost

For the DL approaches, 3 models were developed:

- A CNN model with dual input configuration that used the spectrogram and the mel-spectrogram as inputs
- A CNN model with a single input configuration that used the spectrogram as input
- A CNN model with a single input configuration that used the mel-spectrogram as input

The architecture of the dual input CNN is represented in Figure 4.1, and the architectures of the single input CNNs is the same as represented in the Figure 4.1, but only considering the respective branch before the concatenation and the following layers. All these CNNs were trained with 30 epochs, with a batch size of 16 and 0.001 learning rate (ADAM optimization algorithm [22]). An early stop strategy was also used to avoid overfitting during the training phase was also used (i.e., after 10 consecutive epochs with an increase in the validation loss - validated in the validation set explained above).

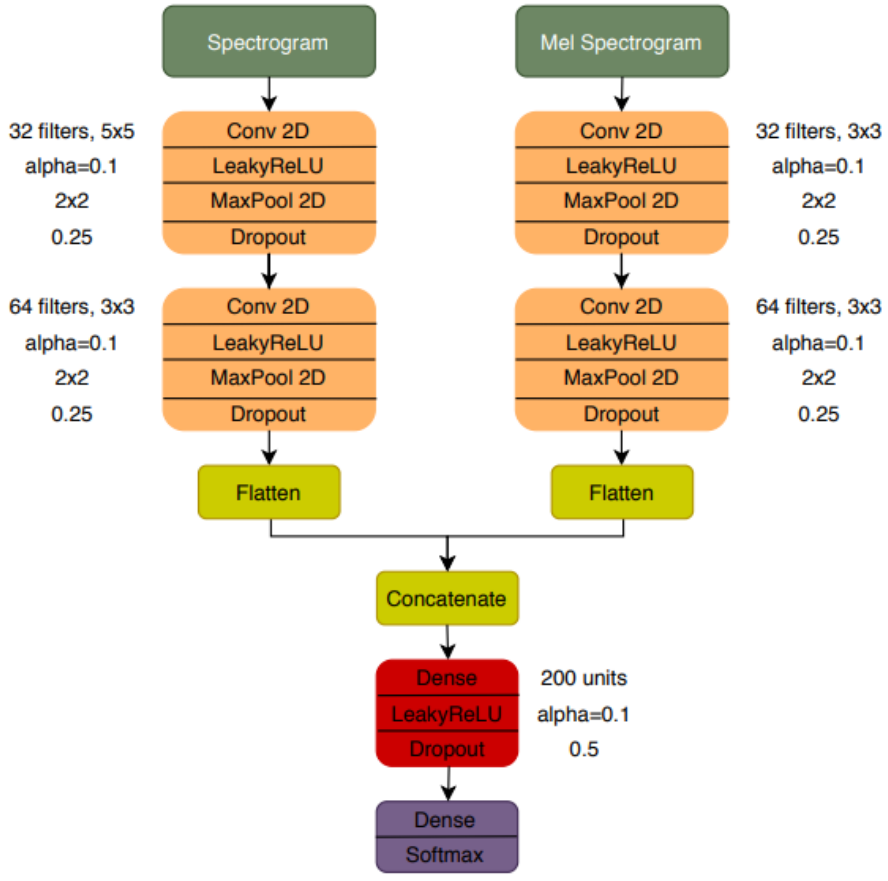


Figure 4.1: Dual Input CNN architecture

4.5 Results

In this section, the performance of the algorithms is analysed. Each experiment is composed of 3-classification problems: 3-classes (crackles vs. wheezes vs. others), 2-class wheezes (wheezes vs. others), and 2-class crackles (crackles vs. others).

The metrics used for the evaluation of the models are explained in Chapter 2, Section 2.4. For the AUC metric, it was only calculated for the binary cases and for the multi-class classification, the evaluation was computed in one-vs-all (crackles vs. the rest, wheezes vs. the rest, and others vs. the rest).

In all the performed comparisons (discussed in the following paragraphs), statistical significance tests were conducted comparing the best model (according to the F1-Score macro) with all the others. When comparing the results for different subpopulations, unpaired tests were performed, namely the unpaired t-test (when the distributions are Gaussian) or the Wilcoxon rank sum test (when the distributions are non-Gaussian). When comparing the results of different algorithms in the same subpopulations, paired tests were performed, namely, the paired T-test (Gaussian distributions) or the Wilcoxon signed rank test (non-Gaussian distributions). In all cases, the Kolmogorov-Smirnov test was employed to test for Gaussianity and the threshold for statistical significance was set to $p < 0.01$. Unless otherwise stated, all the results compared in the paragraphs below are statistically significant.

4.5.1 Respiratory Sound Database

Table 4.1 displays the results obtained by all the classifiers on the test set for the 3-class problem. Table 4.2 displays the results obtained by all the classifiers on the test set for the 2-class problem (crackles vs. others). Table 4.3 displays the results obtained by all the classifiers on the test set for the 2-class problem (wheezes vs. others).

Classifiers	Accuracy	F1 Wheeze	MCC Wheeze	F1 Crackle	MCC Crackle	F1 Other	MCC Other	F1 Macro	MCC Macro
LDA_10MRMR	62.4 ± 0.1	71.0 ± 0.0	67.8 ± 0.0	75.2 ± 0.1	42.5 ± 0.2	17.2 ± 0.1	14.3 ± 0.3	54.5 ± 0.1	41.5 ± 0.2
LDA_100MRMR	65.5 ± 0.0	72.2 ± 0.1	69.7 ± 0.1	76.8 ± 0.1	47.8 ± 0.1	34.9 ± 0.4	22.5 ± 0.1	61.3 ± 0.2	46.7 ± 0.1
LDA_Full	68.8 ± 0.2	72.2 ± 0.1	69.9 ± 0.1	78.3 ± 0.5	53.1 ± 0.7	48.4 ± 1.2	32.4 ± 0.5	66.3 ± 0.6	51.8 ± 0.4
SVMrbf_10MRMR	65.3 ± 0.8	72.8 ± 0.2	69.4 ± 0.3	76.6 ± 0.6	46.8 ± 1.6	32.0 ± 2.8	22.1 ± 1.9	60.5 ± 1.2	46.1 ± 1.3
SVMrbf_100MRMR	67.9 ± 1.2	68.7 ± 3.0	64.1 ± 3.2	77.3 ± 0.5	50.8 ± 1.5	51.1 ± 3.4	30.4 ± 3.0	65.7 ± 2.3	48.4 ± 2.6
SVMrbf_Full	68.5 ± 0.7	67.2 ± 2.9	62.7 ± 2.8	77.0 ± 1.1	51.9 ± 0.8	55.6 ± 3.9	33.0 ± 2.1	66.6 ± 2.6	49.2 ± 1.9
RUSBoost_10MRMR	64.5 ± 0.9	72.5 ± 0.6	69.9 ± 0.7	74.6 ± 0.8	44.1 ± 1.4	39.6 ± 3.2	21.2 ± 2.5	62.2 ± 1.5	45.1 ± 1.5
RUSBoost_100MRMR	68.4 ± 0.7	73.7 ± 0.6	**71.2 ± 0.6**	75.6 ± 1.4	50.6 ± 1.4	54.2 ± 3.5	33.3 ± 2.0	67.8 ± 1.8	51.7 ± 1.3
RUSBoost_Full	69.0 ± 0.6	73.5 ± 0.7	70.3 ± 1.1	75.6 ± 0.8	51.5 ± 0.9	57.3 ± 0.8	34.9 ± 1.0	68.8 ± 0.8	52.2 ± 1.0
CNN_dualInput	**81.6 ± 0.8**	**74.5 ± 1.9**	70.9 ± 1.9	87.8 ± 0.9	**74.6 ± 1.6**	**75.0 ± 1.1**	**62.0 ± 1.5**	**79.1 ± 1.3**	**69.2 ± 1.7**
CNN_Spectrogram	80.2 ± 0.9	72.1 ± 2.9	68.3 ± 3.0	86.9 ± 0.8	72.8 ± 1.5	72.7 ± 2.1	59.5 ± 2.0	77.2 ± 1.9	66.9 ± 2.2
CNN_melSpectrogram	81.4 ± 0.6	74.1 ± 1.7	70.5 ± 1.6	**87.9 ± 0.5**	**74.6 ± 1.1**	73.9 ± 1.1	61.3 ± 1.4	78.6 ± 1.1	68.8 ± 1.4

Table 4.1: Performance results obtained with 3-class problem using the RSD

Regarding the ML approaches, the results were quite similar (even with the Bayesian hyperparameters optimization that has the probabilistic effect on the models). The DL approaches, on the other hand, were a bit different, because the fully-connected layers of the models are seed-dependent, i.e., to obtain the same results, the seeds used were needed. The best model in the 3-class problem was

Classifiers	Accuracy	AUC Crackle	F1 Crackle	MCC Crackle	F1 Other	MCC Other	F1 Macro	MCC Macro
LDA_10MRMR	68.1 ± 0.2	74.7 ± 0.1	76.9 ± 0.1	27.5 ± 0.6	48.4 ± 0.9	27.5 ± 0.6	62.6 ± 0.5	27.5 ± 0.6
LDA_100MRMR	69.8 ± 0.5	76.0 ± 0.4	76.5 ± 0.2	34.6 ± 1.6	58.0 ± 1.6	34.6 ± 1.6	67.2 ± 0.9	34.6 ± 1.6
LDA_Full	68.6 ± 0.6	73.0 ± 1.4	75.3 ± 1.0	32.3 ± 2.1	56.7 ± 2.5	32.3 ± 2.1	66.0 ± 1.8	32.3 ± 2.1
SVMrbf_10MRMR	68.5 ± 0.3	71.6 ± 0.7	78.5 ± 0.2	27.2 ± 1.0	41.5 ± 1.9	27.2 ± 1.0	60.0 ± 1.0	27.2 ± 1.0
SVMrbf_100MRMR	72.4 ± 0.9	78.6 ± 1.7	78.8 ± 0.8	39.5 ± 2.6	60.0 ± 3.9	39.5 ± 2.6	69.4 ± 2.4	39.5 ± 2.6
SVMrbf_Full	71.0 ± 1.0	77.2 ± 1.3	77.2 ± 1.8	37.8 ± 1.8	59.7 ± 4.8	37.8 ± 1.8	68.4 ± 3.3	37.8 ± 1.8
RUSBoost_10MRMR	69.4 ± 0.5	76.0 ± 0.8	75.7 ± 1.0	34.4 ± 0.9	58.6 ± 1.5	34.4 ± 0.9	67.2 ± 1.2	34.4 ± 0.9
RUSBoost_100MRMR	71.2 ± 0.5	79.7 ± 0.5	76.9 ± 0.8	38.6 ± 0.8	61.5 ± 1.0	38.6 ± 0.8	69.2 ± 0.9	38.6 ± 0.8
RUSBoost_Full	71.2 ± 0.5	79.2 ± 0.9	76.8 ± 0.7	38.9 ± 1.1	62.0 ± 1.0	38.9 ± 1.1	69.4 ± 0.8	38.9 ± 1.1
CNN_dualInput	86.3 ± 0.4	83.5 ± 1.4	89.6 ± 0.2	70.4 ± 0.6	79.6 ± 1.6	70.4 ± 0.6	84.6 ± 0.9	70.4 ± 0.6
CNN_Spectrogram	85.3 ± 0.9	82.1 ± 1.5	89.0 ± 0.7	68.4 ± 1.8	77.8 ± 2.1	68.4 ± 1.8	83.4 ± 1.4	68.4 ± 1.8
CNN_melSpectrogram	**87.0 ± 0.7**	**84.0 ± 1.1**	**90.3 ± 0.4**	**72.0 ± 1.4**	**80.4 ± 1.3**	**72.0 ± 1.4**	**85.4 ± 0.8**	**72.0 ± 1.4**

Table 4.2: Performance results obtained with 2-class problem (crackles vs. others) using the RSD

Classifiers	Accuracy	AUC Wheeze	F1 Wheeze	MCC Wheeze	F1 Other	MCC Other	F1 Macro	MCC Macro
LDA_10MRMR	62.4 ± 0.1	62.5 ± 0.2	74.0 ± 0.1	9.3 ± 0.4	32.4 ± 0.5	9.3 ± 0.4	53.2 ± 0.3	9.3 ± 0.4
LDA_100MRMR	56.4 ± 1.6	60.3 ± 1.5	63.1 ± 2.5	11.4 ± 2.1	46.4 ± 1.7	11.4 ± 2.1	54.8 ± 2.1	11.4 ± 2.1
LDA_Full	56.0 ± 1.3	40.8 ± 1.7	62.8 ± 1.5	10.6 ± 3.7	46.0 ± 2.9	10.6 ± 3.7	54.4 ± 2.2	10.6 ± 3.7
SVMrbf_10MRMR	63.4 ± 1.0	63.2 ± 0.8	72.6 ± 1.3	17.5 ± 1.4	44.5 ± 1.6	17.5 ± 1.4	58.6 ± 1.5	17.5 ± 1.4
SVMrbf_100MRMR	66.8 ± 0.9	68.2 ± 1.9	75.0 ± 0.8	25.6 ± 2.7	50.3 ± 3.3	25.6 ± 2.7	62.6 ± 2.0	25.6 ± 2.7
SVMrbf_Full	65.8 ± 1.6	68.0 ± 0.7	73.0 ± 2.4	26.7 ± 2.1	53.2 ± 2.5	26.7 ± 2.1	63.1 ± 2.4	26.7 ± 2.1
RUSBoost_10MRMR	64.6 ± 0.9	67.8 ± 0.7	71.2 ± 1.3	25.7 ± 1.2	53.7 ± 1.1	25.7 ± 1.2	62.4 ± 1.2	25.7 ± 1.2
RUSBoost_100MRMR	64.7 ± 1.1	68.8 ± 1.1	71.8 ± 1.7	25.1 ± 1.4	52.8 ± 1.5	25.1 ± 1.4	62.3 ± 1.6	25.1 ± 1.4
RUSBoost_Full	63.8 ± 1.6	68.6 ± 1.1	70.1 ± 2.5	25.3 ± 2.0	53.9 ± 1.7	25.3 ± 2.0	62.0 ± 2.1	25.3 ± 2.0
CNN_dualInput	72.0 ± 0.9	70.9 ± 1.6	77.6 ± 1.5	40.9 ± 2.4	62.4 ± 2.5	40.9 ± 2.4	70.0 ± 2.0	40.9 ± 2.4
CNN_Spectrogram	**73.0 ± 1.5**	**71.4 ± 1.4**	**78.6 ± 1.9**	**42.2 ± 2.4**	**63.0 ± 2.1**	**42.2 ± 2.4**	**70.8 ± 2.0**	**42.2 ± 2.4**
CNN_melSpectrogram	72.5 ± 1.3	70.9 ± 0.5	78.3 ± 1.9	41.1 ± 0.8	62.3 ± 1.0	41.1 ± 0.8	70.3 ± 1.4	41.1 ± 0.8

Table 4.3: Performance results obtained with 2-class problem (wheezes vs. others) using the RSD

the CNN_dualInput, with F1-Score macro of 79.1% (not statistically significant against the SVMrbf_Full, $p > 0.01$). In the 2-class problem (crackles vs. others), the best model achieved 85.4% F1-Score macro and it was the CNN_melSpectrogram (not statistically significant with the CNN_Spectrogram, $p > 0.01$). Finally, in the last 2-class problem (wheezes vs. others), the best model was the CNN_Spectrogram with F1-Score macro of 70.8% (not statistically significant with the CNN_melSpectrogram and SVMrbf_Full, $p > 0.01$).

4.5.2 Respiratory Sound Database New Annotations

Table 4.4 shows the results obtained for the same classifiers on the test set for the 3-class problem. Table 4.5 shows the results obtained for the same classifiers on the test set for the 2-class problem (crackles vs. others). Table 4.6 shows the results obtained for the same classifiers on the test set for the 2-class problem (wheezes vs. others).

Classifiers	Accuracy	F1 Wheeze	MCC Wheeze	F1 Crackle	MCC Crackle	F1 Other	MCC Other	F1 Macro	MCC Macro
LDA_10MRMR	49.3 ± 0.5	60.8 ± 0.0	59.6 ± 0.0	57.8 ± 1.7	26.8 ± 3.4	34.7 ± 5.4	6.9 ± 1.4	51.1 ± 2.4	31.1 ± 1.6
LDA_100MRMR	44.4 ± 0.0	61.0 ± 0.0	60.1 ± 0.1	60.6 ± 0.0	34.5 ± 0.0	5.2 ± 0.1	1.1 ± 0.2	42.3 ± 0.0	31.9 ± 0.1
LDA_Full	49.0 ± 0.0	61.7 ± 0.1	62.4 ± 0.1	54.9 ± 0.1	22.2 ± 0.1	38.2 ± 0.3	4.6 ± 0.1	51.6 ± 0.2	29.7 ± 0.1
SVMrbf_10MRMR	55.4 ± 0.7	38.2 ± 15.1	36.2 ± 12.6	59.0 ± 0.4	30.0 ± 0.6	54.1 ± 1.2	14.8 ± 1.8	50.4 ± 5.6	27.0 ± 5.0
SVMrbf_100MRMR	65.3 ± 0.5	57.9 ± 3.7	54.2 ± 3.6	63.4 ± 0.4	41.1 ± 0.4	67.9 ± 0.7	31.5 ± 0.8	63.1 ± 1.6	42.3 ± 1.6
SVMrbf_Full	66.5 ± 6.9	50.9 ± 7.6	48.9 ± 6.7	65.7 ± 2.8	45.2 ± 5.4	67.8 ± 13.9	34.4 ± 10.0	61.5 ± 8.1	42.8 ± 7.4
RUSBoost_10MRMR	53.5 ± 0.8	61.6 ± 0.3	62.2 ± 0.3	59.3 ± 1.2	30.8 ± 1.9	44.7 ± 2.6	14.8 ± 1.5	55.2 ± 1.4	35.9 ± 1.2
RUSBoost_100MRMR	62.6 ± 0.7	63.9 ± 0.9	63.7 ± 0.6	63.5 ± 1.0	40.7 ± 1.5	61.5 ± 1.2	29.1 ± 1.2	63.0 ± 1.0	44.5 ± 1.1
RUSBoost_Full	69.9 ± 2.1	**65.3 ± 0.7**	**64.8 ± 0.6**	**72.4 ± 2.5**	**55.5 ± 4.1**	68.8 ± 2.8	44.3 ± 4.3	**68.8 ± 2.0**	**54.9 ± 3.0**
CNN_dualInput	71.8 ± 1.2	59.9 ± 2.8	56.9 ± 2.4	68.6 ± 1.7	51.7 ± 2.0	75.3 ± 1.7	43.6 ± 2.3	67.9 ± 2.1	50.7 ± 2.2
CNN_Spectrogram	69.2 ± 1.2	53.0 ± 8.3	49.9 ± 7.5	63.4 ± 3.7	47.0 ± 2.4	74.3 ± 1.3	37.8 ± 2.7	63.6 ± 4.4	44.9 ± 4.2
CNN_melSpectrogram	**72.2 ± 1.3**	59.6 ± 4.3	56.8 ± 3.4	69.3 ± 1.3	52.3 ± 2.1	**75.6 ± 1.7**	**44.4 ± 2.4**	68.2 ± 2.4	51.2 ± 2.6

Table 4.4: Performance results obtained with 3-class problem using the RSD New Annotations

Classifiers	Accuracy	AUC Crackle	F1 Crackle	MCC Crackle	F1 Other	MCC Other	F1 Macro	MCC Macro
LDA_10MRMR	60.0 ± 0.1	62.5 ± 0.1	56.2 ± 0.3	19.5 ± 0.2	63.0 ± 0.1	19.5 ± 0.2	59.6 ± 0.2	19.5 ± 0.2
LDA_100MRMR	59.6 ± 0.2	64.0 ± 0.1	58.3 ± 0.2	20.3 ± 0.3	60.7 ± 0.3	20.3 ± 0.3	59.5 ± 0.2	20.3 ± 0.3
LDA_Full	57.3 ± 4.2	59.0 ± 4.9	55.0 ± 7.8	15.7 ± 10.3	59.1 ± 1.6	15.7 ± 10.3	57.0 ± 4.7	15.7 ± 10.3
SVMrbf_10MRMR	60.7 ± 0.4	65.5 ± 0.3	59.8 ± 0.6	22.8 ± 0.3	61.5 ± 1.3	22.8 ± 0.3	60.6 ± 1.0	22.8 ± 0.3
SVMrbf_100MRMR	71.7 ± 0.2	78.3 ± 0.3	68.4 ± 0.3	42.9 ± 0.5	74.5 ± 0.1	42.9 ± 0.5	71.4 ± 0.2	42.9 ± 0.5
SVMrbf_Full	66.8 ± 8.4	73.7 ± 6.9	65.4 ± 3.3	34.9 ± 12.0	65.7 ± 19.3	34.9 ± 12.0	65.6 ± 11.3	34.9 ± 12.0
RUSBoost_10MRMR	67.1 ± 0.5	73.2 ± 0.3	64.5 ± 0.4	34.3 ± 0.7	69.3 ± 0.8	34.3 ± 0.7	66.9 ± 0.6	34.3 ± 0.7
RUSBoost_100MRMR	74.7 ± 1.0	83.1 ± 1.0	72.3 ± 0.8	49.3 ± 1.7	76.7 ± 1.2	49.3 ± 1.7	74.5 ± 1.0	49.3 ± 1.7
RUSBoost_Full	**75.4 ± 1.1**	**83.7 ± 1.5**	**73.0 ± 1.2**	**50.8 ± 2.2**	**77.4 ± 1.3**	**50.8 ± 2.2**	**75.2 ± 1.2**	**50.8 ± 2.2**
CNN_dualInput	72.5 ± 1.3	71.5 ± 1.2	66.8 ± 1.7	43.7 ± 2.6	76.4 ± 1.6	43.7 ± 2.6	71.6 ± 1.6	43.7 ± 2.6
CNN_Spectrogram	69.4 ± 3.2	68.7 ± 3.1	63.9 ± 5.5	38.3 ± 5.4	72.9 ± 5.2	38.3 ± 5.4	68.4 ± 5.4	38.3 ± 5.4
CNN_melSpectrogram	72.9 ± 1.2	72.7 ± 1.4	69.4 ± 2.3	45.5 ± 2.7	75.6 ± 1.7	45.5 ± 2.7	72.5 ± 2.0	45.5 ± 2.7

Table 4.5: Performance results obtained with 2-class problem (crackles vs. others) using the RSD New Annotations

Classifiers	Accuracy	AUC Wheeze	F1 Wheeze	MCC Wheeze	F1 Other	MCC Other	F1 Macro	MCC Macro
LDA_10MRMR	60.5 ± 0.8	62.7 ± 1.5	57.6 ± 0.9	20.7 ± 1.6	63.0 ± 0.7	20.7 ± 1.6	60.3 ± 0.8	20.7 ± 1.6
LDA_100MRMR	64.0 ± 0.2	61.4 ± 16.0	59.2 ± 0.1	27.6 ± 0.5	67.8 ± 0.4	27.6 ± 0.5	63.5 ± 0.2	27.6 ± 0.5
LDA_Full	63.5 ± 0.6	32.1 ± 0.8	58.8 ± 0.9	26.5 ± 1.1	67.2 ± 0.7	26.5 ± 1.1	63.0 ± 0.8	26.5 ± 1.1
SVMrbf_10MRMR	65.8 ± 0.9	70.4 ± 1.5	66.4 ± 1.7	32.4 ± 2.1	65.1 ± 1.2	32.4 ± 2.1	65.8 ± 1.4	32.4 ± 2.1
SVMrbf_100MRMR	69.9 ± 0.6	75.6 ± 1.8	67.5 ± 1.4	39.7 ± 1.3	72.0 ± 1.1	39.7 ± 1.3	69.8 ± 1.2	39.7 ± 1.3
SVMrbf_Full	71.0 ± 0.6	**77.7 ± 0.9**	67.3 ± 1.2	41.9 ± 1.3	73.9 ± 1.1	41.9 ± 1.3	70.6 ± 1.2	41.9 ± 1.3
RUSBoost_10MRMR	64.8 ± 0.9	68.8 ± 0.9	62.8 ± 1.0	29.3 ± 1.7	66.5 ± 0.9	29.3 ± 1.7	64.6 ± 1.0	29.3 ± 1.7
RUSBoost_100MRMR	70.5 ± 1.1	76.1 ± 0.9	69.0 ± 1.6	40.9 ± 2.3	71.8 ± 0.8	40.9 ± 2.3	70.4 ± 1.2	40.9 ± 2.3
RUSBoost_Full	69.3 ± 1.4	75.2 ± 1.3	66.7 ± 1.8	38.3 ± 2.8	71.5 ± 1.1	38.3 ± 2.8	69.1 ± 1.5	38.3 ± 2.8
CNN_dualInput	**72.4 ± 1.1**	72.1 ± 1.2	**69.2 ± 2.2**	**44.8 ± 2.1**	**75.0 ± 1.0**	**44.8 ± 2.1**	**72.1 ± 1.6**	**44.8 ± 2.1**
CNN_Spectrogram	67.4 ± 2.2	67.1 ± 2.5	62.9 ± 5.5	35.0 ± 4.3	70.7 ± 2.0	35.0 ± 4.3	66.8 ± 3.8	35.0 ± 4.3
CNN_melSpectrogram	71.7 ± 1.2	71.4 ± 1.4	68.5 ± 2.2	43.2 ± 2.5	74.3 ± 0.8	43.2 ± 2.5	71.4 ± 1.5	43.2 ± 2.5

Table 4.6: Performance results obtained with 2-class problem (wheezes vs. others) using the RSD New Annotations

The best model in the 3-class problem was the RUSBoost_Full, with F1-Score macro of 68.8%. In the 2-class problem (crackles vs. others), the best model achieved 75.2% F1-Score macro and it was once again the RUSBoost_Full (not statistically significant with the LDA_10MRMR and RUSBoost_100MRMR, $p > 0.01$). Finally, in the last 2-class problem (wheezes vs. others), the best model was the CNN_dualInput with F1-Score macro of 72.1% (not statistically significant with the CNN_melSpectrogram, $p > 0.01$).

4.5.3 HF_Lung_V1 Database

Table 4.7 shows the results obtained for the same classifiers on the test set for the 3-class problem. Table 4.8 shows the results obtained for the same classifiers on the test set for the 2-class problem (DAS/crackles vs. others). Table 4.9 shows

the results obtained for the same classifiers on the test set for the 2-class problem (CAS/wheezes vs. others).

Classifiers	Accuracy	F1 Continuous	MCC Continuous	F1 Discontinuous	MCC Discontinuous	F1 Other	MCC Other	F1 Macro	MCC Macro
LDA_10MRMR	70.5 ± 0.2	50.0 ± 0.5	46.0 ± 0.4	48.9 ± 0.5	36.2 ± 0.6	83.4 ± 0.2	58.3 ± 0.5	60.8 ± 0.4	46.8 ± 0.5
LDA_100MRMR	**73.3 ± 0.2**	51.7 ± 0.3	45.9 ± 0.5	**58.3 ± 0.7**	**48.1 ± 0.9**	84.9 ± 0.1	62.5 ± 0.1	65.0 ± 0.4	52.2 ± 0.5
LDA_Full	60.8 ± 3.8	25.7 ± 22.3	17.1 ± 15.9	14.9 ± 15.9	5.3 ± 11.7	78.8 ± 5.5	35.7 ± 30.9	39.8 ± 14.6	19.4 ± 19.5
SVMrbf_10MRMR	70.9 ± 1.2	50.6 ± 2.1	46.7 ± 1.9	50.7 ± 8.3	40.0 ± 5.2	84.4 ± 1.8	62.1 ± 6.2	61.9 ± 4.1	49.6 ± 4.4
SVMrbf_100MRMR	71.4 ± 1.8	47.9 ± 1.5	42.9 ± 2.3	52.2 ± 11.1	43.0 ± 6.3	85.7 ± 2.3	65.9 ± 7.6	61.9 ± 5.0	50.6 ± 5.4
SVMrbf_Full	62.8 ± 7.8	18.2 ± 23.8	16.0 ± 21.1	21.2 ± 27.5	16.8 ± 22.0	78.0 ± 7.3	26.7 ± 34.7	39.1 ± 19.5	19.8 ± 25.9
RUSBoost_10MRMR	71.5 ± 0.6	**54.4 ± 0.9**	**47.5 ± 1.3**	55.7 ± 0.9	45.0 ± 1.2	84.6 ± 0.5	65.6 ± 0.9	64.9 ± 0.8	52.7 ± 1.1
RUSBoost_100MRMR	72.4 ± 0.8	52.3 ± 0.8	46.3 ± 1.2	57.5 ± 1.0	47.5 ± 1.4	**85.8 ± 0.8**	**67.7 ± 1.6**	**65.2 ± 0.9**	**53.8 ± 1.4**
RUSBoost_Full	70.5 ± 1.7	49.3 ± 2.5	43.2 ± 5.7	54.4 ± 5.0	43.5 ± 6.8	84.7 ± 1.1	65.2 ± 2.9	62.8 ± 2.9	50.6 ± 5.1
CNN_dualInput	72.1 ± 0.7	52.6 ± 1.6	46.8 ± 1.6	55.1 ± 2.1	43.9 ± 2.8	85.5 ± 0.6	66.1 ± 1.6	64.4 ± 1.4	52.3 ± 2.0
CNN_Spectrogram	66.7 ± 1.7	43.4 ± 1.0	37.6 ± 1.9	48.5 ± 1.6	35.0 ± 2.1	83.1 ± 1.5	62.7 ± 1.5	58.3 ± 1.4	45.1 ± 1.8
CNN_melSpectrogram	71.0 ± 1.2	51.0 ± 1.5	46.6 ± 1.7	55.0 ± 1.1	44.1 ± 1.4	84.7 ± 1.0	65.2 ± 1.6	63.6 ± 1.2	52.0 ± 1.6

Table 4.7: Performance results obtained with 3-class problem using the HF_Lung_V1

Classifiers	Accuracy	AUC Discontinuous	F1 Discontinuous	MCC Discontinuous	F1 Other	MCC Other	F1 Macro	MCC Macro
LDA_10MRMR	71.0 ± 0.2	67.0 ± 13.8	43.0 ± 0.9	23.7 ± 1.0	80.6 ± 0.1	23.7 ± 1.0	61.8 ± 0.5	23.7 ± 1.0
LDA_100MRMR	85.6 ± 0.0	66.8 ± 40.6	73.2 ± 0.0	63.9 ± 0.0	90.2 ± 0.0	63.9 ± 0.0	81.7 ± 0.0	63.9 ± 0.0
LDA_Full	83.1 ± 2.2	16.6 ± 5.5	65.6 ± 5.5	54.6 ± 6.8	88.8 ± 1.4	54.6 ± 6.8	77.2 ± 3.4	54.6 ± 6.8
SVMrbf_10MRMR	80.0 ± 0.3	84.7 ± 0.1	62.1 ± 0.3	48.8 ± 0.4	86.4 ± 0.2	48.8 ± 0.4	74.2 ± 0.2	48.8 ± 0.4
SVMrbf_100MRMR	**86.7 ± 0.6**	92.6 ± 0.6	75.2 ± 1.1	**66.6 ± 1.5**	**90.9 ± 0.4**	66.6 ± 1.5	**83.1 ± 0.8**	66.6 ± 1.5
SVMrbf_Full	85.2 ± 5.2	88.5 ± 13.6	61.0 ± 32.2	54.5 ± 28.8	90.5 ± 2.4	54.5 ± 28.8	75.8 ± 17.3	54.5 ± 28.8
RUSBoost_10MRMR	82.8 ± 0.7	90.2 ± 0.4	70.3 ± 0.7	60.0 ± 0.9	87.9 ± 0.6	60.0 ± 0.9	79.1 ± 0.6	60.0 ± 0.9
RUSBoost_100MRMR	86.5 ± 0.6	**93.4 ± 0.3**	**75.6 ± 0.7**	62.7 ± 1.0	90.6 ± 0.5	**67.2 ± 1.0**	**83.1 ± 0.6**	**67.2 ± 1.0**
RUSBoost_Full	86.1 ± 1.0	93.3 ± 1.0	75.1 ± 1.6	**66.6 ± 2.2**	90.4 ± 0.8	66.6 ± 2.2	82.8 ± 1.2	66.6 ± 2.2
CNN_dualInput	85.8 ± 0.8	83.0 ± 1.8	72.8 ± 1.6	63.6 ± 1.9	90.4 ± 0.6	63.6 ± 1.9	81.6 ± 1.1	63.6 ± 1.9
CNN_Spectrogram	83.3 ± 1.7	82.0 ± 0.7	70.1 ± 1.3	59.7 ± 1.7	88.4 ± 1.5	59.7 ± 1.7	79.2 ± 1.4	59.7 ± 1.7
CNN_melSpectrogram	85.3 ± 0.9	83.5 ± 0.8	72.8 ± 1.4	63.3 ± 1.9	89.9 ± 0.7	63.3 ± 1.9	81.4 ± 1.0	63.3 ± 1.9

Table 4.8: Performance results obtained with 2-class problem (DAS/crackles vs. others) using the HF_Lung_V1

Classifiers	Accuracy	AUC Continuous	F1 Continuous	MCC Continuous	F1 Other	MCC Other	F1 Macro	MCC Macro
LDA_10MRMR	77.5 ± 0.1	83.4 ± 0.0	50.1 ± 0.1	41.8 ± 0.1	85.4 ± 0.0	41.8 ± 0.1	67.8 ± 0.0	41.8 ± 0.1
LDA_100MRMR	82.6 ± 0.0	90.4 ± 0.0	64.6 ± 0.0	56.5 ± 0.0	88.4 ± 0.0	56.5 ± 0.0	76.5 ± 0.0	56.5 ± 0.0
LDA_Full	81.2 ± 1.1	17.0 ± 3.0	62.8 ± 4.0	52.9 ± 3.2	87.4 ± 0.6	52.9 ± 3.2	75.1 ± 2.3	52.9 ± 3.2
SVMrbf_10MRMR	78.1 ± 0.3	79.9 ± 2.0	52.6 ± 0.4	43.6 ± 0.9	85.7 ± 0.3	43.6 ± 0.9	69.2 ± 0.4	43.6 ± 0.9
SVMrbf_100MRMR	81.8 ± 1.6	86.8 ± 4.0	62.0 ± 3.7	54.5 ± 4.2	88.0 ± 1.0	54.5 ± 4.2	75.0 ± 2.4	54.5 ± 4.2
SVMrbf_Full	82.0 ± 3.6	**90.9 ± 1.4**	60.0 ± 15.7	54.6 ± 11.2	88.3 ± 1.9	54.6 ± 11.2	74.2 ± 8.8	54.6 ± 11.2
RUSBoost_10MRMR	79.5 ± 0.8	85.4 ± 1.1	60.8 ± 2.3	48.6 ± 2.2	86.2 ± 0.4	48.6 ± 2.2	73.5 ± 1.3	48.6 ± 2.2
RUSBoost_100MRMR	**84.5 ± 0.3**	90.2 ± 1.0	**69.5 ± 0.9**	**61.8 ± 0.8**	**89.6 ± 0.2**	**61.8 ± 0.8**	**79.6 ± 0.6**	**61.8 ± 0.8**
RUSBoost_Full	84.0 ± 0.5	89.8 ± 0.9	69.4 ± 1.4	60.5 ± 1.3	89.2 ± 0.4	60.5 ± 1.3	79.3 ± 0.9	60.5 ± 1.3
CNN_dualInput	84.4 ± 0.5	77.1 ± 1.2	69.4 ± 1.8	61.5 ± 1.1	89.5 ± 0.2	61.5 ± 1.1	79.4 ± 1.0	61.5 ± 1.1
CNN_Spectrogram	81.2 ± 0.7	72.2 ± 1.6	61.2 ± 2.7	52.7 ± 1.9	87.6 ± 0.4	52.7 ± 1.9	74.4 ± 1.6	52.7 ± 1.9
CNN_melSpectrogram	83.8 ± 1.0	76.2 ± 1.3	68.0 ± 2.1	59.9 ± 2.6	89.2 ± 0.7	59.9 ± 2.6	78.6 ± 1.4	59.9 ± 2.6

Table 4.9: Performance results obtained with 2-class problem (CAS/wheezes vs. others) using the HF_Lung_V1

The best model in the 3-class problem was the RUSBoost_100MRMR, with F1-Score macro of 65.2% (not statistically significant with the RUSBoost_Full, $p > 0.01$). In the 2-class problem (DAS/crackles vs. others), the best model achieved 75.2% F1-Score macro and it was once again the RUSBoost_100MRMR (not statistically significant with the CNN_Spectrogram, $p > 0.01$). Finally, in the last 2-class problem (CAS/wheezes vs. others), the best model was also the RUSBoost_100MRMR with F1-Score macro of 79.6% (not statistically significant with the LDA_Full, CNN_dualInput and CNN_melSpectrogram, $p > 0.01$).

4.5.4 Comparison between RSD and RSD New Annotations

Since the datasets do not have the same annotations, we cannot compare them directly. As the dataset with the New Annotations is supposed to have corrected

labels, it was expected to have better results, but overall, that did not happen. For the 3-class problem, the best model was the CNN_dualInput with F1-Score macro of 79.1% and decreased its performance by 11.2%, being surpassed by the RUSBoost_Full with F1-Score macro of 68.6%. Regarding the 2-class problem for Crackles, the best model was the CNN_melSpectrogram with F1-Score macro of 85.4% and decreased its performance by 12.9%, being surpassed by the RUSBoost_Full with F1-Score macro of 75.2%.
As for the 2-class problem for Wheezes, the best model was the CNN_Spectrogram with F1-Score macro of 73.0% and decreased its performance by 4.0%, being outperformed by the CNN_dualInput with F1-Score macro of 72.1%, surpassing the best model of the RSD. Even though the overall performance decreased, in general, the DL approaches in both datasets achieved better results. Henceforth, the **RSD New Annotations dataset will not be used**, since the difference between these datasets is not much and it is not public, so it is not possible to compare results with other studies.

4.5.5 Comparison between RSD and HF_Lung_V1 datasets

Regarding the results of training and testing using only one dataset, the results are similar, but if we analyse carefully the annotations of the HF_Lung_V1, we can understand that smaller events, such as crackles, are annotated in the same way as the respiratory cycles, which means these events (that are supposed to have a short duration) are annotated as having a longer duration.

Classifiers	Accuracy	AUC Continuous	F1 Continuous	MCC Continuous	F1 Other	MCC Other	F1 Macro	MCC Macro
LDA_10MRMR	X	X	X	X	X	X	X	X
LDA_100MRMR	X	X	X	X	X	X	X	X
LDA_Full	X	X	X	X	X	X	X	X
SVMrbf_10MRMR	**77.5 ± 5.7**	**83.0 ± 5.2**	**65.6 ± 5.5**	**49.5 ± 9.5**	**83.1 ± 5.1**	**49.5 ± 9.5**	**74.4 ± 5.3**	**49.5 ± 9.5**
SVMrbf_100MRMR	68.8 ± 8.8	81.8 ± 2.0	62.0 ± 3.9	42.5 ± 6.6	72.5 ± 13.3	42.5 ± 6.6	67.2 ± 8.6	42.5 ± 6.6
SVMrbf_Full	70.8 ± 3.1	79.2 ± 3.1	60.1 ± 2.7	40.2 ± 3.8	76.7 ± 4.2	40.2 ± 3.8	68.4 ± 3.4	40.2 ± 3.8
RUSBoost_10MRMR	67.7 ± 2.6	71.9 ± 2.7	54.6 ± 2.2	31.1 ± 4.0	74.9 ± 2.5	31.1 ± 4.0	64.8 ± 2.4	31.1 ± 4.0
RUSBoost_100MRMR	54.1 ± 2.8	76.0 ± 4.0	53.6 ± 1.4	26.8 ± 3.3	54.4 ± 4.8	26.8 ± 3.3	54.0 ± 3.1	26.8 ± 3.3
RUSBoost_Full	74.2 ± 4.6	81.9 ± 2.4	63.7 ± 3.3	45.9 ± 5.9	79.8 ± 4.7	45.9 ± 5.9	71.8 ± 4.0	45.9 ± 5.9
CNN_dualInput	53.5 ± 6.8	61.9 ± 3.2	52.2 ± 1.9	23.8 ± 4.8	53.4 ± 12.5	23.8 ± 4.8	52.8 ± 7.2	23.8 ± 4.8
CNN_Spectrogram	45.5 ± 3.3	54.9 ± 2.6	46.6 ± 2.5	10.2 ± 5.5	43.8 ± 7.0	10.2 ± 5.5	45.2 ± 4.8	10.2 ± 5.5
CNN_melSpectrogram	57.0 ± 3.8	63.1 ± 1.9	52.6 ± 1.7	24.9 ± 3.1	60.2 ± 6.4	24.9 ± 3.1	56.4 ± 4.0	24.9 ± 3.1

Table 4.10: Performance results obtained with 2-class problem (CAS/wheezes vs. others) with the models trained with the RSD and tested with HF_Lung_V1

Classifiers	Accuracy	AUC Discontinuous	F1 Discontinuous	MCC Discontinuous	F1 Other	MCC Other	F1 Macro	MCC Macro
LDA_10MRMR	X	X	X	X	X	X	X	X
LDA_100MRMR	X	X	X	X	X	X	X	X
LDA_Full	X	X	X	X	X	X	X	X
SVMrbf_10MRMR	52.2 ± 7.7	62.2 ± 5.1	41.9 ± 4.5	15.2 ± 7.0	57.9 ± 13.7	15.2 ± 7.0	49.9 ± 9.1	15.2 ± 7.0
SVMrbf_100MRMR	61.0 ± 17.6	**78.6 ± 3.4**	31.3 ± 17.9	17.2 ± 12.3	64.8 ± 27.9	17.2 ± 12.3	48.0 ± 22.9	17.2 ± 12.3
SVMrbf_Full	37.5 ± 5.4	48.5 ± 6.6	37.1 ± 3.3	2.1 ± 5.7	35.9 ± 15.3	2.1 ± 5.7	36.5 ± 9.3	2.1 ± 5.7
RUSBoost_10MRMR	51.0 ± 5.0	74.3 ± 1.6	47.9 ± 1.7	27.4 ± 2.7	53.4 ± 8.0	27.4 ± 2.7	50.6 ± 4.8	27.4 ± 2.7
RUSBoost_100MRMR	**67.6 ± 6.1**	78.0 ± 3.7	**54.4 ± 3.2**	**36.9 ± 5.3**	74.5 ± 6.8	**36.9 ± 5.3**	**64.4 ± 5.0**	**36.9 ± 5.3**
RUSBoost_Full	62.6 ± 9.4	75.5 ± 6.2	51.3 ± 5.1	32.8 ± 8.8	**68.6 ± 10.5**	32.8 ± 8.8	60.0 ± 7.8	32.8 ± 8.8
CNN_dualInput	24.7 ± 0.2	50.0 ± 0.1	39.5 ± 0.0	1.0 ± 0.9	0.4 ± 0.5	1.0 ± 0.9	20.0 ± 0.2	1.0 ± 0.9
CNN_Spectrogram	24.6 ± 0.0	50.0 ± 0.0	39.5 ± 0.0	1.0 ± 0.4	0.1 ± 0.1	1.0 ± 0.4	19.8 ± 0.0	1.0 ± 0.4
CNN_melSpectrogram	24.7 ± 0.1	50.0 ± 0.1	39.5 ± 0.1	0.4 ± 2.0	0.5 ± 0.4	0.4 ± 2.0	20.0 ± 0.2	0.4 ± 2.0

Table 4.11: Performance results obtained with 2-class problem (DAS/crackles vs. others) with the models trained with the RSD and tested with HF_Lung_V1

For both crossing situations, the expected outcome was the same, the decrease of performance, specially in DAS/crackles, since these are annotated differently in

Classifiers	Accuracy	AUC Wheeze	F1 Wheeze	MCC Wheeze	F1 Other	MCC Other	F1 Macro	MCC Macro
LDA_10MRMR	X	X	X	X	X	X	X	X
LDA_100MRMR	X	X	X	X	X	X	X	X
LDA_Full	X	X	X	X	X	X	X	X
SVMrbf_10MRMR	**58.6 ± 1.5**	60.0 ± 0.4	65.0 ± 2.1	16.1 ± 1.6	49.3 ± 0.6	16.1 ± 1.6	57.2 ± 1.4	16.1 ± 1.6
SVMrbf_100MRMR	57.0 ± 4.6	60.6 ± 1.2	62.0 ± 9.1	14.2 ± 4.6	46.9 ± 9.5	14.2 ± 4.6	54.4 ± 9.3	14.2 ± 4.6
SVMrbf_Full	55.1 ± 4.1	61.9 ± 2.9	57.1 ± 8.4	16.8 ± 3.9	51.9 ± 2.2	16.8 ± 3.9	54.5 ± 5.3	16.8 ± 3.9
RUSBoost_10MRMR	61.8 ± 1.1	61.2 ± 0.4	**70.3 ± 1.3**	16.6 ± 1.6	46.3 ± 1.2	16.6 ± 1.6	**58.3 ± 1.2**	16.6 ± 1.6
RUSBoost_100MRMR	57.4 ± 1.2	62.1 ± 1.5	63.2 ± 1.4	15.1 ± 2.5	49.3 ± 1.8	15.1 ± 2.5	56.2 ± 1.6	15.1 ± 2.5
RUSBoost_Full	58.3 ± 1.6	61.0 ± 1.9	65.3 ± 1.5	14.3 ± 3.0	47.8 ± 1.7	14.3 ± 3.0	56.6 ± 1.6	14.3 ± 3.0
CNN_dualInput	56.6 ± 3.5	62.4 ± 1.7	56.1 ± 6.6	25.0 ± 2.6	56.7 ± 1.0	25.0 ± 2.6	56.4 ± 3.8	25.0 ± 2.6
CNN_Spectrogram	58.1 ± 2.9	62.2 ± 1.6	59.9 ± 5.0	24.0 ± 2.8	55.8 ± 1.3	24.0 ± 2.8	57.8 ± 3.2	24.0 ± 2.8
CNN_melSpectrogram	58.3 ± 2.8	**64.6 ± 1.5**	57.5 ± 5.0	**29.5 ± 2.4**	**58.8 ± 1.0**	**29.5 ± 2.4**	58.2 ± 3.0	**29.5 ± 2.4**

Table 4.12: Performance results obtained with 2-class problem (CAS/wheezes vs. others) with the models trained with the HF_Lung_V1 and tested with RSD

Classifiers	Accuracy	AUC Crackle	F1 Crackle	MCC Crackle	F1 Other	MCC Other	F1 Macro	MCC Macro
LDA_10MRMR	X	X	X	X	X	X	X	X
LDA_100MRMR	X	X	X	X	X	X	X	X
LDA_Full	X	X	X	X	X	X	X	X
SVMrbf_10MRMR	46.0 ± 0.7	60.2 ± 2.5	33.2 ± 1.9	12.3 ± 0.8	54.7 ± 0.2	12.3 ± 0.8	44.0 ± 1.0	12.3 ± 0.8
SVMrbf_100MRMR	38.6 ± 2.1	43.6 ± 2.1	15.3 ± 20.1	-2.0 ± 8.3	49.2 ± 7.5	-2.0 ± 8.3	32.2 ± 13.8	-2.0 ± 8.3
SVMrbf_Full	37.7 ± 1.8	42.6 ± 4.4	12.0 ± 18.5	-3.1 ± 8.4	49.8 ± 6.7	-3.1 ± 8.4	30.9 ± 12.6	-3.1 ± 8.4
RUSBoost_10MRMR	56.1 ± 0.7	64.5 ± 2.2	55.6 ± 1.4	21.4 ± 0.6	56.6 ± 0.3	21.4 ± 0.6	56.1 ± 0.8	21.4 ± 0.6
RUSBoost_100MRMR	44.4 ± 5.9	56.5 ± 7.6	31.0 ± 11.7	6.9 ± 10.2	53.2 ± 3.1	6.9 ± 10.2	42.1 ± 7.4	6.9 ± 10.2
RUSBoost_Full	57.3 ± 3.0	**67.4 ± 3.5**	57.0 ± 4.6	23.2 ± 5.0	**57.3 ± 2.4**	23.2 ± 5.0	57.2 ± 3.5	23.2 ± 5.0
CNN_dualInput	64.3 ± 2.8	63.9 ± 2.1	69.2 ± 5.8	28.1 ± 3.7	55.5 ± 5.2	28.1 ± 3.7	62.4 ± 5.5	28.1 ± 3.7
CNN_Spectrogram	**65.6 ± 1.2**	57.5 ± 2.6	**76.4 ± 1.7**	19.4 ± 4.3	35.0 ± 10.1	19.4 ± 4.3	55.7 ± 5.9	19.4 ± 4.3
CNN_melSpectrogram	66.5 ± 1.9	63.7 ± 4.0	73.4 ± 3.8	**28.4 ± 6.0**	52.0 ± 10.6	**28.4 ± 6.0**	**62.7 ± 7.2**	**28.4 ± 6.0**

Table 4.13: Performance results obtained with 2-class problem (DAS/crackles vs. others) with the models trained with the HF_Lung_V1 and tested with RSD

both datasets. For all the following tables in this subsection, the results for the LDA model are not presented, since so far, this type of model was the one that achieved the worst results. Regarding the 3-class problem crossing, both tables are in Appendix A. These results were not sufficiently good, so henceforth, when crossing these datasets, **the 3-class problem comparison will not be performed**.

As we can see by the tables 4.11 (models trained with the RSD data and tested with the HF_Lung_V1 data), whose best model was the RUSBoost_100MRMR with F1-Score macro of 64.4%, and table 4.13 (models trained with the HF_Lung_V1 data and tested with the RSD data), whose best model was the CNN_melSpectrogram with F1-Score macro of 62.7%, achieved quite similar results. In contrast, the tables 4.10 (models trained with the RSD data and tested with the HF_Lung_V1 data), whose best model was the SVMrbf_10MRMR with F1-Score macro of 74.4%, and table 4.12 (models trained with the HF_Lung_V1 data and tested with the RSD data), whose best model was the RUSBoost_10MRMR with F1-Score macro of 58.3%, achieved quite different results (16.1% difference between them).

4.6 Stratification

As the sounds in RSD were collected from subjects of different ages, with various diseases, Body-Mass Index (BMI), and sex and recorded with different types of equipment, a stratified analysis of the results of the models was performed to understand in more detail the behaviour of the models for each subpopulation.

In the categorization by age, all patients under 18 were considered children, the others were considered adults. Regarding the BMI categories, they were defined according to the World Health Organization guidelines [23] and since there were only three underweight patients, they were included in the normal weight category. Concerning the diagnosis category, patients with COPD, asthma, and bronchiectasis were considered chronic; patients with LRTI, URTI, bronchiolitis, or pneumonia were considered non-chronic, whereas participants with no diseases were considered healthy. Table 4.14 shows the number of events per class in each category in the data.

Category	Elements	F Train	F Test	C Train	C Test	W Train	W Test	OC Train	OC Test	OW Train	OW Test
Equipment	AKG C417L	361	285	5387	2682	749	482	1170	1115	274	257
	Littmann 3200	5	5	14	55	28	162	26	297	7	65
	Meditron	87	41	273	144	174	81	884	268	198	66
	Littmann C2SE	86	0	322	0	222	0	398	0	96	0
Age (years)	Adults (19-93, 67.7±11.6)	493	345	5927	2810	1110	676	2204	1441	510	329
	Children (0-18, 4.9.±4.6)	46	30	69	52	63	36	274	180	65	48
	Unknown	0	6	0	19	0	13	0	59	0	11
Sex	Male	272	319	2189	2741	728	682	1510	1256	348	292
	Female	267	56	3807	121	445	30	968	365	227	85
	Unknown	0	6	0	19	0	13	0	59	0	11
BMI	Normal (below 25)	235	91	3913	925	567	115	721	360	179	84
	Overweight (25-29.9)	171	189	1216	908	460	437	910	862	207	190
	Obese (above 30)	84	65	784	977	76	124	501	219	107	55
	Unknown	49	36	83	71	70	49	346	239	82	59
Diagnosis	Chronic (64 COPD, 7 Bronchiectasis, 1 Asthma)	459	351	5899	2829	1085	689	1966	1500	455	340
	Non-Chronic (14 URTI, 2 LRTI, 6 Bronchiolitis, 6 Pneumonia)	62	13	77	43	85	36	385	52	90	15
	Healthy	18	17	20	9	3	0	127	128	30	33

Table 4.14: Distribution of events in the train and test sets per equipment, age (range, mean±standard deviation), sex, BMI (range), and diagnosis [F: Files, C: Annotated Crackles, W: Annotated Wheezes, OC: Annotated Other Crackles, OW: Annotated Other Wheezes]

From Table 4.14 we can conclude that:

- The test set contains no audio files recorded using the LittC2SE stethoscope
- The subjects with Unknown Age do not have any files in the training set; a similar situation is observed for patients with Unknown Sex
- For Healthy subjects there are no cases with annotated wheezes and 9 cases with annotated crackles exist in the test set
- The number of events in each stratification category is not balanced between classes since the goal of the original splitting was to guarantee a 60/40 partition of the data according to the number of respiratory cycles (per class), number of patients and number of files

Given the above stratification and the models already trained with the RSD, the test data were divided into the aforementioned categories, applied to the binary classification problems (crackles vs. others, and wheezes vs. others) and the 3-class problem (crackles vs. wheezes vs. others). Healthy subjects were ignored in the analysis of the results, as their test set did not have annotated wheezes and only contained 9 crackles. The files with missing data (6 files with no information regarding age or sex) were also discarded.

Four metrics were used to evaluate the performance of the classification models: accuracy, area under the curve (AUC), F1-Score, and Matthews Correlation

Coefficient (MCC). For the binary classification tasks, we calculated the accuracy, AUC, MCC, and F1 of the positive class (crackles or wheezes) and macro-averaged F1 (F1 Macro), considering that the dataset is unbalanced. For the 3-class problem, we computed the accuracy, the F1 for each class, and F1 Macro.

Table 4.15 shows summarised version of the results for three: SVMrbf with 100 selected features, RUSBoost with all the features, and CNN with dual input, i.e., with a combination of spectrogram and Mel spectrogram inputs. In Appendix B, tables B.1 to B.24 show the results attained by all the models in these categories.

			2 class crackles			2 class wheezes				3 class		
			SVM	Boost	CNN	SVM	Boost	CNN		SVM	Boost	CNN
Equipment	AKGC417L	Acc	70.3 ± 0.5	68.7 ± 0.6	**86.4 ± 0.8**	64.2 ± 1.3	60.7 ± 1.3	**78.2 ± 0.9**	Acc	69.1 ± 0.6	68.4 ± 0.7	**84.5 ± 1.0**
		AUC	62.7 ± 2.5	64.0 ± 1.0	**79.5 ± 2.3**	59.3 ± 1.5	59.4 ± 1.5	**74.5 ± 2.2**	F1 C	80.5 ± 0.5	77.9 ± 0.8	**90.4 ± 0.8**
		F1 C/W	79.4 ± 0.8	77.3 ± 0.8	**90.9 ± 0.5**	73.3 ± 1.9	67.8 ± 2.3	**83.8 ± 1.0**	F1 W	72.2 ± 2.2	76.8 ± 0.7	**83.0 ± 1.5**
		F1 M	62.8 ± 3.3	63.5 ± 1.2	**81.9 ± 1.8**	59.3 ± 3.3	58.5 ± 2.6	**75.1 ± 2.2**	F1 O	38.8 ± 4.3	45.3 ± 1.3	**71.9 ± 1.9**
		MCC M	26.3 ± 3.6	27.2 ± 1.5	**66.2 ± 1.8**	19.4 ± 2.3	18.1 ± 2.8	**51.2 ± 2.2**	F1 M	63.8 ± 2.3	66.7 ± 0.9	**81.8 ± 1.4**
	Litt3200	Acc	77.3 ± 6.7	78.2 ± 1.9	**90.9 ± 1.3**	**71.5 ± 3.4**	65.2 ± 3.0	51.4 ± 1.8	Acc	55.7 ± 5.9	**68.9 ± 1.4**	65.1 ± 1.8
		AUC	71.2 ± 3.1	67.0 ± 3.4	**86.3 ± 0.7**	62.4 ± 1.7	**62.5 ± 2.2**	57.7 ± 3.3	F1 C	9.6 ± 4.8	2.0 ± 3.0	**16.1 ± 7.1**
		F1 C/W	46.7 ± 4.0	42.0 ± 4.8	**80.4 ± 2.1**	**80.7 ± 2.9**	73.7 ± 3.0	57.1 ± 2.0	F1 W	62.7 ± 6.3	**71.2 ± 2.1**	50.1 ± 6.0
		F1 M	66.0 ± 4.8	64.3 ± 3.1	**87.3 ± 1.5**	**63.1 ± 2.3**	60.9 ± 2.8	50.5 ± 2.7	F1 O	65.0 ± 5.4	78.1 ± 1.4	**78.2 ± 1.7**
		MCC M	36.1 ± 5.1	29.8 ± 5.9	**74.7 ± 3.1**	**27.2 ± 5.5**	23.4 ± 4.3	14.0 ± 6.0	F1 M	45.8 ± 5.5	**50.4 ± 2.2**	48.1 ± 4.9
	Meditron	Acc	86.9 ± 1.0	**87.7 ± 1.4**	85.2 ± 1.5	72.2 ± 3.0	77.4 ± 2.4	**79.6 ± 1.8**	Acc	70.7 ± 2.5	**73.6 ± 1.7**	71.6 ± 1.8
		AUC	**88.6 ± 0.9**	86.5 ± 1.8	86.3 ± 0.8	72.6 ± 2.9	78.4 ± 2.2	**80.4 ± 1.8**	F1 C	56.3 ± 2.5	**58.5 ± 5.8**	57.5 ± 3.8
		F1 C/W	**83.5 ± 1.2**	82.4 ± 2.2	81.3 ± 1.2	73.0 ± 3.8	77.0 ± 3.1	**80.3 ± 2.2**	F1 W	59.6 ± 3.3	59.5 ± 1.5	**61.5 ± 3.3**
		F1 M	86.4 ± 1.1	**86.5 ± 1.6**	84.5 ± 1.4	72.1 ± 3.3	77.4 ± 2.5	**79.6 ± 2.0**	F1 O	80.8 ± 2.9	**84.2 ± 1.3**	80.7 ± 1.2
		MCC M	**74.2 ± 1.9**	73.0 ± 3.2	70.4 ± 1.9	45.1 ± 5.5	57.1 ± 3.9	**60.5 ± 3.5**	F1 M	65.6 ± 2.9	**67.4 ± 2.9**	66.6 ± 2.8
Age	Adults	Acc	71.5 ± 0.9	69.9 ± 0.6	**86.7 ± 0.8**	65.6 ± 0.6	61.4 ± 1.6	**70.5 ± 1.0**	Acc	67.2 ± 0.9	68.1 ± 0.5	**81.9 ± 0.8**
		AUC	66.9 ± 2.3	67.4 ± 0.7	**82.5 ± 1.9**	59.8 ± 1.6	60.0 ± 1.3	**68.7 ± 1.9**	F1 C	77.8 ± 0.4	76.0 ± 0.7	**88.7 ± 0.8**
		F1 C/W	79.0 ± 0.8	76.8 ± 0.7	**90.5 ± 0.5**	75.0 ± 0.7	69.0 ± 2.4	**77.0 ± 1.6**	F1 W	68.7 ± 2.7	73.9 ± 0.9	**74.4 ± 1.9**
		F1 M	67.1 ± 2.8	67.1 ± 1.0	**84.2 ± 1.5**	60.0 ± 2.2	58.8 ± 2.3	**67.7 ± 2.3**	F1 O	45.2 ± 3.6	52.4 ± 0.9	**73.5 ± 1.1**
		MCC M	34.9 ± 3.2	34.3 ± 1.2	**70.0 ± 1.7**	20.2 ± 2.7	19.0 ± 2.3	**36.3 ± 2.9**	F1 M	63.9 ± 2.2	67.4 ± 0.8	**78.9 ± 1.3**
	Child	Acc	80.6 ± 1.5	**84.4 ± 1.8**	82.9 ± 2.1	81.2 ± 3.9	**86.9 ± 2.1**	86.0 ± 3.2	Acc	77.2 ± 4.2	78.2 ± 1.5	**80.6 ± 3.0**
		AUC	84.2 ± 1.5	80.8 ± 3.4	**86.4 ± 1.4**	81.6 ± 3.8	**86.9 ± 2.5**	86.1 ± 3.6	F1 C	65.3 ± 5.1	64.1 ± 4.0	**67.1 ± 2.3**
		F1 C/W	67.7 ± 2.0	68.0 ± 4.3	**70.9 ± 2.0**	79.4 ± 4.1	**84.9 ± 2.9**	84.0 ± 4.2	F1 W	72.7 ± 4.9	71.0 ± 2.0	**82.8 ± 3.2**
		F1 M	76.9 ± 1.6	78.9 ± 2.8	**79.4 ± 1.9**	81.1 ± 4.0	**86.7 ± 2.3**	85.7 ± 3.6	F1 O	82.4 ± 3.8	83.7 ± 1.4	**84.8 ± 3.0**
		MCC M	58.9 ± 2.8	58.2 ± 5.6	**63.4 ± 2.4**	62.8 ± 7.6	**73.6 ± 4.5**	72.3 ± 6.7	F1 M	73.5 ± 4.6	72.9 ± 2.5	**78.2 ± 2.8**
Sex	Male	Acc	72.0 ± 0.8	70.4 ± 0.7	**87.3 ± 0.7**	68.7 ± 1.0	64.4 ± 1.9	**72.4 ± 1.3**	Acc	69.2 ± 0.7	69.4 ± 0.7	**82.7 ± 1.0**
		AUC	66.6 ± 1.8	67.8 ± 0.8	**82.4 ± 1.9**	62.2 ± 1.6	62.8 ± 1.1	**70.4 ± 1.9**	F1 C	79.4 ± 0.4	77.1 ± 0.8	**89.4 ± 0.9**
		F1 C/W	79.8 ± 1.0	77.6 ± 0.8	**91.2 ± 0.4**	77.8 ± 1.0	72.4 ± 2.4	**79.2 ± 1.8**	F1 W	71.6 ± 2.7	**76.9 ± 0.8**	76.8 ± 2.1
		F1 M	66.8 ± 2.3	67.0 ± 1.0	**84.2 ± 1.4**	62.2 ± 2.2	61.0 ± 2.1	**68.9 ± 2.2**	F1 O	46.1 ± 2.9	51.9 ± 0.9	**72.7 ± 1.1**
		MCC M	34.0 ± 2.2	34.4 ± 1.4	**69.9 ± 1.6**	24.7 ± 2.7	23.9 ± 2.0	**38.9 ± 2.6**	F1 M	65.7 ± 2.0	68.6 ± 0.8	**79.6 ± 1.4**
	Female	Acc	72.0 ± 5.4	72.9 ± 1.7	**80.3 ± 3.1**	50.8 ± 4.2	54.6 ± 3.1	**66.0 ± 3.4**	Acc	55.7 ± 4.5	63.1 ± 1.2	**74.6 ± 2.3**
		AUC	78.3 ± 3.2	76.3 ± 2.3	**84.1 ± 1.1**	48.0 ± 5.7	46.4 ± 3.3	**62.3 ± 1.7**	F1 C	49.2 ± 2.4	53.8 ± 2.4	**64.2 ± 1.9**
		F1 C/W	62.0 ± 3.7	60.4 ± 2.3	**70.0 ± 2.7**	30.8 ± 6.6	24.9 ± 5.6	**45.5 ± 2.3**	F1 W	29.0 ± 4.9	34.4 ± 3.0	**41.7 ± 4.7**
		F1 M	69.9 ± 4.8	69.9 ± 1.9	**77.7 ± 2.8**	46.3 ± 5.2	46.1 ± 4.4	**60.3 ± 2.9**	F1 O	62.9 ± 5.6	72.1 ± 1.3	**81.9 ± 2.1**
		MCC M	49.3 ± 5.3	45.8 ± 3.8	**60.5 ± 3.0**	-3.5 ± 10.1	-6.7 ± 6.2	**22.8 ± 3.1**	F1 M	47.0 ± 4.3	53.4 ± 2.2	**62.6 ± 2.9**
BMI	Normal	Acc	79.0 ± 1.7	80.0 ± 1.4	**88.1 ± 1.3**	63.0 ± 2.7	63.1 ± 1.7	**79.4 ± 2.5**	Acc	75.0 ± 0.8	76.9 ± 1.0	**87.2 ± 0.9**
		AUC	65.9 ± 3.7	70.4 ± 1.5	**80.2 ± 3.0**	61.1 ± 2.8	62.4 ± 1.8	**78.1 ± 2.6**	F1 C	85.9 ± 0.3	86.7 ± 0.9	**92.6 ± 0.6**
		F1 C/W	86.8 ± 0.8	86.9 ± 1.1	**92.3 ± 0.8**	69.5 ± 3.0	67.6 ± 1.9	**82.9 ± 2.2**	F1 W	69.2 ± 2.8	72.2 ± 1.0	**82.2 ± 1.7**
		F1 M	67.7 ± 4.6	72.3 ± 1.8	**83.3 ± 2.6**	60.9 ± 4.4	62.3 ± 2.3	**78.4 ± 2.9**	F1 O	41.5 ± 5.0	51.4 ± 2.2	**75.1 ± 2.4**
		MCC M	41.8 ± 6.4	46.6 ± 3.4	**69.7 ± 3.4**	23.1 ± 5.5	24.8 ± 3.6	**57.6 ± 5.1**	F1 M	65.5 ± 2.7	70.1 ± 1.4	**83.3 ± 1.6**
	Overweight	Acc	74.2 ± 2.8	75.1 ± 1.2	**85.5 ± 1.6**	**68.3 ± 2.0**	64.2 ± 2.4	67.9 ± 2.2	Acc	64.9 ± 1.9	70.3 ± 0.8	**78.0 ± 0.7**
		AUC	73.7 ± 2.9	74.7 ± 1.3	**85.3 ± 1.7**	59.5 ± 2.6	60.6 ± 2.2	**66.9 ± 1.2**	F1 C	72.0 ± 0.9	75.5 ± 0.6	**83.4 ± 0.9**
		F1 C/W	78.5 ± 1.4	78.7 ± 0.7	**87.0 ± 1.0**	**78.2 ± 1.4**	73.0 ± 2.5	75.0 ± 2.8	F1 W	69.9 ± 1.8	**74.0 ± 1.1**	70.4 ± 1.7
		F1 M	73.0 ± 3.7	74.4 ± 1.5	**85.3 ± 1.8**	59.9 ± 2.9	59.7 ± 2.9	**65.0 ± 2.2**	F1 O	52.2 ± 4.3	61.9 ± 1.7	**76.0 ± 1.3**
		MCC M	51.2 ± 4.0	52.0 ± 2.0	**72.0 ± 2.5**	20.5 ± 5.5	20.3 ± 4.3	**31.8 ± 2.2**	F1 M	64.7 ± 2.3	70.5 ± 1.1	**76.6 ± 1.3**
	Obese	Acc	59.5 ± 5.3	51.6 ± 3.3	**88.1 ± 2.5**	59.3 ± 3.3	50.0 ± 4.0	**75.0 ± 3.8**	Acc	62.9 ± 1.7	54.8 ± 2.2	**83.4 ± 2.6**
		AUC	62.5 ± 3.4	61.7 ± 2.2	**79.8 ± 0.8**	57.9 ± 1.1	54.7 ± 2.9	**72.6 ± 3.7**	F1 C	75.5 ± 1.7	62.7 ± 2.8	**90.7 ± 1.8**
		F1 C/W	69.7 ± 5.2	60.6 ± 3.8	**92.6 ± 1.8**	67.4 ± 4.1	53.6 ± 7.4	**81.3 ± 3.3**	F1 W	62.8 ± 9.2	74.8 ± 1.9	**76.6 ± 4.0**
		F1 M	53.8 ± 4.2	48.8 ± 2.8	**80.1 ± 2.4**	56.2 ± 3.0	49.3 ± 5.2	**71.7 ± 4.1**	F1 O	33.0 ± 2.2	36.0 ± 1.1	**61.8 ± 2.6**
		MCC M	19.6 ± 5.2	18.3 ± 3.4	**61.1 ± 4.9**	15.0 ± 2.6	9.0 ± 5.5	**43.9 ± 7.4**	F1 M	57.1 ± 4.4	57.8 ± 1.9	**76.4 ± 2.8**
Diagnosis	Chronic	Acc	71.9 ± 0.9	70.5 ± 0.6	**86.6 ± 0.9**	65.6 ± 0.7	61.9 ± 1.6	**70.9 ± 1.0**	Acc	67.4 ± 0.9	68.4 ± 0.5	**81.8 ± 0.9**
		AUC	67.8 ± 2.1	68.3 ± 0.7	**82.7 ± 1.9**	60.2 ± 1.6	60.7 ± 1.2	**69.3 ± 1.8**	F1 C	77.6 ± 0.4	75.9 ± 0.8	**88.3 ± 0.9**
		F1 C/W	79.1 ± 0.8	76.9 ± 0.7	**90.4 ± 0.5**	74.8 ± 0.7	69.2 ± 2.5	**77.3 ± 1.6**	F1 W	68.5 ± 2.7	73.6 ± 0.8	**74.2 ± 1.9**
		F1 M	68.1 ± 2.6	67.9 ± 0.9	**84.3 ± 1.5**	60.3 ± 2.1	59.5 ± 2.3	**68.3 ± 2.2**	F1 O	47.1 ± 3.4	54.1 ± 0.9	**73.9 ± 1.2**
		MCC M	36.7 ± 3.0	36.0 ± 1.2	**70.2 ± 1.8**	20.8 ± 2.7	20.4 ± 2.2	**37.4 ± 2.7**	F1 M	64.4 ± 2.2	67.9 ± 0.8	**78.8 ± 1.3**
	Non-Chronic	Acc	81.6 ± 2.5	80.5 ± 2.6	**85.3 ± 3.3**	82.4 ± 3.7	84.1 ± 3.6	**85.5 ± 4.4**	Acc	77.8 ± 3.5	78.2 ± 2.2	**84.7 ± 1.7**
		AUC	82.4 ± 2.5	80.7 ± 2.8	**86.0 ± 2.8**	80.9 ± 4.0	82.3 ± 2.5	**83.5 ± 5.1**	F1 C	78.7 ± 3.6	75.4 ± 4.1	**84.0 ± 1.7**
		F1 C/W	81.8 ± 2.4	79.2 ± 3.2	**85.4 ± 2.6**	87.0 ± 3.0	88.4 ± 3.0	**89.6 ± 3.2**	F1 W	83.4 ± 4.7	85.8 ± 2.0	**90.4 ± 3.0**
		F1 M	81.6 ± 2.6	80.4 ± 2.8	**85.3 ± 3.3**	79.5 ± 4.2	81.5 ± 3.5	**82.9 ± 5.1**	F1 O	73.6 ± 4.1	75.0 ± 2.5	**81.6 ± 2.6**
		MCC M	65.1 ± 4.9	61.3 ± 5.5	**72.4 ± 5.0**	60.1 ± 7.4	63.5 ± 6.2	**66.0 ± 10.1**	F1 M	78.6 ± 4.1	78.7 ± 2.9	**85.3 ± 2.4**

Table 4.15: Results (Acc: Accuracy, C: Crackle, W: Wheeze, O: Other, SVM: SVMrbf_100MRMR, Boost: RUSBoost_Full, CNN: CNN_dualInput, M: Macro)

In all the performed comparisons (discussed in the following paragraphs), statistical significance tests were conducted. When comparing the results for different subpopulations, unpaired tests were performed, namely the unpaired t-

test (when the distributions are Gaussian) or the Wilcoxon rank sum test (when the distributions are non-Gaussian). When comparing the results of different algorithms in the same subpopulations, paired tests were performed, namely, the paired T-test (Gaussian distributions) or the Wilcoxon signed rank test (non-Gaussian distributions). In all cases, the Kolmogorov-Smirnov test was employed to test for Gaussianity and the threshold for statistical significance was set to $p < 0.01$. Unless otherwise stated, all the results compared in the paragraphs below are statistically significant.

Looking at Table 4.15, we can observe that, for the three types of classification problems, the model that obtained the best results overall was the CNN, except for four cases: Children in wheezes classification, where the Boost performed better; Litt3200 stethoscope in wheezes classification, where the SVM outperformed the CNN; Meditron stethoscope in crackles classification, where the SVM model also outperformed the CNN; and finally, in the 3-class classification problem, where the SVM also performed better than the CNN in the Litt3200 and Meditron.

For the Equipment category, in wheezes classification, the results are quite similar between all 3 stethoscopes/microphones, with a slight advantage for the AKGC417L microphone (with F1 wheezes of 83.8% in the AKGC417L microphone, maximum F1 wheezes of 80.7% in the Litt3200, and F1 wheezes of 80.3% in the Meditron). In crackles classification, the results are higher for the AKGC417L microphone and Meditron, while in the Litt3200, the results are worse (F1 crackles of 90.9% in the AKGC417L microphone, F1 crackles of 80.4% in the Litt3200, and a maximum F1 crackles of 83.5% in the Meditron). In the 3-class problem, the AKGC417L microphone also achieved better results than the other two (F1 macro of 81.8% in the AKGC417L microphone, a maximum F1 macro of 50.4% in the Litt3200, and F1 macro of 67.4% in the Meditron). Overall, the AKGC417L microphone achieved better results, since this microphone is more sensitive, it has no filters, and its training and test sets are larger than the sets for the other types of equipment.

When we look at the Age of the subjects, in wheezes classification, Children' achieved better results than Adults (with F1 wheezes of 84.9% in Children using the Boost model and F1 wheezes of 77.0% in Adults using the CNN model). Even though the Boost achieved better results than the CNN in wheezes classification in Children, the difference was not statistically significant ($p > 0.01$). In crackles classification, the reverse occurs, with Adults outperforming Children (with F1 crackles of 90.5% in Adults and F1 crackles of 70.9% in Children). Regarding the 3-class problem, Adults once again outperformed Children (with F1 macro of 78.9% and F1 macro of 78.2%, respectively).

Regarding the Sex category, Male subjects achieved superior results in wheezes classification than Female subjects (F1 wheezes of 79.2% in Males and F1 wheezes of 45.5% in Females). In the classification of crackles, the same occurs, Male subjects also achieved superior results than Female Subjects (F1 crackles of 91.2% in Males and F1 crackles of 70.0% in Females). As for the 3-class classification problem, even though there is still an advantage for the Male subjects, the difference is lower (F1 macro of 79.6% in Males and F1 macro of 62.6% in Females). Even

though Female subjects have a large number of annotated crackles in the training set, these differences can be explained by the unbalanced data in both crackles and wheezes in Female subjects between train and test sets. Overall, the results on the classification of crackles were superior to the results on the classification of wheezes, as there are more annotated crackles in both training and test sets than wheezes.

Regarding the BMI category, in the classification of wheezes, Obese and Normal BMI subjects achieved better results than Overweight BMI subjects (with F1 wheezes of 81.3% in Obese BMI subjects, F1 wheezes of 82.9% in Normal BMI subjects, and maximum F1 wheezes of 78.2% in Overweight BMI subjects). In the crackles classification, Obese and Normal BMI subjects achieved better results than Overweight BMI subjects (with F1 crackles of 92.6% in Obese BMI subjects, F1 crackles of 92.3% in Normal BMI subjects, and F1 crackles of 87.0% in Overweight BMI subjects - except for the Obese and Normal CNNs where $p > 0.01$). In the 3-class problem, Obese and Overweight BMI subjects achieved similar results (in terms of F1 macro, with 76.4% and 76.6%, respectively), while the Normal BMI subjects achieved superior results (with F1 macro of 83.3% - except for the Obese and Normal CNNs, where $p > 0.01$). Overall, crackles classification achieved better results than wheezes classification, maybe due to having more annotated crackles than wheezes in the training set, which benefits the CNN model.

In the Diagnosis category, the subjects with a Non-Chronic diagnosis achieved better results in the wheezes classification than the subjects with a Chronic diagnosis (F1 wheezes of 89.5% and F1 wheezes of 77.3%, respectively). In the classification of crackles, the reverse occurs and the subjects with Chronic diagnosis surpassed the subjects with Non-Chronic diagnosis (F1 crackles of 90.4% and F1 crackles of 85.4%, respectively). As for the 3-class classification problem, the same as the wheezes classification happened: the subjects with Non-Chronic diagnosis once again achieved better results than those with Chronic diagnosis (F1 macro of 85.3% and F1 macro of 78.8%, respectively). Regarding crackles classification, that difference may be explained by the fact that the AKGC417L microphone has more sensibility than any other equipment, and in the training set of the subjects with Non-Chronic diagnosis, there are only files where the equipment used were the LittC2SE and Meditron (less sensibility in general), while in the training set of the subjects with Chronic diagnosis, most of the files were recorded using the AKGC417L microphone.

Chapter 5

Segmentation of adventitious events

As explained before, this thesis was divided into two parts, the classification of adventitious events and the segmentation of adventitious sounds. In this section, a detailed description of the segmentation part was made.

Segmentation of ARS is a complex task, especially the post-processing analysis since the first part is the same as the already explained classification of ARS. As a consequence of that, only **the binary problems are going to be evaluated** (crackles vs. normal sounds, and wheezes vs. normal sounds).

Regarding the segmentation, two approaches were tested:

- CNN model to classify individual frames
- RNN/LSTM model to classify a group of frames

Usually, between these two approaches, the one that can achieve better results is the second one, since it takes into account the past and the future at any given point, while the first approach only assesses each frame individually.

5.1 Dataset

One of the datasets used is the RSD, explained in Section 3.1. The data was divided into Train-Test (TT), as already explained before, but while training the models, a random validation set containing 25% of the training set was generated.

The HF_Lung_V1 Database was also used and also explained in Section 3.1. Since this dataset is divided into continuous (wheezes, stridors, and rhonchus) and discontinuous (crackles) sounds, the analysis was done using this division. Also, a random validation set containing 25% of the training set was generated.

In both datasets, in contrast to the classification problem, no events "other" were added, since in this case, we cannot correctly replicate the characteristics of the events in analysis. To surpass this problem, in the first approach, a parameter

was passed to the models to have the classifier heavily weigh the few examples of the event that are available. The second approach, since it was a replication, it was not done anything regarding the unbalanced data.

In the second approach, the RSD only using the AKGC417L files is going to be used. Henceforth, this dataset will be denominated by RSD_AKGC417L.

5.2 Segmentation using individual frame classification

As explained before, this approach is will consist in creating a CNN model to classify individual frames, without any relation to the previous or the next frames.

5.2.1 Feature Extraction

Unlike the classification problem, no ML models were tested, since overall they achieved worst results than the DL models in classification.

The features for the DL models are extracted by window (3 window methods - Hamming, Blackman-Haris and rectangular; 3 window sizes - 32ms, 64ms and 128ms with 75% overlap), and the spectrogram and the mel-spectrogram used. In these experiments, only the Blackman-Haris window method was used (the window sizes used were the same).

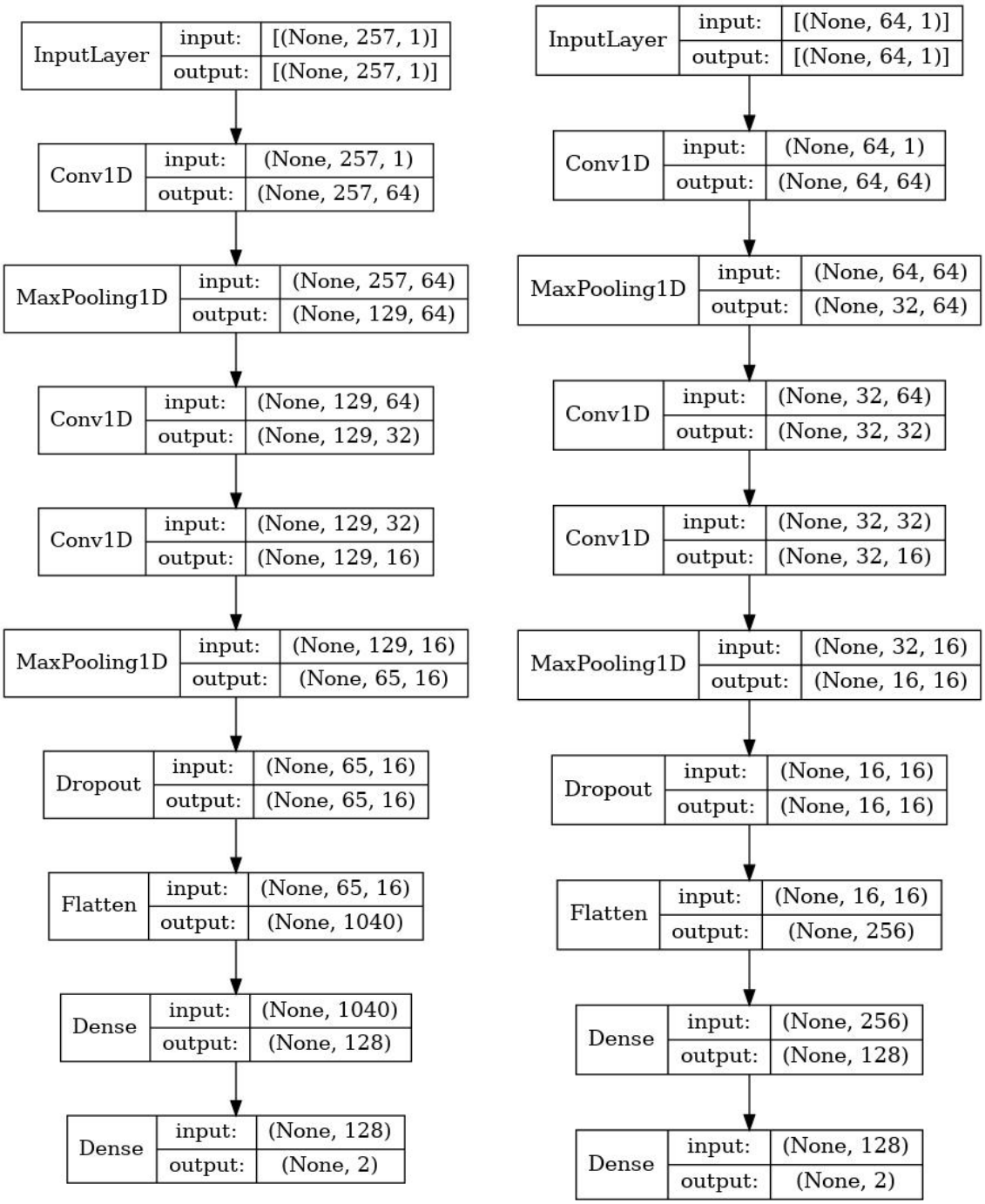
5.2.2 Classifiers

All the classifiers in which the RSD was dataset chosen were trained 10 times with different seeds, whilst when the HF_Lung_V1 dataset was chosen were trained 5 times with different seeds, since the size difference is quite significant and there are some computational limitations. Also due to computational limitations and size differences between datasets, when the HF_Lung_V1 dataset was used to train the model, the 32ms window size was not used.

For the DL approaches, 2 models were developed:

- A CNN model with single input configuration that used the spectrogram as input (Figure 5.1a)

- A CNN model with single input configuration that used the mel-spectrogram as input (Figure 5.1b)



(a) Spectrogram

(b) Mel-Spectrogram

Figure 5.1: CNNs architecture using Spectrogram as input (left) and using Mel-Spectrogram as input (right)

All these CNNs were trained with 30 epochs, with a batch size of 32 and 0.001 learning rate (ADAM optimisation algorithm [22]). It also used a model checkpoint that saves the one with the lowest validation loss to avoid overfitting during the training phase.

5.2.3 Post-processing

The post-processing analysis is similar to the one used in [6] and the metrics used for the evaluation of the models are explained in subsection 2.4. Starting this process, the probability of the frames is converted to an integer value (0 as being a normal sound frame or 1 as being an event sound frame) with three different thresholds: 0.25, 0.5 and 0.75. Next, sequential frames with the same classification are aggregated into an event and then compared with the real annotations (the energy peaks were not computed as it was in the article), using the Jaccard Index and Overlap Coefficient. Although JI is widely used for segmentation tasks, OC metric is more relevant for this problem, as JI is insensitive to the length of the segments [24], but the results for both metrics are going to be presented, in pair with 3 metrics (Recall, Precision, and F1-Score). Also, for a better understanding of the performance of the models, the results for the individual frames are presented with the same metrics.

5.2.4 Results

HF_Lung_V1 Database

Table 5.1 displays the results obtained by all the classifiers on the test set for the 2-class problem (DAS vs. normal sounds). Table 5.2 displays the results obtained by all the classifiers on the test set for the 2-class problem (CAS vs. normal sounds). The following tables do not have results when the window size is 32ms, due to computational limitations.

	Recall_Mel	Precision_Mel	F1-Score_Mel	Recall_Spec	Precision_Spec	F1-Score_Spec
Class_0.25_64ms	96.7 ± 2.2	5.9 ± 0.4	11.2 ± 0.7	**96.8 ± 2.0**	5.8 ± 0.2	10.9 ± 0.4
Class_0.5_64ms	75.7 ± 5.6	10.2 ± 1.2	17.9 ± 1.7	69.4 ± 10.4	11.8 ± 2.3	19.9 ± 2.6
Class_0.75_64ms	11.7 ± 6.2	25.1 ± 3.0	14.6 ± 5.0	9.1 ± 9.0	25.4 ± 2.7	11.0 ± 7.7
Class_0.25_128ms	**96.9 ± 0.7**	6.2 ± 0.2	11.6 ± 0.3	96.4 ± 1.8	6.0 ± 0.3	11.2 ± 0.5
Class_0.5_128ms	75.4 ± 2.1	11.8 ± 0.5	**20.3 ± 0.7**	68.2 ± 8.7	13.0 ± 2.2	**21.6 ± 2.7**
Class_0.75_128ms	14.6 ± 3.9	**27.3 ± 1.5**	18.6 ± 3.0	9.3 ± 5.9	**26.7 ± 2.1**	12.7 ± 5.8
Seg_JI_0.25_64ms	24.3 ± 2.4	0.7 ± 0.1	1.4 ± 0.2	25.2 ± 1.8	0.9 ± 0.1	1.7 ± 0.2
Seg_JI_0.5_64ms	9.8 ± 2.5	0.4 ± 0.0	0.8 ± 0.1	9.0 ± 3.7	0.4 ± 0.1	0.8 ± 0.2
Seg_JI_0.75_64ms	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.1 ± 0.2	0.0 ± 0.0	0.0 ± 0.1
Seg_JI_0.25_128ms	**32.0 ± 2.6**	1.0 ± 0.0	1.9 ± 0.1	**30.5 ± 3.5**	**1.1 ± 0.1**	**2.2 ± 0.1**
Seg_JI_0.5_128ms	21.9 ± 2.0	**1.1 ± 0.0**	**2.1 ± 0.1**	18.9 ± 5.1	**1.1 ± 0.1**	2.1 ± 0.1
Seg_JI_0.75_128ms	0.5 ± 0.3	0.1 ± 0.1	0.2 ± 0.1	0.4 ± 0.5	0.1 ± 0.1	0.2 ± 0.2
Seg_Over_0.25_64ms	89.8 ± 1.1	2.6 ± 0.3	5.1 ± 0.6	89.4 ± 0.8	3.0 ± 0.3	5.9 ± 0.5
Seg_Over_0.5_64ms	**97.1 ± 1.0**	4.2 ± 0.8	8.0 ± 1.5	**97.5 ± 1.0**	5.3 ± 1.9	9.9 ± 3.4
Seg_Over_0.75_64ms	91.6 ± 6.0	**22.6 ± 3.9**	**35.9 ± 4.5**	78.4 ± 16.4	23.7 ± 3.8	**35.8 ± 4.7**
Seg_Over_0.25_128ms	90.8 ± 0.6	2.8 ± 0.2	5.4 ± 0.4	91.0 ± 0.6	3.4 ± 0.6	6.5 ± 1.0
Seg_Over_0.5_128ms	96.8 ± 0.3	4.8 ± 0.3	9.2 ± 0.6	97.2 ± 1.1	5.9 ± 1.5	11.1 ± 2.7
Seg_Over_0.75_128ms	85.9 ± 4.3	22.4 ± 1.9	35.5 ± 2.1	70.3 ± 16.9	**24.6 ± 3.3**	35.3 ± 2.4

Table 5.1: Performance results of CNN on DAS vs. normal sounds problem using the HF_Lung_V1 (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, 0.25/0.5/0.75: Threshold, 32/64/128ms: Window size, *_Mel: CNN with mel-spectrogram as input, *_Spec: CNN with spectrogram as input)

	Recall_Mel	Precision_Mel	F1-Score_Mel	Recall_Spec	Precision_Spec	F1-Score_Spec
Class_0.25_64ms	**94.7 ± 2.5**	6.9 ± 0.2	12.8 ± 0.4	**94.9 ± 4.1**	6.8 ± 0.5	12.7 ± 0.8
Class_0.5_64ms	59.1 ± 7.0	11.8 ± 0.7	19.5 ± 0.9	63.3 ± 10.8	11.0 ± 1.4	18.6 ± 1.6
Class_0.75_64ms	16.6 ± 2.6	23.7 ± 1.8	19.3 ± 1.6	19.4 ± 4.0	20.0 ± 2.3	19.2 ± 1.6
Class_0.25_128ms	93.5 ± 3.8	7.1 ± 0.4	13.1 ± 0.7	92.7 ± 2.2	7.2 ± 0.2	13.4 ± 0.4
Class_0.5_128ms	56.6 ± 6.5	12.4 ± 1.2	20.2 ± 1.5	51.1 ± 6.7	14.1 ± 0.9	22.0 ± 0.5
Class_0.75_128ms	20.7 ± 3.7	**25.1 ± 4.2**	**22.3 ± 2.0**	19.3 ± 3.4	**27.7 ± 2.8**	**22.4 ± 1.7**
Seg_JI_0.25_64ms	26.3 ± 4.8	0.9 ± 0.1	1.8 ± 0.2	22.4 ± 6.4	0.9 ± 0.1	1.8 ± 0.3
Seg_JI_0.5_64ms	6.5 ± 1.2	0.4 ± 0.0	0.8 ± 0.1	8.2 ± 3.4	0.5 ± 0.1	0.9 ± 0.1
Seg_JI_0.75_64ms	2.2 ± 0.1	0.4 ± 0.0	0.7 ± 0.0	2.0 ± 0.5	0.3 ± 0.0	0.5 ± 0.1
Seg_JI_0.25_128ms	**30.5 ± 4.2**	**1.5 ± 0.3**	**2.9 ± 0.5**	**32.2 ± 1.0**	**1.5 ± 0.1**	**2.8 ± 0.1**
Seg_JI_0.5_128ms	13.7 ± 2.0	1.0 ± 0.0	1.9 ± 0.1	13.2 ± 2.5	1.1 ± 0.1	2.1 ± 0.2
Seg_JI_0.75_128ms	6.1 ± 0.8	1.3 ± 0.2	2.1 ± 0.2	6.0 ± 1.2	**1.5 ± 0.1**	2.3 ± 0.2
Seg_Over_0.25_64ms	91.3 ± 1.0	3.3 ± 0.4	6.3 ± 0.8	93.1 ± 1.8	4.0 ± 1.1	7.7 ± 2.1
Seg_Over_0.5_64ms	**98.0 ± 0.4**	6.3 ± 0.6	11.7 ± 1.1	97.7 ± 1.0	5.8 ± 1.2	10.9 ± 2.1
Seg_Over_0.75_64ms	90.7 ± 3.7	14.3 ± 1.2	24.6 ± 1.8	93.8 ± 2.5	13.0 ± 2.0	22.7 ± 2.9
Seg_Over_0.25_128ms	92.7 ± 0.5	4.5 ± 0.5	8.5 ± 0.9	93.0 ± 0.3	4.1 ± 0.2	7.9 ± 0.3
Seg_Over_0.5_128ms	**98.0 ± 0.4**	7.1 ± 0.8	13.3 ± 1.4	**97.8 ± 0.3**	8.1 ± 1.0	14.9 ± 1.7
Seg_Over_0.75_128ms	83.6 ± 6.1	**15.5 ± 2.9**	**26.0 ± 4.0**	79.6 ± 6.2	**16.8 ± 1.9**	**27.7 ± 2.2**

Table 5.2: Performance results of CNN on CAS vs. normal sounds problem using the HF_Lung_V1 (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, 0.25/0.5/0.75: Threshold, 32/64/128ms: Window size, *_Mel: CNN with mel-spectrogram as input, *_Spec: CNN with spectrogram as input)

After analysing the tables, we can understand that the results are quite low (better performance when the OC is used), even before the post-processing in the classification of the individual frames. Regarding the DAS vs. normal sounds, the best results were attained when the threshold was 0.75, using the OC, a window size of 64ms, and the mel-spectrogram as input, with an F1-Score of 35.9%. Concerning the CAS vs. normal sounds, the best results were attained when the threshold was 0.75, using the OC, a window size of 128ms, and the spectrogram as input, with an F1-Score of 27.7%. Both of these tables help us better understand the difference between the 2 metrics OC and JI since the difference between results is significant. Given the lower overall results with this approach, also **these results are just going to be used as a baseline for the next approach**.

RSD

Table 5.3 displays the results obtained by all the classifiers on the test set for the 2-class problem (crackles vs. normal sounds). Table 5.4 displays the results obtained by all the classifiers on the test set for the 2-class problem (wheezes vs. normal sounds).

	Recall_Mel	Precision_Mel	F1-Score_Mel	Recall_Spec	Precision_Spec	F1-Score_Spec
Class_0.25_32ms	**97.2 ± 1.6**	1.4 ± 0.0	2.8 ± 0.1	**98.1 ± 1.9**	1.3 ± 0.0	2.6 ± 0.0
Class_0.5_32ms	79.8 ± 11.1	2.4 ± 0.6	4.6 ± 1.1	84.4 ± 13.0	2.0 ± 0.6	3.9 ± 1.1
Class_0.75_32ms	31.4 ± 17.2	4.5 ± 2.5	7.7 ± 4.1	13.8 ± 11.9	3.8 ± 3.1	5.9 ± 4.8
Class_0.25_64ms	93.4 ± 4.9	1.6 ± 0.1	3.2 ± 0.3	96.4 ± 1.4	1.4 ± 0.1	2.8 ± 0.1
Class_0.5_64ms	70.2 ± 8.4	3.9 ± 0.5	7.4 ± 0.8	77.3 ± 5.6	3.2 ± 0.7	6.1 ± 1.2
Class_0.75_64ms	38.0 ± 9.8	**7.6 ± 0.8**	**12.5 ± 1.0**	41.6 ± 4.8	**7.1 ± 0.9**	**12.1 ± 1.2**
Class_0.25_128ms	94.4 ± 2.3	1.5 ± 0.1	2.9 ± 0.3	96.6 ± 3.3	1.3 ± 0.2	2.7 ± 0.4
Class_0.5_128ms	74.7 ± 8.8	3.0 ± 0.4	5.7 ± 0.8	82.1 ± 17.0	2.6 ± 1.5	4.9 ± 2.8
Class_0.75_128ms	36.0 ± 14.8	5.4 ± 0.9	8.8 ± 0.5	24.3 ± 21.3	3.7 ± 3.3	6.2 ± 5.3
Seg_JI_0.25_32ms	12.9 ± 7.0	0.5 ± 0.2	0.9 ± 0.4	7.2 ± 6.0	0.4 ± 0.3	0.7 ± 0.6
Seg_JI_0.5_32ms	35.7 ± 17.9	1.1 ± 0.6	2.1 ± 1.1	25.1 ± 20.5	0.7 ± 0.6	1.4 ± 1.2
Seg_JI_0.75_32ms	27.3 ± 14.4	2.4 ± 1.3	4.5 ± 2.3	12.6 ± 10.6	2.0 ± 1.6	3.4 ± 2.8
Seg_JI_0.25_64ms	19.0 ± 5.3	0.7 ± 0.2	1.3 ± 0.4	15.4 ± 1.2	0.8 ± 0.1	1.5 ± 0.2
Seg_JI_0.5_64ms	**48.3 ± 2.1**	3.2 ± 0.6	5.9 ± 1.1	**48.0 ± 2.9**	2.4 ± 0.7	4.6 ± 1.3
Seg_JI_0.75_64ms	38.2 ± 7.6	**7.5 ± 0.8**	**12.4 ± 1.1**	41.2 ± 3.3	**7.2 ± 0.9**	**12.2 ± 1.1**
Seg_JI_0.25_128ms	11.6 ± 3.5	0.5 ± 0.1	1.0 ± 0.2	10.7 ± 9.5	0.5 ± 0.4	0.9 ± 0.8
Seg_JI_0.5_128ms	32.3 ± 3.1	2.4 ± 0.7	4.5 ± 1.3	25.9 ± 21.5	2.2 ± 2.4	4.1 ± 4.2
Seg_JI_0.75_128ms	25.9 ± 7.2	6.4 ± 1.4	9.8 ± 1.0	21.6 ± 18.2	4.6 ± 4.0	7.4 ± 6.2
Seg_Over_0.25_32ms	**97.2 ± 1.5**	**10.0 ± 13.3**	15.8 ± 18.8	**95.8 ± 4.2**	**17.3 ± 15.7**	26.5 ± 22.0
Seg_Over_0.5_32ms	94.1 ± 3.4	9.5 ± 13.5	15.0 ± 19.3	93.8 ± 5.4	16.1 ± 16.6	24.4 ± 23.7
Seg_Over_0.75_32ms	59.0 ± 30.3	5.2 ± 2.9	9.4 ± 5.1	28.8 ± 24.0	4.3 ± 3.5	7.5 ± 6.1
Seg_Over_0.25_64ms	92.6 ± 3.9	3.3 ± 0.2	6.4 ± 0.4	94.0 ± 1.9	4.6 ± 1.1	8.8 ± 2.0
Seg_Over_0.5_64ms	80.5 ± 6.7	5.1 ± 0.6	9.6 ± 1.1	85.8 ± 3.9	4.2 ± 0.8	7.9 ± 1.5
Seg_Over_0.75_64ms	51.7 ± 11.5	9.8 ± 0.9	**16.3 ± 1.2**	55.8 ± 5.3	9.5 ± 1.0	16.2 ± 1.3
Seg_Over_0.25_128ms	89.5 ± 2.3	3.9 ± 0.3	7.6 ± 0.5	94.7 ± 4.8	16.7 ± 15.4	25.7 ± 21.8
Seg_Over_0.5_128ms	71.1 ± 7.3	5.0 ± 0.6	9.4 ± 0.9	82.7 ± 15.9	**17.3 ± 14.4**	**26.8 ± 20.5**
Seg_Over_0.75_128ms	36.1 ± 13.7	8.2 ± 0.7	12.8 ± 0.9	26.0 ± 22.5	5.3 ± 4.4	8.5 ± 7.0

Table 5.3: Performance results of CNN on crackles vs. normal sounds problem using the RSD (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, 0.25/0.5/0.75: Threshold, 32/64/128ms: Window size, *_Mel: CNN with mel-spectrogram as input, *_Spec: CNN with spectrogram as input)

	Recall_Mel	Precision_Mel	F1-Score_Mel	Recall_Spec	Precision_Spec	F1-Score_Spec
Class_0.25_32ms	**88.3 ± 2.0**	7.1 ± 0.3	13.2 ± 0.4	89.4 ± 2.6	6.9 ± 0.3	12.7 ± 0.5
Class_0.5_32ms	56.9 ± 4.0	9.9 ± 0.4	16.8 ± 0.4	57.1 ± 3.9	9.7 ± 0.4	16.6 ± 0.5
Class_0.75_32ms	21.0 ± 3.7	11.0 ± 0.2	14.3 ± 0.7	22.4 ± 6.0	10.9 ± 0.3	14.5 ± 0.9
Class_0.25_64ms	84.6 ± 2.5	7.7 ± 0.3	14.1 ± 0.5	**90.1 ± 4.1**	7.2 ± 0.5	13.2 ± 0.8
Class_0.5_64ms	50.3 ± 3.8	11.6 ± 0.6	18.8 ± 0.6	59.6 ± 5.0	11.4 ± 0.4	19.0 ± 0.4
Class_0.75_64ms	21.1 ± 2.9	14.3 ± 0.7	17.0 ± 0.7	27.7 ± 3.3	14.5 ± 0.6	19.0 ± 0.9
Class_0.25_128ms	85.7 ± 2.5	8.1 ± 0.4	14.8 ± 0.7	89.7 ± 4.0	7.8 ± 0.8	14.3 ± 1.2
Class_0.5_128ms	54.9 ± 3.8	13.4 ± 0.4	**21.6 ± 0.5**	61.6 ± 5.4	13.7 ± 1.5	22.3 ± 1.5
Class_0.75_128ms	28.3 ± 3.4	**17.5 ± 0.7**	21.5 ± 1.2	34.1 ± 4.0	**18.9 ± 1.5**	**24.2 ± 1.3**
Seg_JI_0.25_32ms	11.5 ± 1.7	0.4 ± 0.0	0.7 ± 0.1	11.8 ± 2.1	0.4 ± 0.0	0.7 ± 0.1
Seg_JI_0.5_32ms	3.8 ± 0.6	0.3 ± 0.0	0.5 ± 0.1	3.4 ± 0.8	0.2 ± 0.0	0.5 ± 0.1
Seg_JI_0.75_32ms	1.1 ± 0.1	0.1 ± 0.0	0.3 ± 0.0	1.0 ± 0.2	0.1 ± 0.0	0.2 ± 0.0
Seg_JI_0.25_64ms	15.5 ± 1.3	0.7 ± 0.0	1.3 ± 0.1	**21.3 ± 3.5**	0.7 ± 0.0	1.4 ± 0.0
Seg_JI_0.5_64ms	6.3 ± 0.4	0.6 ± 0.0	1.0 ± 0.0	8.5 ± 0.7	0.7 ± 0.0	1.2 ± 0.0
Seg_JI_0.75_64ms	3.3 ± 0.4	0.6 ± 0.1	1.0 ± 0.2	3.7 ± 0.1	0.6 ± 0.1	1.1 ± 0.1
Seg_JI_0.25_128ms	**26.1 ± 1.8**	1.4 ± 0.1	2.6 ± 0.1	31.1 ± 1.9	1.4 ± 0.1	2.7 ± 0.2
Seg_JI_0.5_128ms	15.6 ± 2.0	1.6 ± 0.2	**3.0 ± 0.3**	17.9 ± 1.9	1.9 ± 0.1	3.4 ± 0.2
Seg_JI_0.75_128ms	8.2 ± 1.0	**1.8 ± 0.1**	**3.0 ± 0.2**	9.9 ± 1.2	**2.3 ± 0.3**	**3.7 ± 0.4**
Seg_Over_0.25_32ms	**95.4 ± 0.4**	3.1 ± 0.3	6.0 ± 0.5	**94.9 ± 0.3**	2.9 ± 0.3	5.7 ± 0.6
Seg_Over_0.5_32ms	94.4 ± 0.4	6.3 ± 0.4	11.8 ± 0.7	**94.9 ± 0.3**	6.4 ± 0.6	11.9 ± 1.1
Seg_Over_0.75_32ms	91.7 ± 1.6	10.5 ± 0.6	18.9 ± 1.0	91.4 ± 0.7	10.8 ± 0.9	19.3 ± 1.4
Seg_Over_0.25_64ms	93.6 ± 1.0	4.0 ± 0.4	7.6 ± 0.6	94.6 ± 0.4	3.3 ± 0.5	6.3 ± 1.0
Seg_Over_0.5_64ms	91.9 ± 0.7	7.6 ± 0.6	14.0 ± 1.0	93.1 ± 0.9	6.9 ± 0.6	12.8 ± 1.0
Seg_Over_0.75_64ms	84.9 ± 2.8	12.9 ± 0.9	**22.4 ± 1.3**	87.1 ± 2.4	12.5 ± 1.0	21.9 ± 1.4
Seg_Over_0.25_128ms	94.2 ± 1.6	4.8 ± 0.3	9.2 ± 0.6	94.7 ± 0.9	4.1 ± 0.5	7.9 ± 0.9
Seg_Over_0.5_128ms	88.5 ± 1.9	8.7 ± 0.4	15.8 ± 0.6	90.2 ± 1.5	8.9 ± 1.2	16.1 ± 1.9
Seg_Over_0.75_128ms	76.8 ± 3.2	**14.8 ± 1.0**	24.8 ± 1.3	78.8 ± 2.8	**16.0 ± 2.1**	**26.5 ± 2.8**

Table 5.4: Performance results of CNN on wheezes vs. normal sounds problem using the RSD (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, 0.25/0.5/0.75: Threshold, 32/64/128ms: Window size, *_Mel: CNN with mel-spectrogram as input, *_Spec: CNN with spectrogram as input)

As well as using the HF_Lung_V1, we can understand that the results are quite low (better performance when the OC is used), even before the post-processing in the classification of the individual frames. Regarding the crackles vs. normal sounds, the best results were attained when the threshold was 0.5, using the OC, a window size of 128ms, and the spectrogram as input, with an F1-Score of 26.8%. However, for the combination of all the parameters, overall, the best results were obtained when the threshold was 0.75 and the window size of 64ms, which is expected, since these events are short are the models classify almost everything as being an event, so a higher threshold controls better which is considered an event or not. Concerning the wheezes vs. normal sounds, the best results were attained when the threshold was 0.75, using the OC, a window size of 128ms, and the spectrogram as input, with an F1-Score of 26.5%, which is also expected, since these events are longer and, similarly to the previous models, almost everything is classified as being an event, so a higher threshold controls better which is considered an event or not. Given the lower overall results with this approach, **these results are just going to be used as a baseline for the next approach**.

5.3 Segmentation using sequential frame classification

In this approach, the architecture of the model used was a replication of one of the models on the paper [6].

5.3.1 Feature Extraction

Regarding the feature extraction, all the recordings were resampled to 4000 Hz, then applied a high-pass filter (at 80 Hz) to eliminate the heart sound noise, and STFT with Hanning window with a size of 256ms and 64ms hop length, without additional zero-padding. After this process, a 15s sound signal is transformed into the corresponding spectrogram. Following the spectrogram extraction, more features are extracted: MFCCs (20 static coefficients, 20 delta coefficients, and 20 acceleration coefficients) and Energy Summation (four frequency bands). In the end, a 938 x 193 feature matrix combines all of those above. To conclude, a min-max normalization is performed on each feature (values between 0 and 1). Since most of the files of RSD and RSD_AKGC417L are longer than 15s, the files were divided into 15s chunks with 75% overlap, and if necessary, a zero-padding was added to the feature matrix in order to keep the size constant. In contrast, the HF_Lung_V1 has almost every recording with 15s, so the sounds longer than 15s were discarded (similarly to the [6]), and the recordings with shorter than 15s were also discarded. Also, the recordings with no events were discarded in the training and test sets in the datasets.

5.3.2 Classifiers

All the classifiers in both datasets were trained with a single seed, due to some computational limitations.

The model used was a CNN-BiLSTM with some parameters slightly different, also due to computational limitations (Figure 5.2) (e.g., in the article, this same model has 6,959,809 parameters in total, while this replicated model has 2,987,746 parameters in total).

In [6], a 5-fold Cross Validation was used, but once again due to computational limitations, it was used a Train-Test approach was used with a 25% random validation set. These classifiers were trained with 100 epochs, with a batch size of 16 and a 0.0001 learning rate (ADAM optimisation algorithm [22]). It also used a model checkpoint that saves the one with the lowest validation loss during the training phase, and an early stop strategy (i.e., after 50 consecutive epochs with an increase in the validation loss) to avoid overfitting.

5.3.3 Post-processing

The post-processing analysis is similar to the one used in [6] and the metrics used for the evaluation of the models are explained in subsection 2.4. As there are going to be common segments between these 15s chunks in the RSD and RSD_AKGC417L, an approach was used to turn the output of each chunk of each file into a single output array, explained in Figure 5.3. Even though in Figure 5.3, the final values are calculated with the average, the median is also going to be tested. This process is not performed in the HF_Lung_V1, as most of the files have 15s.

Following that conversion to a single array per file, the probability of the frames is converted to an integer value (0 as being a normal sound frame or 1 as being an event sound frame) with three different thresholds: 0.25, 0.5 and 0.75. Next, sequential frames with the same classification are aggregated into an event and then compared with the real annotations, using the Jaccard Index and Overlap Coefficient. Although JI is widely used for segmentation tasks, OC metric is more relevant for this problem, as JI is insensitive to the length of the segments [24], but the results for both metrics are going to be presented, in pair with 3 metrics (Recall, Precision, and F1-Score). Also, for a better understanding of the performance of the models, the results for the individual frames are presented with the same metrics.

5.3.4 Results

Given that the chunks without events were discarded and the recordings longer and shorter than 15s in the HF_Lung_V1 dataset, the training and test sets are smaller. Table 5.5 shows the difference between using the full datasets against the discarded mentioned files in the number of chunks of all sets.

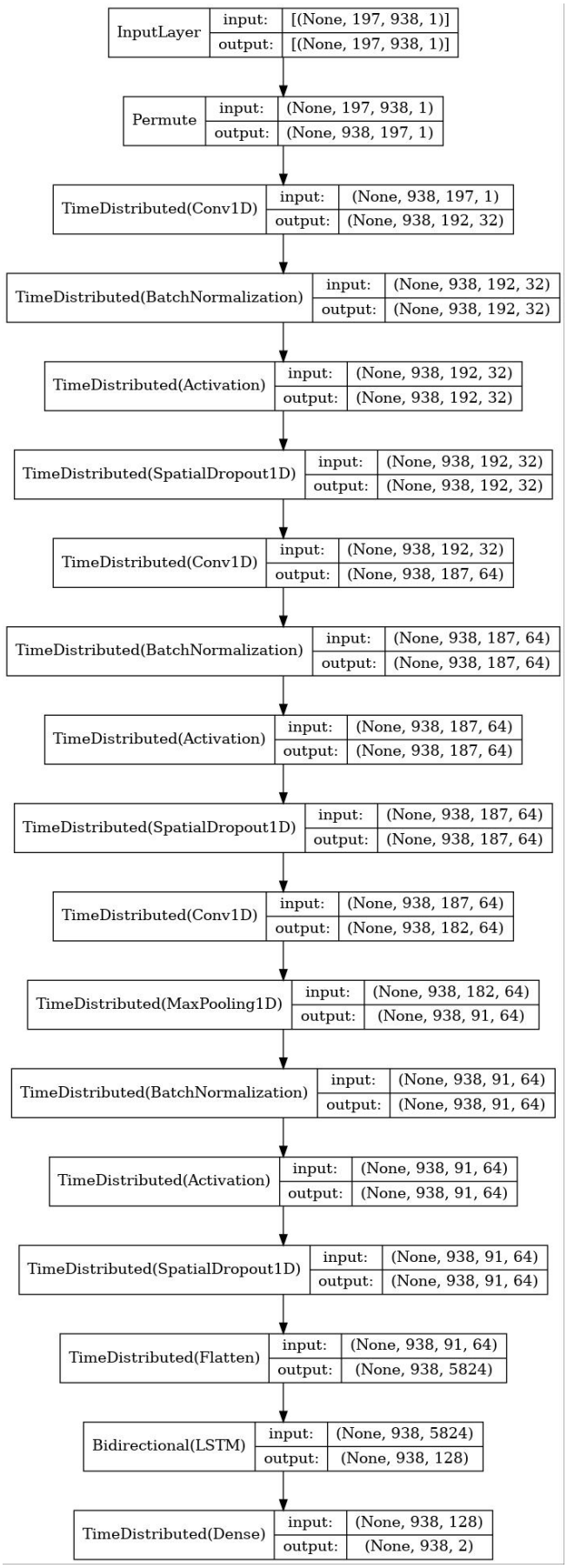


Figure 5.2: Replication of CNN-BiLSTM of [6]

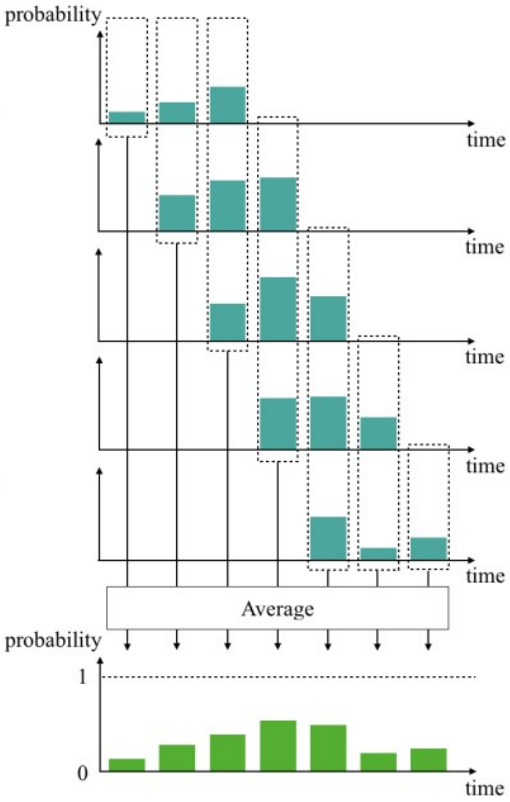


Figure 5.3: Beginning of post-processing for files with more than 15s [25]

Dataset	#Original Wheezes & Crackles /CAS & DAS	#Wheezes/CAS	#Crackles/DAS
HF_Lung_V1 Train	5856	2278	2306
HF_Lung_V1 Validation	1953	760	769
HF_Lung_V1 Test	1956	661	368
RSD Train	1445	669	714
RSD Validation	482	224	239
RSD Test	1176	511	469
RSD_AKGC417L Train	812	310	342
RSD_AKGC417L Validation	271	104	114
RSD_AKGC417L Test	855	291	333

Table 5.5: Distribution of annotated events in the training, validation and test sets before and after the removal of the files (#: Number of annotated events of)

HF_Lung_V1 Dataset

Table 5.6 displays the results obtained by the classifier on the test set for the 2-class problem (DAS vs. normal sounds). Table 5.7 displays the results obtained by the classifier on the test set for the 2-class problem (CAS vs. normal sounds).

	Recall	**Precision**	**F1-Score**
Class_0.25	**87.8**	51.5	**64.9**
Class_0.5	60.0	68.3	63.9
Class_0.75	19.4	**80.5**	31.3
Seg_JI_0.25	**48.2**	18.2	26.4
Seg_JI_0.5	28.7	**28.1**	**28.4**
Seg_JI_0.75	5.8	23.9	9.4
Seg_OC_0.25	**97.0**	30.9	46.9
Seg_OC_0.5	91.0	55.4	68.9
Seg_OC_0.75	69.1	**78.8**	**73.6**

Table 5.6: Performance results of CNN-BiLSTM on DAS vs. normal sounds problem using the HF_Lung_V1 (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, 0.25/0.5/0.75: Threshold)

	Recall	**Precision**	**F1-Score**
Class_0.25	**34.7**	65.0	**45.2**
Class_0.5	21.0	78.0	33.1
Class_0.75	12.1	**86.5**	21.3
Seg_JI_0.25	**23.2**	27.0	**25.0**
Seg_JI_0.5	17.0	45.0	24.6
Seg_JI_0.75	9.9	**54.2**	16.8
Seg_OC_0.25	**68.3**	52.1	**59.1**
Seg_OC_0.5	45.8	68.8	55.0
Seg_OC_0.75	31.4	**79.0**	44.9

Table 5.7: Performance results of CNN-BiLSTM on CAS vs. normal sounds problem using the HF_Lung_V1 (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, 0.25/0.5/0.75: Threshold)

When comparing the results obtained from the replication and the results attained in the article, the values that can be compared are *Class_0.5* and *Seg_JI_0.5*. Regarding the DAS vs. normal sounds, the results of the article are superior (F1-Score of 71.2% in *Class_0.5* and F1-Score of 70.8% in *Seg_JI_0.5*, 7.3% and 42.4%, respectively). Regarding the CAS vs. normal sounds, the results of the article are superior (F1-Score of 47.9% in *Class_0.5* and F1-Score of 46.4% in *Seg_JI_0.5*, 14.8% and 21.8%, respectively). As explained before, as this model was replicated with the available information from the article and still it was simplified due to computational limitations, the results are lower. Also, since the post-processing method has some simplifications since energy peaks are not analysed.

Regarding the DAS vs. normal sounds, the best results were attained when the threshold was 0.75 and using the OC (F1-Score of 73.6%). Concerning the CAS vs. normal sounds, the best results were attained when the threshold was 0.25 and using the OC (F1-Score of 59.1%).

RSD

Table 5.8 displays the results obtained by the classifier on the test set for the 2-class problem (crackles vs. normal sounds). Table 5.9 displays the results obtained by the classifier on the test set for the 2-class problem (wheezes vs. normal sounds).

	Recall	**Precision**	**F1-Score**
Class_Average_0.25	**65.7**	43.3	52.2
Class_Average_0.5	46.9	55.6	50.9
Class_Average_0.75	27.7	**66.1**	39.0
Class_Median_0.25	65.5	43.6	**52.4**
Class_Median_0.5	46.9	55.5	50.8
Class_Median_0.75	27.6	65.7	38.9
Seg_JI_Average_0.25	49.2	36.7	42.0
Seg_JI_Average_0.5	37.0	49.1	42.2
Seg_JI_Average_0.75	22.2	**57.6**	32.1
Seg_JI_Median_0.25	**50.9**	36.3	42.4
Seg_JI_Median_0.5	38.8	48.5	**43.1**
Seg_JI_Median_0.75	23.5	56.4	33.2
Seg_OC_Average_0.25	72.0	45.9	56.1
Seg_OC_Average_0.5	57.3	59.9	58.6
Seg_OC_Average_0.75	38.8	**70.3**	50.0
Seg_OC_Median_0.25	**74.7**	45.6	56.6
Seg_OC_Median_0.5	60.2	59.4	**59.8**
Seg_OC_Median_0.75	41.3	69.5	51.8

Table 5.8: Performance results of CNN-BiLSTM on crackles vs. normal sounds problem using the RSD (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, Average/Median: Post-processing method to aggregate the predictions for files longer that 15s, 0.25/0.5/0.75: Threshold)

After analysing the tables, we can understand that the results are reasonably good (a better performance when the OC is used), even before the post-processing in the classification of the individual frames. Regarding the crackles vs. normal sounds, the best results were attained when the threshold was 0.5, using the OC and the median to aggregate the chunks (F1-Score of 59.8%), even though the difference to the same model but using the average is small (F1-Score of 58.6%, 1.2% difference). Concerning the wheezes vs. normal sounds, the best results were attained when the threshold was 0.25, using the OC and the average to aggregate the chunks (F1-Score of 45.8%), but once again the difference to the same model but using the median is small (F1-Score of 45.0%, 0.8% difference). This advantage for the crackles vs. normal sounds can be explained by the larger training set that it has, since a larger dataset benefits better this type of model.

	Recall	**Precision**	**F1-Score**
Class_Average_0.25	**51.1**	28.0	36.1
Class_Average_0.5	34.6	28.2	31.1
Class_Average_0.75	21.2	27.4	23.9
Class_Median_0.25	50.8	28.3	**36.4**
Class_Median_0.5	34.7	**28.7**	31.4
Class_Median_0.75	21.5	28.1	24.3
Seg_JI_Average_0.25	16.2	10.4	12.7
Seg_JI_Average_0.5	13.2	**11.5**	12.3
Seg_JI_Average_0.75	9.4	9.7	9.5
Seg_JI_Median_0.25	**17.7**	10.0	12.7
Seg_JI_Median_0.5	15.1	11.2	**12.9**
Seg_JI_Median_0.75	10.8	9.4	10.1
Seg_OC_Average_0.25	71.0	33.8	**45.8**
Seg_OC_Average_0.5	55.4	**35.2**	43.0
Seg_OC_Average_0.75	40.0	31.5	35.3
Seg_OC_Median_0.25	**75.4**	32.1	45.0
Seg_OC_Median_0.5	64.4	34.9	45.3
Seg_OC_Median_0.75	48.9	32.0	38.7

Table 5.9: Performance results of CNN-BiLSTM on wheezes vs. normal sounds problem using the RSD (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, Average/Median: Post-processing method to aggregate the predictions for files longer that 15s, 0.25/0.5/0.75: Threshold)

RSD tested on files that were recorded with the AKGC417L microphone

In the previous chapter, a stratified analysis was performed to better understand which demographic/equipment achieved better results. Regarding the equipment, the AKGC417L microphone was the one that achieved the better results (especially in the 2-class problem crackles vs. normal sounds where the highest F1-Score was achieved), due to its higher sensibility than any other equipment and less filtering, but also due to the larger training and test sets. For that reason and since it is the most important category in the available stratification, a CNN-BiLSTM model was tested only using the files that were recorded with the AKGC417L microphone.

Table 5.10 displays the results obtained by the classifier on the test set for the 2-class problem (crackles vs. normal sounds). Table 5.11 displays the results obtained by the classifier on the test set for the 2-class problem (wheezes vs. normal sounds).

	Recall	Precision	F1-Score
Class_Average_0.25	**65.0**	53.6	**58.8**
Class_Average_0.5	46.8	67.0	55.1
Class_Average_0.75	28.4	**78.1**	41.7
Class_Median_0.25	64.9	53.7	**58.8**
Class_Median_0.5	46.9	67.1	55.2
Class_Median_0.75	28.4	**78.1**	41.7
Seg_JI_Average_0.25	**52.2**	45.5	**48.6**
Seg_JI_Average_0.5	40.1	61.1	48.4
Seg_JI_Average_0.75	26.0	**72.1**	38.2
Seg_JI_Median_0.25	51.9	45.4	48.5
Seg_JI_Median_0.5	40.0	61.1	48.3
Seg_JI_Median_0.75	25.9	**72.1**	38.1
Seg_OC_Average_0.25	**74.8**	54.5	63.1
Seg_OC_Average_0.5	61.7	70.7	65.9
Seg_OC_Average_0.75	43.6	**81.2**	56.8
Seg_OC_Median_0.25	**74.8**	54.5	63.0
Seg_OC_Median_0.5	61.7	70.8	**66.0**
Seg_OC_Median_0.75	43.5	**81.2**	56.6

Table 5.10: Performance results of CNN-BiLSTM on crackles vs. normal sounds problem tested only on the RSD files recorded with the AKGC417L microphone (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, Average/Median: Post-processing method to aggregate the predictions for files longer that 15s, 0.25/0.5/0.75: Threshold)

	Recall	Precision	F1-Score
Class_Average_0.25	**43.1**	53.6	**47.8**
Class_Average_0.5	27.4	57.5	37.1
Class_Average_0.75	15.9	56.3	24.7
Class_Median_0.25	42.9	53.6	47.7
Class_Median_0.5	27.4	**57.6**	37.2
Class_Median_0.75	15.8	56.4	24.7
Seg_JI_Average_0.25	**21.6**	20.5	**21.0**
Seg_JI_Average_0.5	14.4	**23.8**	17.9
Seg_JI_Average_0.75	8.8	23.1	12.8
Seg_JI_Median_0.25	21.5	20.4	**21.0**
Seg_JI_Median_0.5	14.4	**23.8**	17.9
Seg_JI_Median_0.75	8.9	23.2	12.8
Seg_OC_Average_0.25	**71.5**	46.0	56.0
Seg_OC_Average_0.5	58.5	55.9	**57.2**
Seg_OC_Average_0.75	40.9	**58.1**	48.0
Seg_OC_Median_0.25	71.2	46.0	55.9
Seg_OC_Median_0.5	58.5	55.9	57.1
Seg_OC_Median_0.75	40.7	**58.1**	47.9

Table 5.11: Performance results of CNN-BiLSTM on wheezes vs. normal sounds problem tested only on the RSD files recorded with the AKGC417L microphone (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, Average/Median: Post-processing method to aggregate the predictions for files longer that 15s, 0.25/0.5/0.75: Threshold)

After analysing the tables, we can understand that the results are good (a better performance when the OC is used), even before the post-processing in the classification of the individual frames. Regarding the crackles vs. normal sounds, the best results were attained when the threshold was 0.5, using the OC and the median to aggregate the chunks (F1-Score of 66.0%), even though the difference to the same model but using the average is almost zero (F1-Score of 65.9%, 0.1% difference). Concerning the wheezes vs. normal sounds, the best results were attained when the threshold was 0.5, using the OC and the average to aggregate the chunks (F1-Score of 57.2%), but once again the difference to the same model but using the median is almost zero (F1-Score of 57.1%, 0.1% difference). This advantage for the crackles vs. normal sounds can be explained by the larger training set that it has, since a larger dataset benefits better this type of model.

RSD only using the files that were recorded with the AKGC417L microphone

In the previous chapter, a stratified analysis was performed to better understand which demographic/equipment achieved better results. Regarding the equipment, the AKGC417L microphone was the one that achieved the better results (especially in the 2-class problem crackles vs. normal sounds where the highest F1-Score was achieved), due to its higher sensibility than any other equipment and less filtering, but also due to the larger training and test sets. For that reason and since it is the most important category in the available stratification, a CNN-BiLSTM model was trained and tested only using the files that were recorded with the AKGC417L microphone.

Table 5.12 displays the results obtained by the classifier on the test set for the 2-class problem (crackles vs. normal sounds). Table 5.13 displays the results obtained by the classifier on the test set for the 2-class problem (wheezes vs. normal sounds).

After analysing the tables, we can understand that the results are reasonably good (a better performance when the OC is used), even before the post-processing in the classification of the individual frames. Regarding the crackles vs. normal sounds, the best results were attained when the threshold was 0.5, using the OC and the average to aggregate the chunks (F1-Score of 66.4%), even though the difference to the same model but using the median is almost zero (F1-Score of 66.3%, 0.1% difference). Concerning the wheezes vs. normal sounds, the best results were attained when the threshold was 0.5, using the OC and the median to aggregate the chunks (F1-Score of 58.6%), even though the difference to the same model but using the average is almost zero (F1-Score of 58.4%, 0.2% difference). This difference between crackles vs. normal sounds and wheezes vs. normal sounds results can be explained by the larger train, validation and test sets.

Also, these **results are slightly higher** than the results of the model trained with RSD and tested with the RSD files recorded with the AKGC417L microphone.

	Recall	Precision	F1-Score
Class_Average_0.25	**66.0**	54.4	**59.6**
Class_Average_0.5	47.2	69.6	56.3
Class_Average_0.75	26.6	**79.8**	39.9
Class_Median_0.25	**66.0**	54.4	**59.6**
Class_Median_0.5	47.3	69.5	56.3
Class_Median_0.75	26.8	**79.9**	40.2
Seg_JI_Average_0.25	54.7	44.0	48.7
Seg_JI_Average_0.5	41.4	61.6	**49.5**
Seg_JI_Average_0.75	25.2	73.2	37.5
Seg_JI_Median_0.25	**54.9**	44.0	48.9
Seg_JI_Median_0.5	41.4	61.6	**49.5**
Seg_JI_Median_0.75	25.4	**73.3**	37.7
Seg_OC_Average_0.25	**77.2**	52.6	62.5
Seg_OC_Average_0.5	62.5	70.8	**66.4**
Seg_OC_Average_0.75	41.3	81.7	54.9
Seg_OC_Median_0.25	77.1	52.5	62.5
Seg_OC_Median_0.5	62.4	70.8	66.3
Seg_OC_Median_0.75	41.6	**81.8**	55.2

Table 5.12: Performance results of CNN-BiLSTM on crackles vs. normal sounds problem using the RSD_AKGC417L (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, Average/Median: Post-processing method to aggregate the predictions for files longer that 15s, 0.25/0.5/0.75: Threshold)

	Recall	Precision	F1-Score
Class_Average_0.25	34.8	56.0	42.9
Class_Average_0.5	23.5	60.7	33.9
Class_Average_0.75	14.4	60.7	23.3
Class_Median_0.25	**34.9**	56.1	**43.0**
Class_Median_0.5	23.6	**60.8**	33.9
Class_Median_0.75	14.4	60.7	23.3
Seg_JI_Average_0.25	16.0	15.1	**15.6**
Seg_JI_Average_0.5	10.7	18.7	13.6
Seg_JI_Average_0.75	7.5	18.4	10.7
Seg_JI_Median_0.25	**16.1**	15.2	**15.6**
Seg_JI_Median_0.5	10.6	**18.8**	13.5
Seg_JI_Median_0.75	7.5	18.4	10.6
Seg_OC_Average_0.25	**70.5**	44.0	54.2
Seg_OC_Average_0.5	60.3	56.6	58.4
Seg_OC_Average_0.75	49.4	**59.8**	54.1
Seg_OC_Median_0.25	**70.5**	44.1	54.3
Seg_OC_Median_0.5	60.5	56.9	**58.6**
Seg_OC_Median_0.75	49.4	59.7	54.1

Table 5.13: Performance results of CNN-BiLSTM on wheezes vs. normal sounds problem using the RSD_AKGC417L (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, Average/Median: Post-processing method to aggregate the predictions for files longer that 15s, 0.25/0.5/0.75: Threshold)

Comparison between RSD and HF_Lung_V1 datasets

Tables 5.14 and 5.15 display the results obtained by the model when trained with the RSD and tested with HF_Lung_V1 dataset for both 2-class problems. Tables 5.16 and 5.17 display the results obtained by the model when trained with the RSD and tested with HF_Lung_V1 dataset for both 2-class problems.

	Recall	**Precision**	**F1-Score**
Class_0.25	**13.8**	**43.7**	**21.0**
Class_0.5	6.8	40.5	11.7
Class_0.75	2.7	35.1	5.0
Seg_JI_0.25	**0.1**	**0.1**	**0.1**
Seg_JI_0.5	0.0	0.0	0.0
Seg_JI_0.75	0.0	0.0	0.0
Seg_OC_0.25	**94.2**	**43.0**	**59.0**
Seg_OC_0.5	82.8	40.6	54.5
Seg_OC_0.75	57.8	36.8	45.0

Table 5.14: Performance results obtained with 2-class problem (DAS/crackles vs. normal sounds) with the model trained with the RSD and tested with HF_Lung_V1 (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, 0.25/0.5/0.75: Threshold)

	Recall	**Precision**	**F1-Score**
Class_0.25	**57.4**	39.9	**47.1**
Class_0.5	37.9	45.9	41.5
Class_0.75	21.4	**52.8**	30.4
Seg_JI_0.25	**25.3**	**10.8**	**15.1**
Seg_JI_0.5	17.4	**10.8**	13.3
Seg_JI_0.75	8.9	9.1	9.0
Seg_OC_0.25	**90.8**	30.3	45.5
Seg_OC_0.5	83.1	36.6	50.8
Seg_OC_0.75	68.5	**43.6**	**53.3**

Table 5.15: Performance results obtained with 2-class problem (CAS/wheezes vs. normal sounds) with the model trained with the RSD and tested with HF_Lung_V1 (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, 0.25/0.5/0.75: Threshold)

	Recall	Precision	F1-Score
Class_Average_0.25	**75.4**	7.7	14.0
Class_Average_0.5	45.4	**8.9**	**14.8**
Class_Average_0.75	8.7	5.4	6.6
Class_Median_0.25	75.3	7.7	14.0
Class_Median_0.5	45.6	8.8	**14.8**
Class_Median_0.75	9.3	5.4	6.8
Seg_JI_Average_0.25	**3.3**	1.8	**2.3**
Seg_JI_Average_0.5	1.8	2.1	2.0
Seg_JI_Average_0.75	1.0	2.0	1.3
Seg_JI_Median_0.25	3.2	1.5	2.1
Seg_JI_Median_0.5	2.2	**2.3**	2.2
Seg_JI_Median_0.75	0.9	1.7	1.2
Seg_OC_Average_0.25	41.6	18.7	**25.8**
Seg_OC_Average_0.5	20.3	**19.3**	19.8
Seg_OC_Average_0.75	7.7	13.6	9.8
Seg_OC_Median_0.25	**44.7**	17.8	25.4
Seg_OC_Median_0.5	22.5	**19.3**	20.7
Seg_OC_Median_0.75	8.6	13.5	10.5

Table 5.16: Performance results obtained with 2-class problem (DAS/crackles vs. normal sounds) with the model trained with the HF_Lung_V1 and tested with RSD (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, Average/Median: Post-processing method to aggregate the predictions for files longer that 15s, 0.25/0.5/0.75: Threshold)

	Recall	Precision	F1-Score
Class_Average_0.25	**45.8**	40.3	**42.9**
Class_Average_0.5	26.4	45.9	33.5
Class_Average_0.75	13.3	51.1	21.1
Class_Median_0.25	45.4	40.4	42.8
Class_Median_0.5	26.3	46.3	33.5
Class_Median_0.75	13.1	**51.2**	20.9
Seg_JI_Average_0.25	16.8	17.1	17.0
Seg_JI_Average_0.5	10.4	**21.6**	14.0
Seg_JI_Average_0.75	4.8	18.1	7.6
Seg_JI_Median_0.25	**20.1**	16.6	**18.2**
Seg_JI_Median_0.5	12.3	21.1	15.6
Seg_JI_Median_0.75	5.8	18.2	8.8
Seg_OC_Average_0.25	47.1	36.6	41.2
Seg_OC_Average_0.5	30.5	44.8	36.3
Seg_OC_Average_0.75	20.4	48.1	28.6
Seg_OC_Median_0.25	**55.8**	35.6	**43.4**
Seg_OC_Median_0.5	38.4	45.4	41.6
Seg_OC_Median_0.75	25.2	**49.2**	33.4

Table 5.17: Performance results obtained with 2-class problem (CAS/wheezes vs. normal sounds) with the model trained with the HF_Lung_V1 and tested with RSD (Class: Classification, Seg_JI: Segmentation with JI, Seg_OC: Segmentation with OC, Average/Median: Post-processing method to aggregate the predictions for files longer that 15s, 0.25/0.5/0.75: Threshold)

As already explained, the events in the HF_Lung_V1 dataset are annotated in the same way as the respiratory cycles, meaning the crackles, which have a short duration, are having a longer duration. Overall, when the models were trained with the RSD and tested with HF_Lung_V1, the results were higher for both 2-class problems than when the opposite was tested. As expected, the best results for all combinations were achieved with the OC. For the results of the model trained with the RSD and tested with HF_Lung_V1, the results on the 2-class problem DAS/crackles vs. normal sounds, the best result was achieved with a 0.25 threshold with 59.0% of F1-Score, while the best results on the other 2-class problem were achieved with a 0.75 threshold with 53.3% of F1-Score, but when the models were trained with the HF_Lung_V1 and tested with RSD, the best result was achieved in the 2-class problem CAS/wheezes vs. normal sounds using a 0.5 threshold with median, with 43.4% of F1-Score, while the other 2-class problem achieved using a 0.25 threshold with average, with 25.8% of F1-Score. Regarding the DAS/crackles vs. normal sounds, the results are better than the CAS/wheezes vs. normal sounds, but **none of these models performed well enough to consider them good at generalising** for anything dataset, especially due to the different types of annotations.

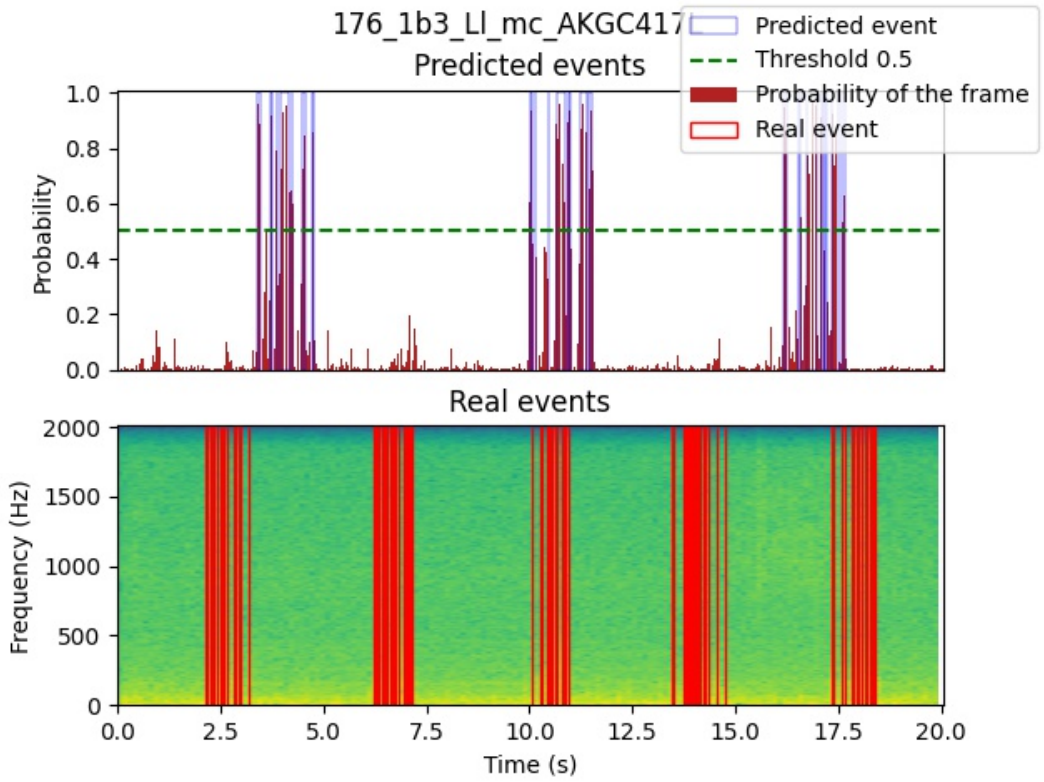
Comparison between RSD and RSD_AKGC417L

When comparing Tables 5.8 and 5.12 and their best results, using the RSD_AKGC417L on the 2-class problem crackles vs. normal sounds is better. The best model using the full RSD achieved an F1-Score of 59.8%, while the best model using the RSD_AKGC417L achieved an F1-Score of 66.4%, with a 6.6% difference. Figure 5.4 shows the output of both models. When comparing Tables 5.9 and 5.13 and their best results, using the RSD_AKGC417L on the 2-class problem wheezes vs. normal sounds is better. The best model using the full RSD achieved an F1-Score of 45.8%, while the best model using the RSD_AKGC417L achieved an F1-Score of 58.6%, with a 12.8% difference. Figure 5.5 shows the output of both models.

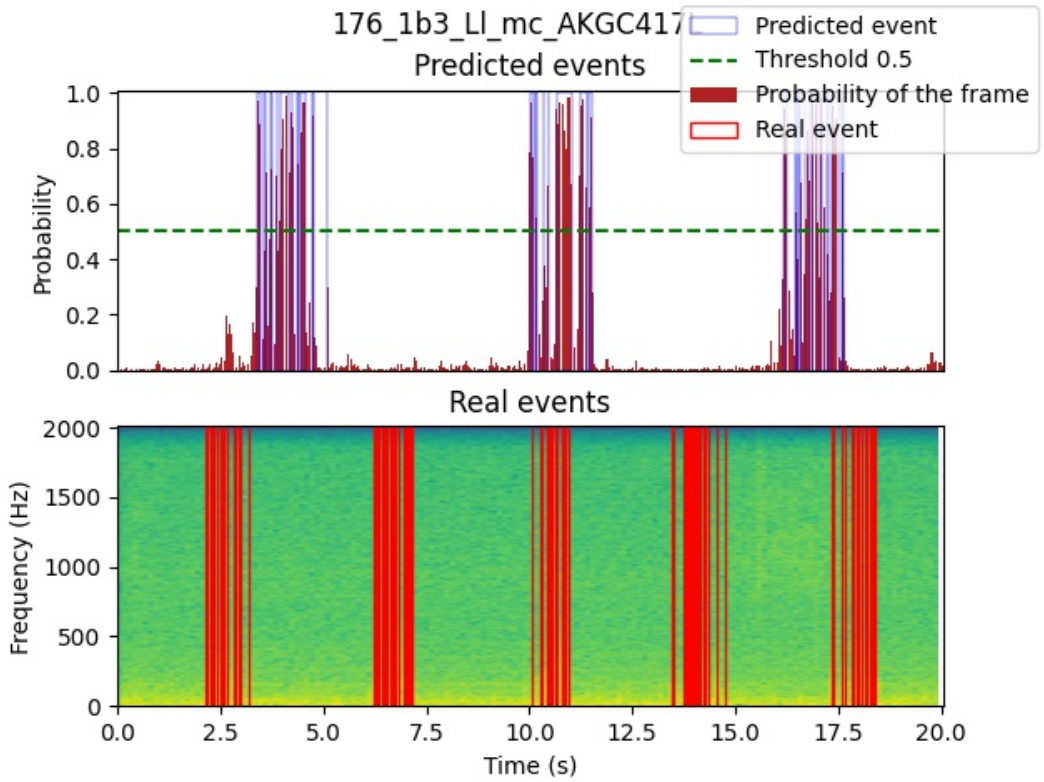
This overall advantage for the RSD_AKGC417L microphone can be explained by the quality of the recordings, even though the train, validation and test sets are smaller than the RSD. Also, given the lower number of recordings with the others stethoscopes, as these type of models require larger amount of data to correctly learn the characteristics of those recordings, the model cannot produce good results. With this experiment, it was proved that **the files recorded using the AKGC417L microphone in the RSD can produce better results due to the higher quality of the recordings**, meaning the number of samples in the training set of each recording is important to be large.

5.4 Comparison between both approaches

When comparing both approaches, only the segmentation metric (JI or OC) and thresholds (0.25, 0.5 and 0.75) can be compared, since these parameters are the only thing in common between both approaches. Starting by comparing the

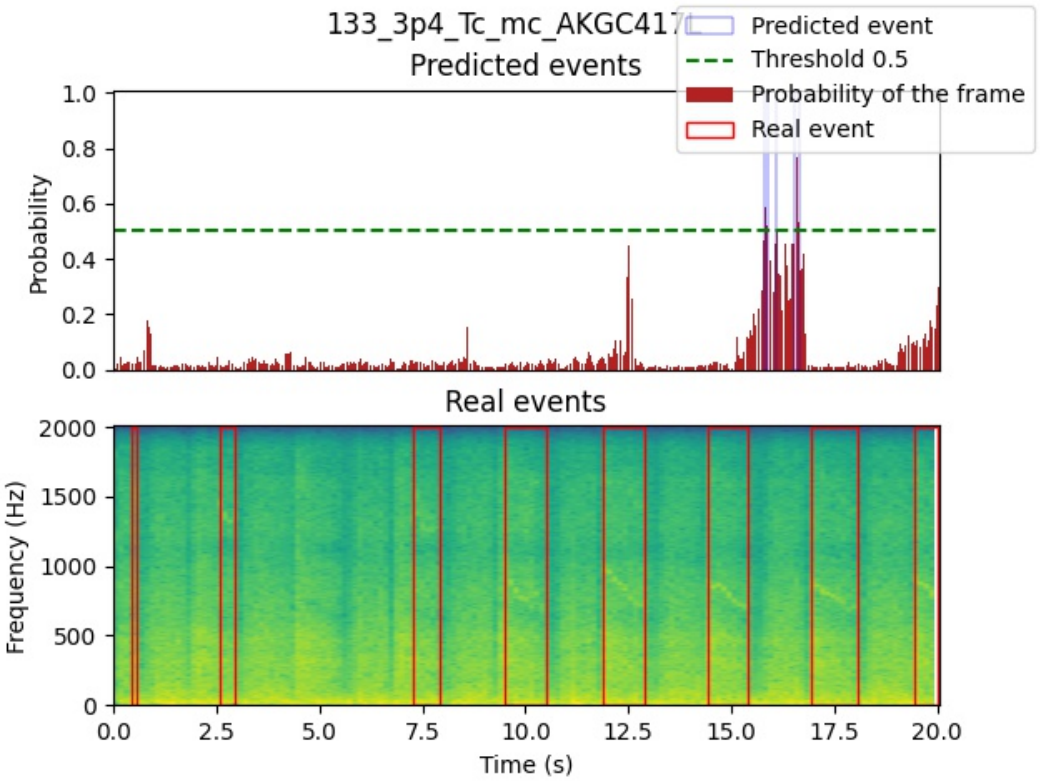


(a) Best model output for CNN-BiLSTM with average as aggregation of chunks method and threshold of 0.5 using the RSD_AKGC417L

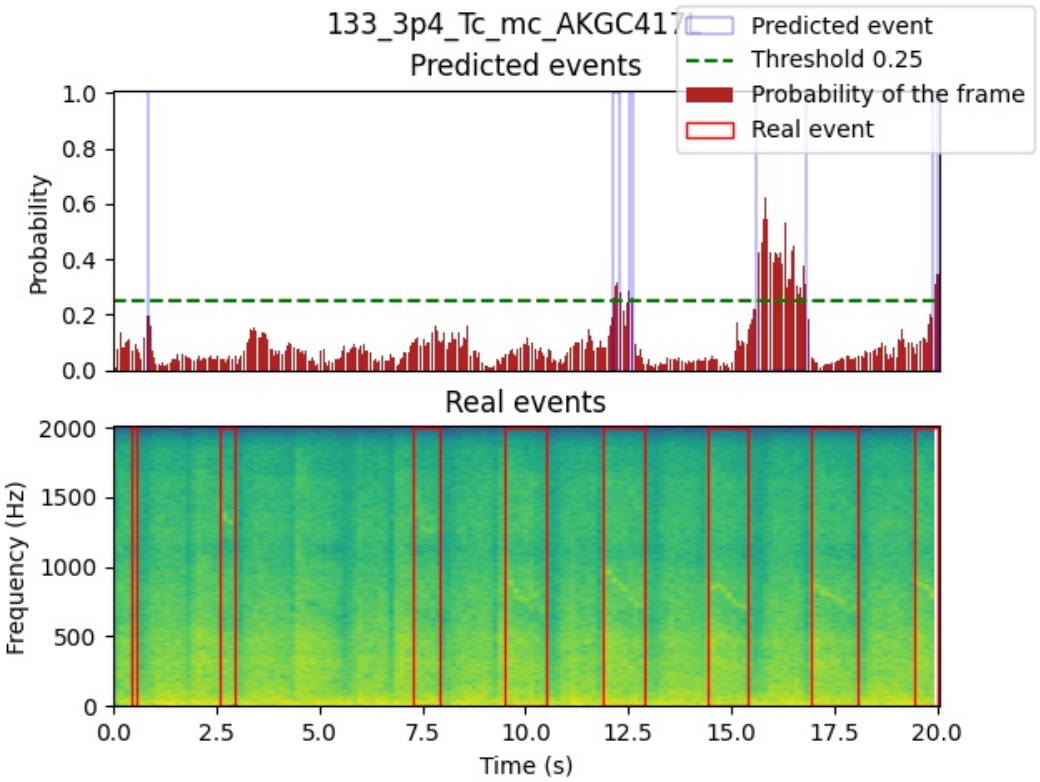


(b) Best model output for CNN-BiLSTM with median as aggregation of chunks method and threshold of 0.5 using the RSD

Figure 5.4: Best models output on crackles vs. normal sounds using the RSD_AKGC417L and RSD datasets



(a) Best model output for CNN-BiLSTM with median as aggregation of chunks method and threshold of 0.5 using the RSD_AKGC417L



(b) Best model output for CNN-BiLSTM with average as aggregation of chunks method and threshold of 0.25 using the RSD

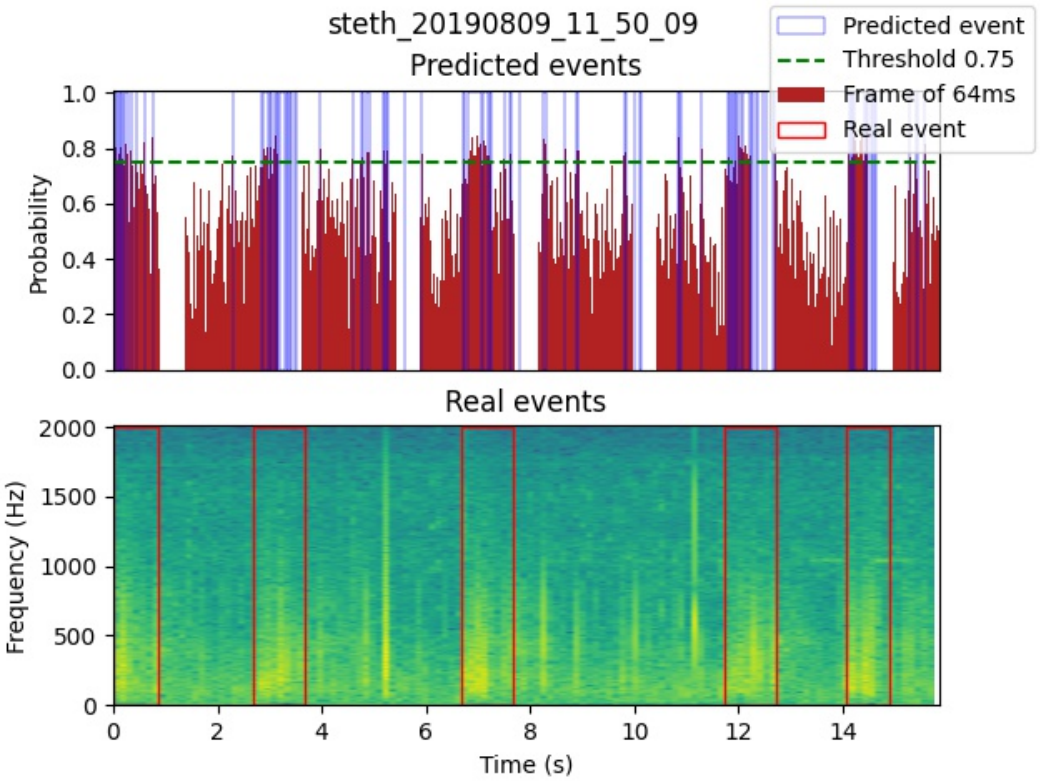
Figure 5.5: Best models output on wheezes vs. normal sounds using the RSD_AKGC417L and RSD datasets

results of the CNN and CNN-BiLSTM on CAS vs. normal sounds using the HF_Lung_V1, the best results on both approaches was using the OC, but the CNN achieved higher results using the 0.75 threshold since this model classifies almost every frame as having an event. Figure 5.6 shows a graphical representation of the output for the same file of the best combination of parameters for the models of each approach. Figure 5.6a shows that quite a lot of the frames are classified as having an event, while Figure 5.6b shows that most of the frames have a lower probability of having an event, and some of the real events are not detected (reducing the number of FN, increasing the Recall, and consequently, the F1-Score), but when the model thinks there is an event presented, it has a high degree of confidence.

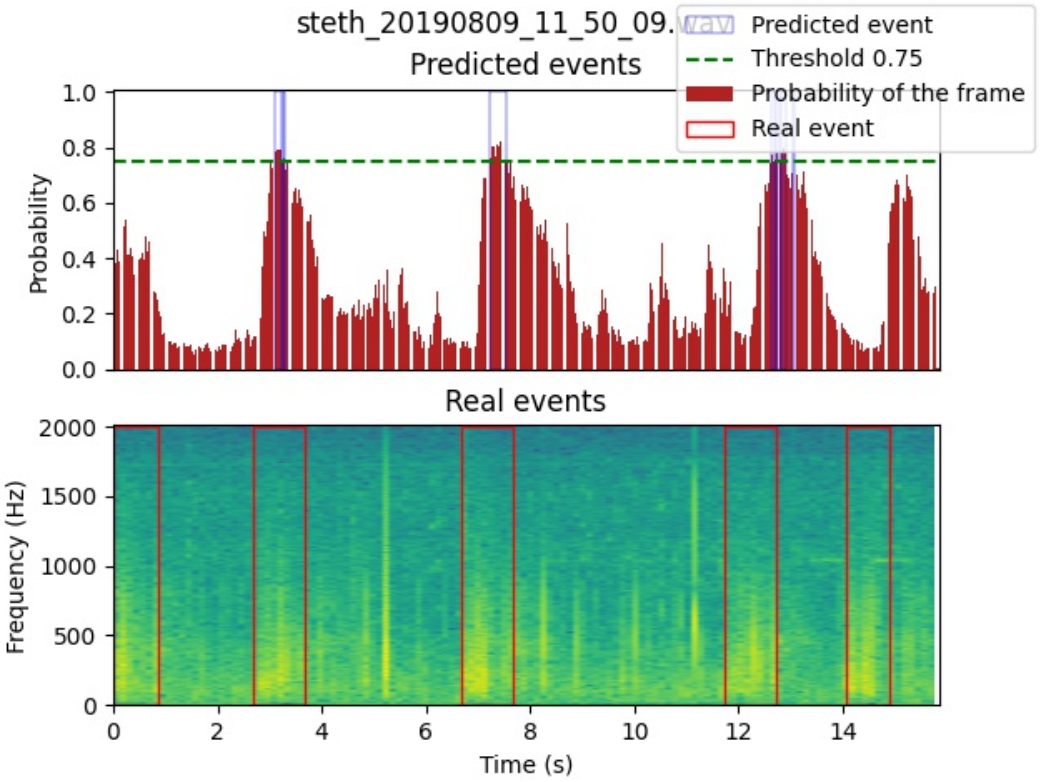
When comparing the results of the CNN and CNN-BiLSTM on CAS vs. normal sounds using the HF_Lung_V1, the best models on both approaches is also using the OC, but once again, the CNN achieved higher results using the 0.75 threshold, since this model classifies almost every frame as having an event. Figure 5.7 shows a graphical representation of the output for the same file of the best combination of parameters for the models of each approach. Figure 5.7a shows that quite a lot of the frames are classified as having an event, while Figure 5.7b shows that most of the frames have a lower probability of having an event, and some of the real events are not detected (reducing the number of FN, increasing the Recall, and consequently, the F1-Score), but when the model thinks there is an event presented, it has a high degree of confidence.

The next comparison is between the results of the CNN and CNN-BiLSTM on crackles vs. normal sounds using the RSD. In this case, the best results on both approaches were attained with the OC and using a 0.5 threshold and the CNN achieved higher results than the CNN-BiLSTM (F1-Score of 26.8% vs. F1-Score of 22.3%, respectively). Figure 5.8 shows a graphical representation of the output for the same file of the best combination of parameters for the models of each approach. Figure 5.8a shows that quite a lot of the frames are classified as having an event, while Figure 5.8b shows that most of the frames have a lower probability of having an event, and some of the real events are not detected (reducing the number of FN, increasing the Recall, and consequently, the F1-Score), but when the model thinks there is an event presented, it has a high degree of confidence.

Finally, the last comparison is between the results of the CNN and CNN-BiLSTM on wheezes vs. normal sounds using the RSD. The best combination of parameters was using the OC but 0.5 threshold in the CNN-BiLSTM and 0.75 in the CNN, as this last model classifies almost every frame as having an event. Figure 5.9 shows a graphical representation of the output for the same file of the best combination of parameters for the models of each approach. Figure 5.9a shows that quite a lot of the frames are classified as having an event, while Figure 5.9b shows that most of the frames have a lower probability of having an event, and some of the real events are not detected (reducing the number of FN, increasing the Recall, and consequently, the F1-Score), but when the model thinks there is an event presented, it has a high degree of confidence.

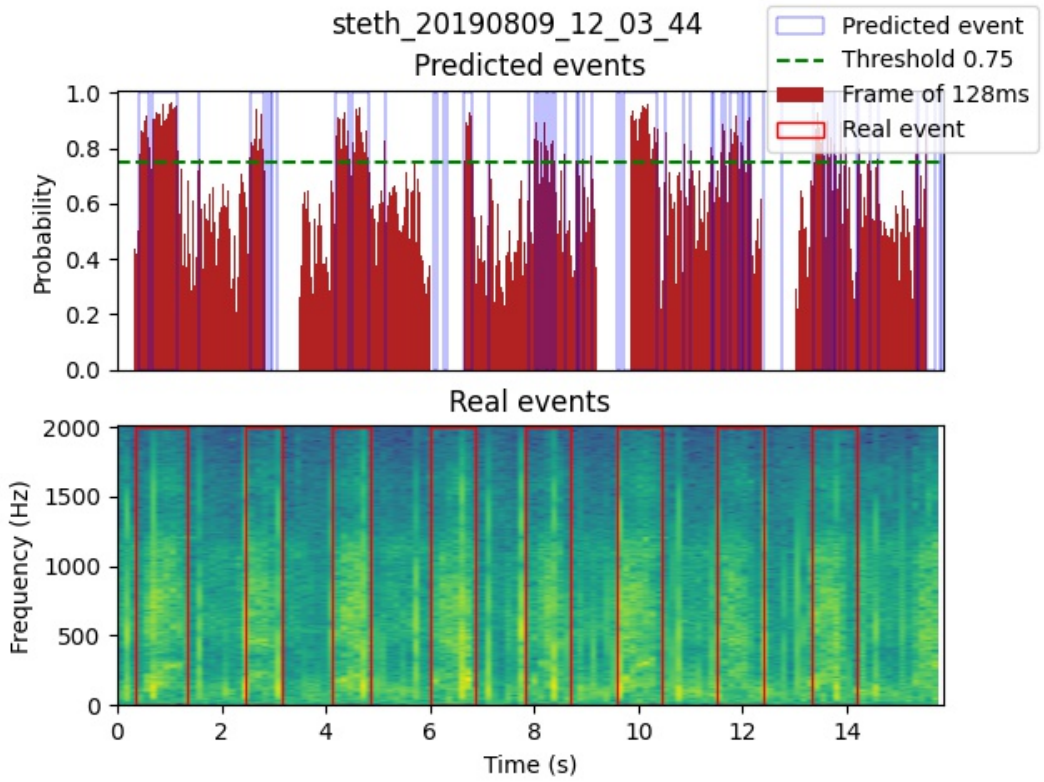


(a) Best model output for CNN with mel-spectrogram as input, threshold of 0.75 and window size of 64ms

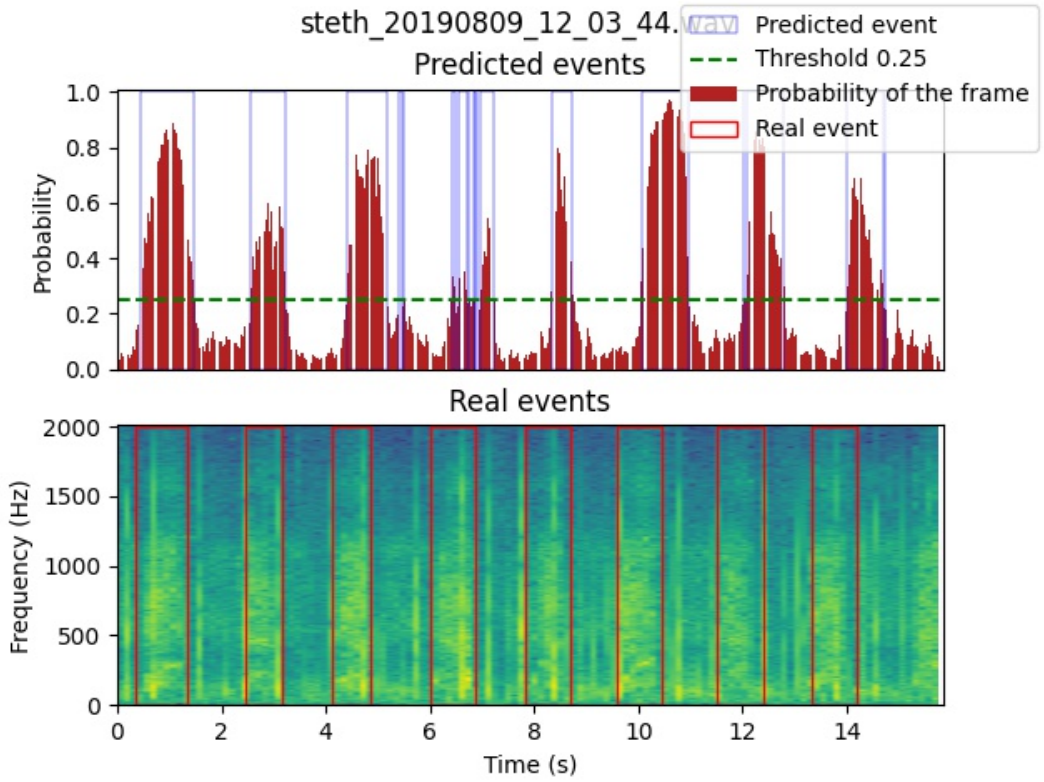


(b) Best model output for CNN-BiLSTM with threshold of 0.75

Figure 5.6: Best models output on DAS vs. normal sounds using the HF_Lung_V1

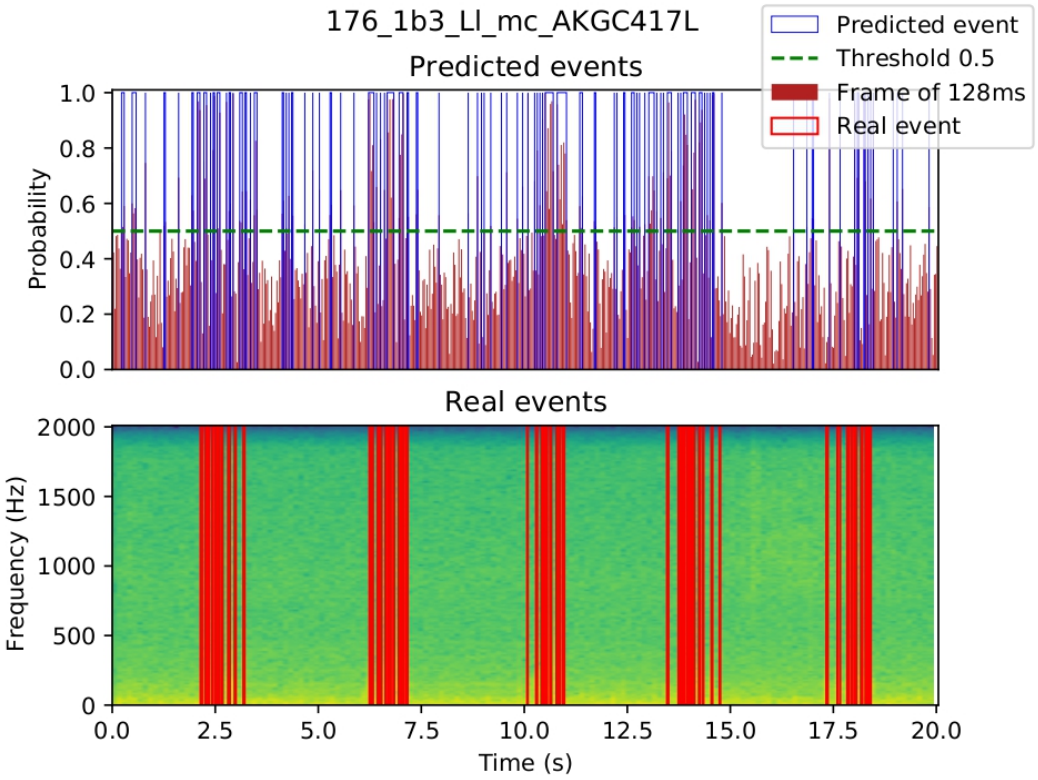


(a) Best model output for CNN with spectrogram as input, threshold of 0.75 and window size of 128ms

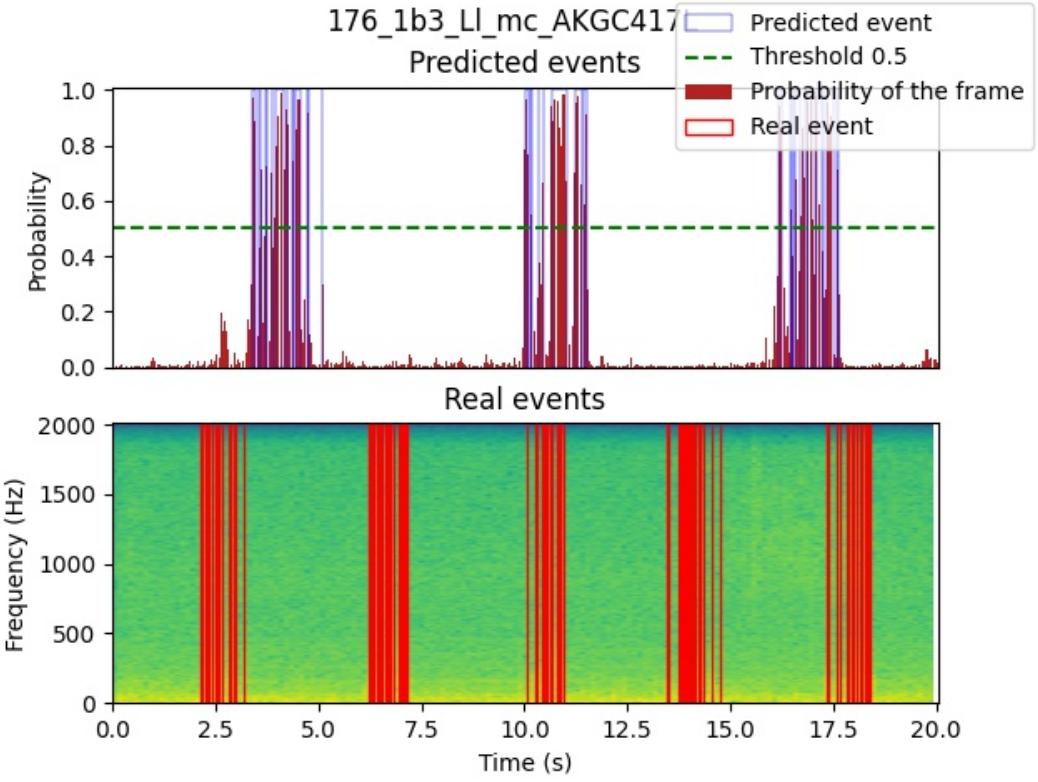


(b) Best model output for CNN-BiLSTM with threshold of 0.25

Figure 5.7: Best models output on CAS vs. normal sounds using the HF_Lung_V1

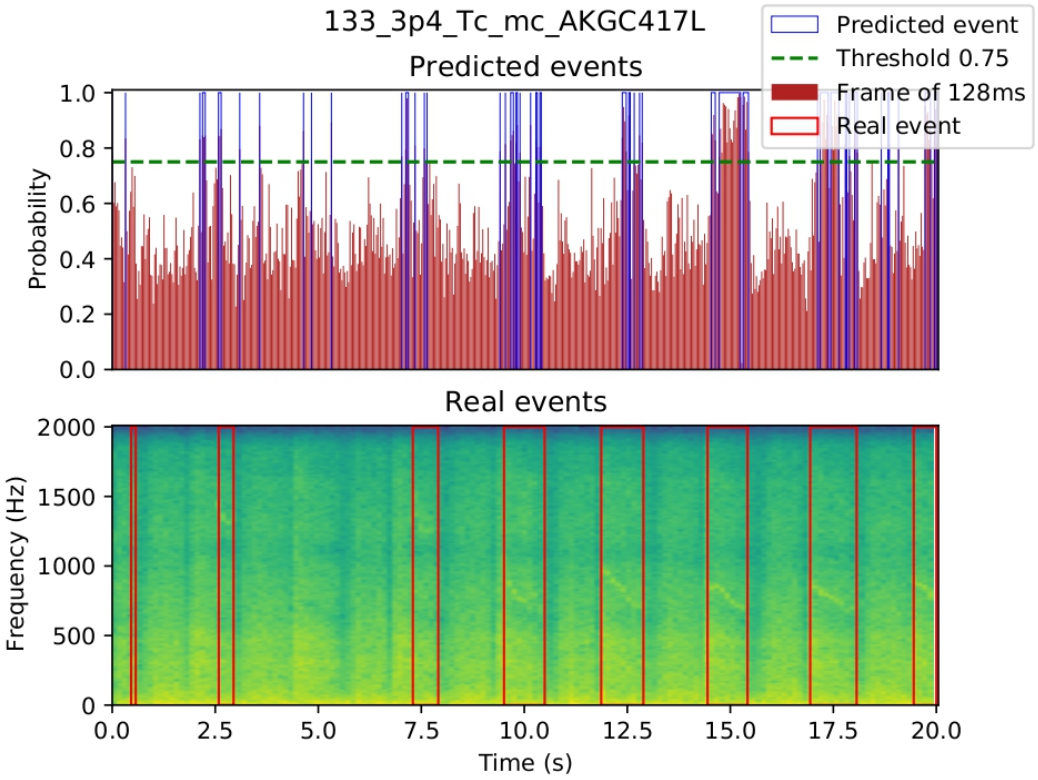


(a) Best model output for CNN with spectrogram as input, threshold of 0.5 and window size of 128ms

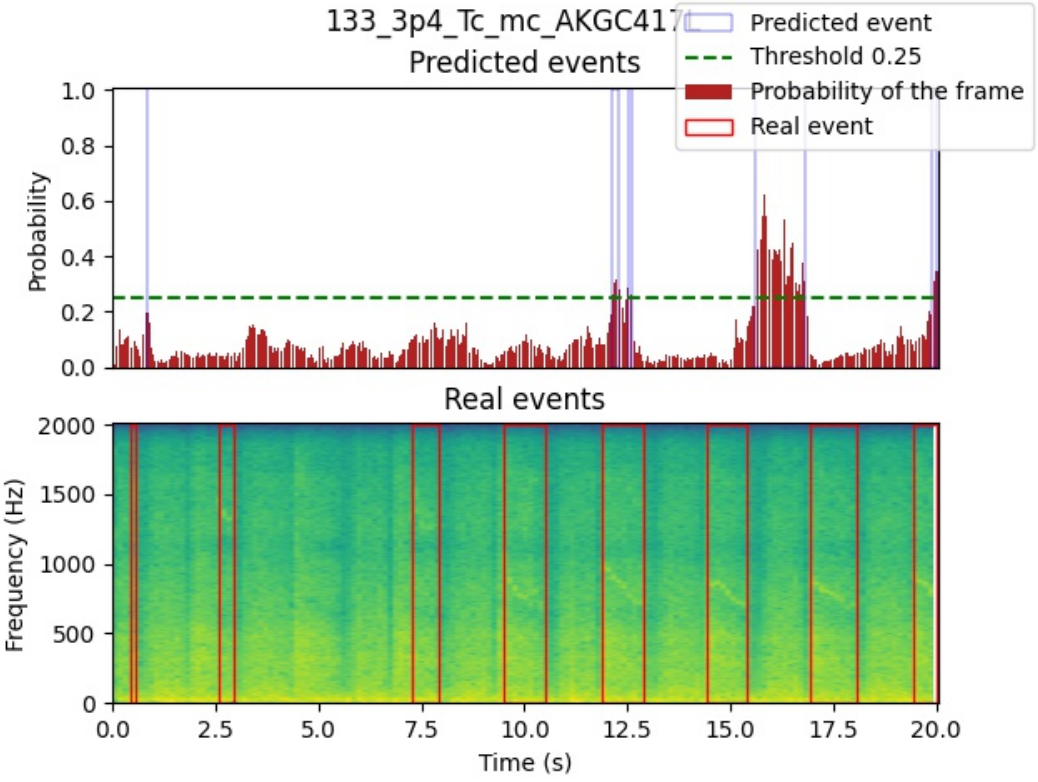


(b) Best model output for CNN-BiLSTM with median as aggregation of chunks method and threshold of 0.5

Figure 5.8: Best models output on crackles vs. normal sounds using the RSD



(a) Best model output for CNN with spectrogram as input, threshold of 0.75 and window size of 128ms



(b) Best model output for CNN-BiLSTM with average as aggregation of chunks method and threshold of 0.25

Figure 5.9: Best models output on wheezes vs. normal sounds using the RSD

Chapter 6

Conclusions and Future Work

This chapter concludes the thesis, summarising all the work done so far, as well as proposing suggestions for possible future work in this field of study.

6.1 Conclusions

Respiratory pathologies are a big concern within the medical community. This thesis aims to apply deep learning DL approaches to segment and classify adventitious respiratory sounds ARS. Since the final goal is quite broad, the thesis was divided into major problems: classification of ARS and segmentation of ARS.

There are more studies regarding the classification of ARS than the segmentation of ARS. All the studies presented in Chapter 3 have at least one usage of a DL approach, which suggests the power that these type of models can achieve.

Two datasets were employed throughout this thesis: RSD and HF_Lung_V1 (three if we consider the RSD New Annotations, although it is quite similar to the RSD). These datasets have different advantages in each of them, one has a large amount of data, while the other has more information regarding the demographic information, meaning a more detailed analysis can be performed (and was performed, with the acceptance of the article "Classification of Adventitious Respiratory Sound Events: A Stratified Analysis"). Both datasets were very important for this thesis but both have some negative aspect that is going to be addressed in the following section since it is something that can be improved for possible future work.

Regarding the classification of ARS, multiple models were trained. The classical Machine Learning approaches, even though they are out of the scope of this thesis, were important to help understand how the DL models can achieve better results. Overall, CNNs achieve the best results, but are not good enough to be used in a real scenario, as the results can be improved to be more accurate. The stratified analysis performed in the RSD demonstrated that some characteristics help achieve better results in this task, such as sounds recorded with the AKGC417L microphone, as well as the recordings of Male and Normal BMI subjects.

Concerning the segmentation of ARS, only DL models were trained, with 2 different approaches. Performing segmentation of ARS with the classification of individual frames proved to be a weak approach, since each frame is evaluated individually, without checking memory information, e.g., any relation between previous and next frames. Therefore, so it was used as a baseline to compare with the model used in the second approach. In contrast, the segmentation of ARS with sequential frames achieved better results, as in this approach, the relation between the past and the future is taken into account. Since it was created taking into account the longer annotations of the HF_Lung_V1 dataset, shorter events such as the crackles of the RSD had some difficulties in correctly classifying them. For every model tested, the segmentation metric that achieved the best results was always the OC, due to its less strict nature. The influence of the threshold depends on the average value of the predictions for each frame, i.e., if the model classifies most of the frames as containing an event, a higher threshold produces better results, whilst if a model does not have that much confidence in the classification of the frames, a lower threshold will be beneficial. Regarding the aggregation of chunks and the usage of average or median, the difference is not significant in most cases.

This dissertation proved that DL can achieve good results in classification and segmentation of ARS, but it is still a challenging task, and this thesis is a good step in the direction to improving the current results.

6.2 Future Work

Since this is a complex work and there is plenty of room for improvement, there are many research opportunities. In the following paragraphs, some ideas for a possible future work are going to be suggested.

The first and major suggestion for a possible future work is the creation of new datasets/improvement of the ones that already exist since this is a key issue in these studies. The datasets used in this thesis (RSD and HF_Lung_V1) are quite good and both of them have their advantages and disadvantages, but one negative aspect common to both datasets is the quality of the annotations (lack of *golden annotations*). In the HF_Lung_V1, the annotated crackles/DAS, which are supposed to be quite short in duration, have on average 0.89s, while the wheezes/CAS on average have a longer duration than the annotated crackles/DAS (0.85s), and that is not corrected in reality; whilst in the RSD, some of the annotations are not accurate. Also, in the HF_Lung_V1 dataset, there is a lack of demographic information, which was good to know in the RSD, because it allowed us to better comprehend which characteristics achieve better results. Since the splitting of RSD currently available was done according to the number of respiratory cycles and the number of events on each set, a new splitting can be created based on the demographic information to try to balance the number of events between categories and sets to achieve a more balanced partition of the data in both sets. Regarding the creation of new datasets, a wider variety of diseases/comorbidities and events could be present/annotated. The ones already present in the datasets

used are already diverse, but more diversity could be helpful in achieving the final goal of this problem. Also, more events could be present/annotated to have diversity in the datasets, e.g., fine and coarse crackles.

Concerning the features extraction, trying other new features that have not been experimented could have achieved better results than the ones attained like different window sizes accordingly to the events that are being classified/segmented; or another pre-processing method could be beneficial to achieve better results, such as have a denoising phase on the recordings. In conclusion, the more help the models have, the better results could be achieved.

In the classification of ARS, the DL models proved to be better than the classical ML approaches, but none of the DL models testes take into account the past and/or the future, like the CNN-BiLSTM used in the segmentation task.

In the segmentation of ARS, after the classification of the frames is done, the formation of the segments has other alternatives that can improve the results, such as using a threshold for aggregating the segments (e.g., a minimum of 5 consecutive frames with the same classification - even though it depends on the window size and the events in study), as well as other types of segmentation metrics (e.g., Dice Coefficient).

References

[1] WHO, "The top 10 causes of death," https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, 2020, accessed: 2021-12-21.

[2] PORDATA, "Óbitos por algumas causas de morte (%)," https://www.pordata.pt/Portugal/%c3%93bitos+por+algumas+causas+de+morte+(percentagem)-758, 2019, accessed: 2021-12-21.

[3] L. Patel, D. Gandhi, and D. Beddow, "Controversies on the stethoscope during covid-19: A necessary tool or an unnecessary evil?" *The American Journal of the Medical Sciences*, vol. 361, no. 2, p. 278, 2021.

[4] B. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques *et al.*, "A respiratory sound database for the development of automated classification," in *International Conference on Biomedical and Health Informatics*. Springer, 2017, pp. 33–37.

[5] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni *et al.*, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiological measurement*, vol. 40, no. 3, p. 035001, 2019.

[6] F.-S. Hsu, S.-R. Huang, C.-W. Huang, C.-J. Huang, Y.-R. Cheng, C.-C. Chen, J. Hsiao, C.-W. Chen, L.-C. Chen, Y.-C. Lai *et al.*, "Benchmarking of eight recurrent neural network variants for breath phase and adventitious sound detection on a self-developed open-access lung sound database-hf_lung_v1," *arXiv preprint arXiv:2102.03049*, 2021.

[7] B. M. Rocha, D. Pessoa, A. Marques, P. Carvalho, and R. P. Paiva, "Automatic classification of adventitious respiratory sounds: A (un) solved problem?" *Sensors*, vol. 21, no. 1, p. 57, 2021.

[8] A. Sovijarvi, F. Dalmasso, J. Vanderschoot, L. Malmberg, G. Righini, and S. Stoneman, "Definition of terms for applications of respiratory sounds," *European Respiratory Review*, vol. 10, no. 77, pp. 597–610, 2000.

[9] A. Marques and A. Oliveira, "Normal versus adventitious respiratory sounds," in *Breath Sounds*. Springer, 2018, pp. 181–206.

[10] N. Caka, "What are the spectral and temporal features in speech signal?" 03 2015, accessed: 2021-12-28.

[11] O. Lartillot, "Mirtoolbox 1.7. 2 user's manual," *Oslo University*, 2019.

[12] OpenAI, "Openai five defeats dota 2 world champions," https://openai.com/blog/openai-five-defeats-dota-2-world-champions/, 2019, accessed: 2021-12-28.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] C. Jácome, J. Ravn, E. Holsbø, J. C. Aviles-Solis, H. Melbye, and L. Ailo Bongo, "Convolutional neural network for breathing phase detection in lung sounds," *Sensors*, vol. 19, no. 8, p. 1798, 2019.

[15] B. M. Rocha, D. Pessoa, A. Marques, P. Carvalho, and R. P. Paiva, "Influence of event duration on automatic wheeze classification," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 7462–7469.

[16] 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*, pp. 1–13, 2021.

[17] M. Aykanat, Ö. Kılıç, B. Kurt, and S. Saryal, "Classification of lung sounds using convolutional neural networks," *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, pp. 1–9, 2017.

[18] J. Acharya and A. Basu, "Deep neural network for respiratory sound classification in wearable devices enabled by patient specific model tuning," *IEEE transactions on biomedical circuits and systems*, vol. 14, no. 3, pp. 535–544, 2020.

[19] D. Perna and A. Tagarelli, "Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks," in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2019, pp. 50–55.

[20] C.-H. Hsiao, T.-W. Lin, C.-W. Lin, F.-S. Hsu, F. Y.-S. Lin, C.-W. Chen, and C.-M. 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 & Biology Society (EMBC)*. IEEE, 2020, pp. 754–759.

[21] E. Messner, M. Fediuk, P. Swatek, S. Scheidl, F.-M. Smolle-Juttner, H. Olschewski, and F. 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)*. IEEE, 2018, pp. 356–359.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] WHO, "A healthy lifestyle - who recommendations," https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations, 2010, accessed: 2022-07-23.

[24] B. Rees, "Similarity in graphs: Jaccard versus the overlap coefficient," https://medium.com/rapids-ai/similarity-in-graphs-jaccard-versus-the-overlap-coefficient-610e083b877d, 2021, accessed: 2022-08-01.

[25] H. Lim, J.-S. Park, and Y. Han, "Rare sound event detection using 1d convolutional recurrent neural networks." in *DCASE*, 2017, pp. 80–84.

Appendices

Appendix A

Results for the 3-class problem trained with RSD and tested with HF_Lung_V1 and vice-versa

Classifiers	**Accuracy**	**F1 Continuous**	**MCC Continuous**	**F1 Discontinuous**	**MCC Discontinuous**	**F1 Other**	**MCC Other**	**F1 Macro**	**MCC Macro**
LDA_10MRMR	X	X	X	X	X	X	X	X	X
LDA_100MRMR	X	X	X	X	X	X	X	X	X
LDA_Full	X	X	X	X	X	X	X	X	X
SVMrbf_10MRMR	33.6 ± 1.6	43.3 ± 0.6	17.3 ± 1.3	1.2 ± 0.8	1.1 ± 1.8	27.6 ± 3.9	4.7 ± 1.5	24.0 ± 1.8	7.7 ± 1.5
SVMrbf_100MRMR	47.9 ± 3.2	47.7 ± 1.0	26.5 ± 1.9	0.5 ± 1.6	0.1 ± 2.6	57.6 ± 4.4	**31.6 ± 4.7**	35.3 ± 2.3	19.4 ± 3.1
SVMrbf_Full	**55.4 ± 6.3**	**51.9 ± 3.8**	**33.0 ± 5.8**	0.1 ± 0.1	-0.2 ± 0.5	**66.6 ± 7.9**	31.3 ± 7.0	**39.5 ± 3.9**	**21.5 ± 4.4**
RUSBoost_10MRMR	25.1 ± 0.8	39.3 ± 0.1	-3.6 ± 2.9	0.1 ± 0.1	-0.1 ± 0.5	3.2 ± 3.4	-1.6 ± 2.5	14.2 ± 1.2	-1.8 ± 2.0
RUSBoost_100MRMR	27.1 ± 2.2	40.2 ± 0.6	5.8 ± 4.0	0.0 ± 0.0	-0.1 ± 0.3	8.6 ± 7.5	7.9 ± 4.4	16.3 ± 2.7	4.5 ± 2.9
RUSBoost_Full	34.4 ± 5.8	42.9 ± 2.3	15.9 ± 6.6	0.0 ± 0.0	0.0 ± 0.0	28.6 ± 15.0	20.4 ± 6.4	23.8 ± 5.8	12.1 ± 4.3
CNN_dualInput	31.3 ± 5.2	0.1 ± 0.1	-1.5 ± 0.5	29.4 ± 4.0	2.0 ± 7.2	39.9 ± 12.9	11.7 ± 2.5	23.1 ± 5.7	4.1 ± 3.4
CNN_Spectrogram	29.0 ± 6.8	1.5 ± 2.2	-1.4 ± 2.2	**32.2 ± 2.2**	**8.5 ± 4.1**	30.4 ± 16.9	8.2 ± 3.0	21.4 ± 7.1	5.1 ± 3.1
CNN_melSpectrogram	33.7 ± 6.8	0.2 ± 0.2	-0.6 ± 1.4	26.6 ± 4.7	-2.3 ± 6.7	45.2 ± 18.2	12.0 ± 3.3	24.0 ± 7.7	3.0 ± 3.8

Table A.1: Performance results obtained with 3-class problem with the models trained with the RSD and tested with HF_Lung_V1

Classifiers	Accuracy	F1 Wheeze	MCC Wheeze	F1 Crackle	MCC Crackle	F1 Other	MCC Other	F1 Macro	MCC Macro
LDA_10MRMR	X	X	X	X	X	X	X	X	X
LDA_100MRMR	X	X	X	X	X	X	X	X	X
LDA_Full	X	X	X	X	X	X	X	X	X
SVMrbf_10MRMR	**44.3 ± 8.1**	24.1 ± 3.8	**8.8 ± 8.2**	**51.5 ± 16.5**	**17.3 ± 7.6**	40.8 ± 11.8	12.6 ± 9.8	**38.8 ± 10.7**	**12.9 ± 8.5**
SVMrbf_100MRMR	28.5 ± 6.2	17.7 ± 3.2	-3.0 ± 13.4	1.3 ± 1.2	-16.6 ± 2.1	47.6 ± 9.1	7.1 ± 1.5	22.2 ± 4.5	-4.2 ± 5.7
SVMrbf_Full	33.3 ± 7.2	7.7 ± 12.7	1.9 ± 17.3	2.4 ± 5.1	-3.6 ± 5.5	49.5 ± 8.7	1.4 ± 2.2	19.9 ± 8.8	-0.1 ± 8.3
RUSBoost_10MRMR	30.4 ± 1.9	18.5 ± 1.1	-5.2 ± 3.2	35.5 ± 1.1	10.9 ± 1.0	38.2 ± 3.5	15.9 ± 2.2	30.7 ± 1.9	7.2 ± 2.1
RUSBoost_100MRMR	29.5 ± 3.1	17.2 ± 1.8	-6.7 ± 4.8	21.9 ± 6.7	1.2 ± 4.5	47.2 ± 4.0	18.5 ± 3.9	28.8 ± 4.2	4.3 ± 4.4
RUSBoost_Full	35.7 ± 6.4	**18.8 ± 4.0**	-0.9 ± 8.5	26.1 ± 10.2	1.6 ± 6.6	53.4 ± 4.9	22.9 ± 5.1	32.8 ± 6.4	7.9 ± 6.7
CNN_dualInput	36.9 ± 6.8	8.5 ± 2.3	-9.1 ± 8.5	18.2 ± 16.3	-4.9 ± 14.1	**57.3 ± 0.7**	23.7 ± 2.4	28.0 ± 6.4	3.2 ± 8.3
CNN_Spectrogram	27.2 ± 9.0	6.8 ± 2.5	-24.9 ± 15.4	15.1 ± 9.8	-7.9 ± 8.6	48.5 ± 9.9	19.3 ± 4.5	23.5 ± 7.4	-4.5 ± 9.5
CNN_melSpectrogram	29.3 ± 3.6	7.4 ± 1.8	-17.9 ± 7.8	1.7 ± 0.9	-19.8 ± 1.1	57.1 ± 2.0	**25.2 ± 4.0**	22.1 ± 1.6	-4.2 ± 4.3

Table A.2: Performance results obtained with 3-class problem with the models trained with the HF_Lung_V1 and tested with RSD

Appendix B

Complete results of the stratification of RSD

Classifiers	Accuracy	AUC Wheeze	F1 Wheeze	MCC Wheeze	F1 Other	MCC Other
LDA_10MRMR	68.9 ± 0.8	72.1 ± 0.7	72.3 ± 0.5	47.4 ± 1.2	64.7 ± 1.3	47.4 ± 1.2
LDA_100MRMR	83.6 ± 1.4	83.2 ± 1.9	80.7 ± 2.3	66.8 ± 3.3	85.7 ± 1.0	66.8 ± 3.3
LDA_Full	83.6 ± 4.4	84.1 ± 5.5	81.5 ± 6.8	68.0 ± 10.2	85.0 ± 3.0	68.0 ± 10.2
SVMrbf_10MRMR	74.9 ± 1.5	76.1 ± 2.1	74.2 ± 2.7	52.4 ± 4.4	75.4 ± 1.7	52.4 ± 4.4
SVMrbf_100MRMR	81.2 ± 3.9	81.6 ± 3.8	79.4 ± 4.1	62.8 ± 7.6	82.7 ± 3.8	62.8 ± 7.6
SVMrbf_Full	77.0 ± 4.6	77.4 ± 5.1	74.7 ± 5.8	54.4 ± 10.1	78.9 ± 3.9	54.4 ± 10.1
RUSBoost_10MRMR	80.8 ± 2.6	81.0 ± 2.5	78.7 ± 2.7	61.8 ± 5.1	82.5 ± 2.6	61.8 ± 5.1
RUSBoost_100MRMR	85.0 ± 2.6	85.0 ± 2.9	82.8 ± 3.4	69.8 ± 5.2	86.7 ± 2.2	69.8 ± 5.2
RUSBoost_Full	86.9 ± 2.1	86.9 ± 2.5	84.9 ± 2.9	73.6 ± 4.5	88.4 ± 1.7	73.6 ± 4.5
CNN_dualInput	86.0 ± 3.2	86.1 ± 3.6	84.0 ± 4.2	72.3 ± 6.7	87.4 ± 2.9	72.3 ± 6.7
CNN_Spectrogram	77.3 ± 6.2	79.4 ± 5.3	78.4 ± 4.3	60.1 ± 8.8	75.7 ± 8.8	60.1 ± 8.8
CNN_melSpectrogram	88.3 ± 3.3	88.3 ± 3.0	86.7 ± 3.4	76.6 ± 6.3	89.6 ± 3.2	76.6 ± 6.3

Table B.1: Stratification for Children (wheezes vs. others)

Classifiers	Accuracy	AUC Crackle	F1 Crackle	MCC Crackle	F1 Other	MCC Other
LDA_10MRMR	70.1 ± 1.1	78.5 ± 0.6	58.4 ± 0.8	47.5 ± 1.0	76.6 ± 1.2	47.5 ± 1.0
LDA_100MRMR	85.2 ± 1.6	84.9 ± 1.1	71.9 ± 1.4	63.7 ± 1.7	89.9 ± 1.3	63.7 ± 1.7
LDA_Full	77.0 ± 1.6	75.3 ± 8.3	56.7 ± 11.0	44.8 ± 12.3	84.1 ± 1.3	44.8 ± 12.3
SVMrbf_10MRMR	69.4 ± 1.4	78.3 ± 0.9	58.0 ± 1.1	47.1 ± 1.4	76.0 ± 1.4	47.1 ± 1.4
SVMrbf_100MRMR	80.6 ± 1.5	84.2 ± 1.5	67.7 ± 2.0	58.9 ± 2.8	86.1 ± 1.2	58.9 ± 2.8
SVMrbf_Full	81.8 ± 3.0	85.0 ± 2.0	69.2 ± 3.2	60.7 ± 4.2	87.0 ± 2.5	60.7 ± 4.2
RUSBoost_10MRMR	83.3 ± 2.7	83.2 ± 2.5	69.2 ± 4.0	59.9 ± 5.3	88.6 ± 2.0	59.9 ± 5.3
RUSBoost_100MRMR	86.8 ± 1.9	84.5 ± 2.5	73.2 ± 3.6	65.0 ± 4.7	91.3 ± 1.3	65.0 ± 4.7
RUSBoost_Full	84.4 ± 1.8	80.8 ± 3.4	68.0 ± 4.3	58.2 ± 5.6	89.7 ± 1.2	58.2 ± 5.6
CNN_dualInput	82.9 ± 2.1	86.4 ± 1.4	70.9 ± 2.0	63.4 ± 2.4	87.8 ± 1.8	63.4 ± 2.4
CNN_Spectrogram	80.7 ± 2.4	85.5 ± 1.9	68.7 ± 2.6	60.8 ± 3.4	86.1 ± 2.0	60.8 ± 3.4
CNN_melSpectrogram	81.4 ± 2.3	86.2 ± 1.9	69.6 ± 2.9	62.0 ± 3.8	86.6 ± 1.9	62.0 ± 3.8

Table B.2: Stratification for Children (crackles vs. others)

Classifiers	Accuracy	AUC Wheeze	F1 Wheeze	MCC Wheeze	F1 Other	MCC Other
LDA_10MRMR	61.8 ± 0.1	51.4 ± 0.2	74.2 ± 0.1	3.4 ± 0.4	26.8 ± 0.7	3.4 ± 0.4
LDA_100MRMR	54.4 ± 1.6	52.8 ± 1.1	62.8 ± 2.4	5.3 ± 2.1	40.8 ± 2.0	5.3 ± 2.1
LDA_Full	53.3 ± 1.6	52.2 ± 2.6	61.5 ± 1.0	4.1 ± 4.9	40.6 ± 3.5	4.1 ± 4.9
SVMrbf_10MRMR	62.1 ± 1.1	55.6 ± 0.9	72.5 ± 1.2	11.6 ± 1.8	38.9 ± 1.9	11.6 ± 1.8
SVMrbf_100MRMR	65.6 ± 0.6	59.8 ± 1.6	75.0 ± 0.7	20.2 ± 2.7	44.9 ± 3.7	20.2 ± 2.7
SVMrbf_Full	64.3 ± 1.4	61.2 ± 1.3	72.5 ± 2.2	21.9 ± 2.2	48.7 ± 2.9	21.9 ± 2.2
RUSBoost_10MRMR	62.8 ± 0.9	60.8 ± 0.6	70.6 ± 1.3	20.7 ± 1.1	49.2 ± 1.2	20.7 ± 1.1
RUSBoost_100MRMR	62.6 ± 1.0	59.7 ± 0.9	71.0 ± 1.6	18.8 ± 1.6	47.2 ± 1.9	18.8 ± 1.6
RUSBoost_Full	61.4 ± 1.6	60.0 ± 1.3	69.0 ± 2.4	19.0 ± 2.3	48.5 ± 2.2	19.0 ± 2.3
CNN_dualInput	70.5 ± 1.0	68.7 ± 1.9	77.0 ± 1.6	36.3 ± 2.9	58.3 ± 3.0	36.3 ± 2.9
CNN_Spectrogram	72.4 ± 1.8	70.5 ± 1.4	78.7 ± 2.1	40.0 ± 2.5	60.6 ± 1.9	40.0 ± 2.5
CNN_melSpectrogram	71.0 ± 1.2	68.6 ± 0.8	77.7 ± 1.8	36.5 ± 0.9	58.1 ± 1.5	36.5 ± 0.9

Table B.3: Stratification for Adults (wheezes vs. others)

Classifiers	Accuracy	AUC Crackle	F1 Crackle	MCC Crackle	F1 Other	MCC Other
LDA_10MRMR	67.8 ± 0.2	60.2 ± 0.4	77.5 ± 0.1	22.9 ± 0.7	43.5 ± 0.9	22.9 ± 0.7
LDA_100MRMR	68.6 ± 0.3	64.4 ± 0.9	76.5 ± 0.2	29.1 ± 1.5	52.6 ± 1.6	29.1 ± 1.5
LDA_Full	67.9 ± 0.5	64.1 ± 1.4	75.8 ± 0.8	28.2 ± 2.2	52.3 ± 2.5	28.2 ± 2.2
SVMrbf_10MRMR	68.3 ± 0.3	57.7 ± 0.7	79.1 ± 0.2	20.6 ± 1.1	34.5 ± 2.1	20.6 ± 1.1
SVMrbf_100MRMR	71.5 ± 0.9	66.9 ± 2.3	79.0 ± 0.8	34.9 ± 3.2	55.2 ± 4.7	34.9 ± 3.2
SVMrbf_Full	70.0 ± 1.0	66.5 ± 2.4	77.3 ± 1.8	33.4 ± 2.5	55.0 ± 5.7	33.4 ± 2.5
RUSBoost_10MRMR	68.3 ± 0.5	64.9 ± 0.7	75.8 ± 1.0	29.7 ± 0.9	53.7 ± 1.6	29.7 ± 0.9
RUSBoost_100MRMR	69.8 ± 0.6	66.9 ± 0.7	76.9 ± 0.8	33.5 ± 0.9	56.5 ± 1.3	33.5 ± 0.9
RUSBoost_Full	69.9 ± 0.6	67.4 ± 0.7	76.8 ± 0.7	34.3 ± 1.2	57.4 ± 1.2	34.3 ± 1.2
CNN_dualInput	86.7 ± 0.8	82.5 ± 1.9	90.5 ± 0.5	70.0 ± 1.7	77.9 ± 2.4	70.0 ± 1.7
CNN_Spectrogram	85.7 ± 0.9	81.6 ± 1.6	89.7 ± 0.8	67.6 ± 1.9	76.5 ± 2.1	67.6 ± 1.9
CNN_melSpectrogram	87.3 ± 0.6	83.4 ± 1.1	90.9 ± 0.4	71.3 ± 1.4	79.2 ± 1.3	71.3 ± 1.4

Table B.4: Stratification for Adults (crackles vs. others)

Classifiers	Accuracy	AUC Wheeze	F1 Wheeze	MCC Wheeze	F1 Other	MCC Other
LDA_10MRMR	44.8 ± 0.5	39.3 ± 0.4	57.3 ± 0.7	-20.1 ± 0.6	21.8 ± 0.5	-20.1 ± 0.6
LDA_100MRMR	50.5 ± 4.0	53.6 ± 2.2	55.6 ± 6.1	6.8 ± 4.1	43.2 ± 2.7	6.8 ± 4.1
LDA_Full	43.6 ± 4.2	48.0 ± 5.4	47.4 ± 3.5	-3.7 ± 10.3	39.3 ± 5.2	-3.7 ± 10.3
SVMrbf_10MRMR	47.9 ± 1.7	48.8 ± 1.7	55.2 ± 2.5	-2.3 ± 3.1	37.5 ± 2.2	-2.3 ± 3.1
SVMrbf_100MRMR	59.3 ± 3.3	57.9 ± 1.1	67.4 ± 4.1	15.0 ± 2.6	44.9 ± 1.8	15.0 ± 2.6
SVMrbf_Full	56.1 ± 1.6	55.7 ± 2.3	64.1 ± 2.1	10.5 ± 4.3	43.1 ± 3.5	10.5 ± 4.3
RUSBoost_10MRMR	49.6 ± 0.9	54.0 ± 1.1	53.7 ± 2.2	7.6 ± 2.1	44.4 ± 1.6	7.6 ± 2.1
RUSBoost_100MRMR	52.2 ± 1.8	55.6 ± 1.3	57.5 ± 3.1	10.4 ± 2.5	45.2 ± 1.7	10.4 ± 2.5
RUSBoost_Full	50.0 ± 4.0	54.7 ± 2.9	53.6 ± 7.4	9.0 ± 5.5	45.0 ± 3.0	9.0 ± 5.5
CNN_dualInput	75.0 ± 3.8	72.6 ± 3.7	81.3 ± 3.3	43.9 ± 7.4	62.0 ± 4.8	43.9 ± 7.4
CNN_Spectrogram	75.8 ± 2.5	70.7 ± 2.3	82.7 ± 2.3	42.6 ± 4.6	59.3 ± 3.4	42.6 ± 4.6
CNN_melSpectrogram	74.6 ± 3.3	71.5 ± 2.2	81.1 ± 3.2	42.6 ± 5.1	60.6 ± 2.9	42.6 ± 5.1

Table B.5: Stratification for Obese (wheezes vs. others)

Classifiers	Accuracy	AUC Crackle	F1 Crackle	MCC Crackle	F1 Other	MCC Other
LDA_10MRMR	56.8 ± 0.5	51.8 ± 0.4	69.3 ± 0.6	2.8 ± 0.6	27.1 ± 0.5	2.8 ± 0.6
LDA_100MRMR	51.9 ± 1.3	59.1 ± 0.6	61.8 ± 1.8	14.2 ± 1.0	34.9 ± 0.5	14.2 ± 1.0
LDA_Full	53.1 ± 2.9	60.1 ± 4.6	63.1 ± 2.5	15.8 ± 7.2	35.7 ± 3.8	15.8 ± 7.2
SVMrbf_10MRMR	68.4 ± 1.5	53.4 ± 0.7	79.9 ± 1.2	6.2 ± 1.3	25.7 ± 0.9	6.2 ± 1.3
SVMrbf_100MRMR	59.5 ± 5.3	62.5 ± 3.4	69.7 ± 5.2	19.6 ± 5.2	37.8 ± 3.2	19.6 ± 5.2
SVMrbf_Full	53.0 ± 9.1	57.3 ± 2.4	62.8 ± 9.4	11.9 ± 4.6	33.5 ± 1.9	11.9 ± 4.6
RUSBoost_10MRMR	49.9 ± 3.8	55.4 ± 0.8	60.2 ± 4.9	8.5 ± 1.2	31.9 ± 0.7	8.5 ± 1.2
RUSBoost_100MRMR	51.3 ± 3.7	60.8 ± 1.4	60.5 ± 4.3	17.0 ± 2.1	36.4 ± 1.2	17.0 ± 2.1
RUSBoost_Full	51.6 ± 3.3	61.7 ± 2.2	60.6 ± 3.8	18.3 ± 3.4	37.0 ± 1.8	18.3 ± 3.4
CNN_dualInput	88.1 ± 2.5	79.8 ± 0.8	92.6 ± 1.8	61.1 ± 4.9	67.6 ± 2.9	61.1 ± 4.9
CNN_Spectrogram	86.2 ± 2.5	78.6 ± 1.2	91.4 ± 1.8	56.0 ± 4.0	64.1 ± 2.9	56.0 ± 4.0
CNN_melSpectrogram	88.9 ± 2.4	79.1 ± 1.5	93.3 ± 1.7	62.5 ± 5.3	68.0 ± 3.2	62.5 ± 5.3

Table B.6: Stratification for Obese (crackles vs. others)

Classifiers	Accuracy	AUC Wheeze	F1 Wheeze	MCC Wheeze	F1 Other	MCC Other
LDA_10MRMR	65.6 ± 0.2	51.9 ± 0.3	77.8 ± 0.1	4.9 ± 0.8	23.1 ± 1.1	4.9 ± 0.8
LDA_100MRMR	53.1 ± 1.5	51.0 ± 1.2	62.5 ± 2.0	1.8 ± 2.2	37.1 ± 2.0	1.8 ± 2.2
LDA_Full	52.6 ± 0.4	50.9 ± 1.5	61.8 ± 1.0	1.7 ± 2.8	37.3 ± 2.5	1.7 ± 2.8
SVMrbf_10MRMR	65.6 ± 1.3	56.1 ± 0.8	76.5 ± 1.3	13.3 ± 1.7	35.8 ± 2.1	13.3 ± 1.7
SVMrbf_100MRMR	68.3 ± 2.0	59.5 ± 2.6	78.2 ± 1.4	20.5 ± 5.5	41.5 ± 4.4	20.5 ± 5.5
SVMrbf_Full	66.1 ± 2.3	61.4 ± 2.1	75.0 ± 2.7	22.4 ± 4.0	46.6 ± 3.8	22.4 ± 4.0
RUSBoost_10MRMR	66.0 ± 1.4	61.9 ± 0.9	74.7 ± 1.5	22.9 ± 1.8	47.8 ± 1.4	22.9 ± 1.8
RUSBoost_100MRMR	65.0 ± 1.2	59.6 ± 1.0	74.4 ± 1.5	18.8 ± 1.7	44.1 ± 2.2	18.8 ± 1.7
RUSBoost_Full	64.2 ± 2.4	60.6 ± 2.2	73.0 ± 2.5	20.3 ± 4.3	46.4 ± 3.2	20.3 ± 4.3
CNN_dualInput	67.9 ± 2.2	66.9 ± 1.2	75.0 ± 2.8	31.8 ± 2.2	54.9 ± 1.6	31.8 ± 2.2
CNN_Spectrogram	70.1 ± 1.7	68.4 ± 0.9	77.1 ± 2.1	34.9 ± 1.7	56.5 ± 1.2	34.9 ± 1.7
CNN_melSpectrogram	68.8 ± 1.3	66.2 ± 0.9	76.4 ± 1.9	31.0 ± 1.3	53.6 ± 1.5	31.0 ± 1.3

Table B.7: Stratification for Overweight (wheezes vs. others)

Classifiers	Accuracy	AUC Crackle	F1 Crackle	MCC Crackle	F1 Other	MCC Other
LDA_10MRMR	67.2 ± 0.6	66.4 ± 0.6	75.0 ± 0.3	40.8 ± 0.9	52.5 ± 1.3	40.8 ± 0.9
LDA_100MRMR	71.9 ± 1.3	71.3 ± 1.3	77.2 ± 0.8	47.6 ± 2.1	63.3 ± 2.3	47.6 ± 2.1
LDA_Full	70.4 ± 1.9	69.8 ± 2.0	75.8 ± 0.8	43.9 ± 2.7	61.5 ± 4.3	43.9 ± 2.7
SVMrbf_10MRMR	63.7 ± 0.9	62.8 ± 0.9	73.4 ± 0.4	35.9 ± 1.4	42.7 ± 2.4	35.9 ± 1.4
SVMrbf_100MRMR	74.2 ± 2.8	73.7 ± 2.9	78.5 ± 1.4	51.2 ± 4.0	67.4 ± 6.0	51.2 ± 4.0
SVMrbf_Full	74.2 ± 3.5	73.8 ± 3.6	78.1 ± 1.6	50.5 ± 5.1	68.0 ± 7.8	50.5 ± 5.1
RUSBoost_10MRMR	73.5 ± 1.3	73.0 ± 1.3	77.8 ± 0.7	49.3 ± 1.9	67.1 ± 2.6	49.3 ± 1.9
RUSBoost_100MRMR	74.5 ± 1.1	74.1 ± 1.1	78.5 ± 0.6	51.2 ± 1.7	68.8 ± 2.1	51.2 ± 1.7
RUSBoost_Full	75.1 ± 1.2	74.7 ± 1.3	78.7 ± 0.7	52.0 ± 2.0	70.0 ± 2.2	52.0 ± 2.0
CNN_dualInput	85.5 ± 1.6	85.3 ± 1.7	87.0 ± 1.0	72.0 ± 2.5	83.6 ± 2.5	72.0 ± 2.5
CNN_Spectrogram	84.0 ± 1.9	83.7 ± 2.0	86.0 ± 1.2	69.7 ± 2.8	81.4 ± 2.9	69.7 ± 2.8
CNN_melSpectrogram	85.6 ± 1.4	85.4 ± 1.4	87.2 ± 1.0	72.4 ± 2.4	83.7 ± 1.9	72.4 ± 2.4

Table B.8: Stratification for Overweight (crackles vs. others)

Classifiers	Accuracy	AUC Wheeze	F1 Wheeze	MCC Wheeze	F1 Other	MCC Other
LDA_10MRMR	65.3 ± 0.2	60.4 ± 0.3	75.4 ± 0.1	27.5 ± 0.6	41.0 ± 0.6	27.5 ± 0.6
LDA_100MRMR	61.9 ± 0.5	59.7 ± 0.4	69.2 ± 1.0	20.2 ± 0.8	50.1 ± 1.6	20.2 ± 0.8
LDA_Full	64.5 ± 4.2	62.1 ± 4.4	71.6 ± 3.4	25.5 ± 9.1	52.6 ± 5.9	25.5 ± 9.1
SVMrbf_10MRMR	63.6 ± 1.0	60.3 ± 1.3	72.0 ± 0.7	22.8 ± 2.4	47.6 ± 2.9	22.8 ± 2.4
SVMrbf_100MRMR	63.0 ± 2.7	61.1 ± 2.8	69.5 ± 3.0	23.1 ± 5.5	52.2 ± 5.8	23.1 ± 5.5
SVMrbf_Full	66.1 ± 1.4	64.9 ± 1.0	71.2 ± 2.4	30.3 ± 2.1	58.5 ± 2.0	30.3 ± 2.1
RUSBoost_10MRMR	64.7 ± 2.1	63.6 ± 2.2	69.9 ± 1.9	27.4 ± 4.4	57.4 ± 3.0	27.4 ± 4.4
RUSBoost_100MRMR	64.6 ± 1.3	63.3 ± 1.7	69.9 ± 1.3	26.9 ± 3.2	56.7 ± 2.9	26.9 ± 3.2
RUSBoost_Full	63.1 ± 1.7	62.4 ± 1.8	67.6 ± 1.9	24.8 ± 3.6	56.9 ± 2.7	24.8 ± 3.6
CNN_dualInput	79.4 ± 2.5	78.1 ± 2.6	82.9 ± 2.2	57.6 ± 5.1	73.9 ± 3.5	57.6 ± 5.1
CNN_Spectrogram	77.4 ± 2.5	75.5 ± 3.0	81.9 ± 1.8	53.5 ± 5.1	69.9 ± 4.7	53.5 ± 5.1
CNN_melSpectrogram	76.3 ± 1.8	74.0 ± 2.3	81.2 ± 1.0	51.0 ± 3.7	67.8 ± 3.7	51.0 ± 3.7

Table B.9: Stratification for Normal (wheezes vs. others)

Classifiers	Accuracy	AUC Crackle	F1 Crackle	MCC Crackle	F1 Other	MCC Other
LDA_10MRMR	78.8 ± 0.2	64.2 ± 0.3	86.9 ± 0.1	41.6 ± 0.5	44.9 ± 0.8	41.6 ± 0.5
LDA_100MRMR	79.4 ± 0.4	68.7 ± 1.2	86.7 ± 0.1	44.3 ± 1.5	54.7 ± 2.1	44.3 ± 1.5
LDA_Full	78.3 ± 4.0	68.9 ± 2.6	85.6 ± 3.2	43.2 ± 8.3	55.3 ± 4.2	43.2 ± 8.3
SVMrbf_10MRMR	74.6 ± 0.6	56.1 ± 1.2	84.8 ± 0.3	24.5 ± 3.0	23.5 ± 3.6	24.5 ± 3.0
SVMrbf_100MRMR	79.0 ± 1.7	65.9 ± 3.7	86.8 ± 0.8	41.8 ± 6.4	48.6 ± 8.4	41.8 ± 6.4
SVMrbf_Full	80.1 ± 2.2	69.1 ± 5.3	87.2 ± 1.0	45.8 ± 8.6	54.0 ± 12.2	45.8 ± 8.6
RUSBoost_10MRMR	78.2 ± 0.6	68.3 ± 1.7	85.7 ± 0.3	41.6 ± 2.3	54.0 ± 3.0	41.6 ± 2.3
RUSBoost_100MRMR	80.6 ± 1.1	70.3 ± 2.2	87.4 ± 0.6	47.8 ± 3.5	57.4 ± 4.0	47.8 ± 3.5
RUSBoost_Full	80.0 ± 1.4	70.4 ± 1.5	86.9 ± 1.1	46.6 ± 3.4	57.6 ± 2.5	46.6 ± 3.4
CNN_dualInput	88.1 ± 1.3	80.2 ± 3.0	92.3 ± 0.8	69.7 ± 3.4	74.3 ± 4.3	69.7 ± 3.4
CNN_Spectrogram	87.8 ± 1.2	79.7 ± 2.7	92.1 ± 0.7	68.8 ± 3.0	73.6 ± 3.9	68.8 ± 3.0
CNN_melSpectrogram	88.7 ± 0.6	80.5 ± 1.5	92.6 ± 0.4	71.3 ± 1.6	75.3 ± 2.0	71.3 ± 1.6

Table B.10: Stratification for Normal (crackles vs. others)

Classifiers	Accuracy	AUC Wheeze	F1 Wheeze	MCC Wheeze	F1 Other	MCC Other
LDA_10MRMR	39.4 ± 0.6	50.8 ± 1.2	39.1 ± 1.0	1.6 ± 2.5	39.7 ± 0.3	1.6 ± 2.5
LDA_100MRMR	50.9 ± 2.3	45.0 ± 0.7	25.7 ± 0.9	-9.0 ± 1.2	63.2 ± 2.8	-9.0 ± 1.2
LDA_Full	51.7 ± 4.3	45.7 ± 3.1	26.2 ± 3.7	-7.7 ± 5.7	63.9 ± 4.3	-7.7 ± 5.7
SVMrbf_10MRMR	45.9 ± 1.2	50.5 ± 2.1	36.6 ± 2.5	0.9 ± 3.7	52.7 ± 2.5	0.9 ± 3.7
SVMrbf_100MRMR	50.8 ± 4.2	48.0 ± 5.7	30.8 ± 6.6	-3.5 ± 10.1	61.7 ± 3.8	-3.5 ± 10.1
SVMrbf_Full	53.0 ± 2.3	49.5 ± 4.2	31.3 ± 6.0	-1.0 ± 7.4	64.1 ± 3.0	-1.0 ± 7.4
RUSBoost_10MRMR	54.2 ± 2.6	49.6 ± 3.2	31.1 ± 4.4	-0.8 ± 5.8	65.6 ± 2.5	-0.8 ± 5.8
RUSBoost_100MRMR	49.8 ± 1.4	42.3 ± 2.0	21.5 ± 3.6	-14.0 ± 3.7	63.0 ± 1.8	-14.0 ± 3.7
RUSBoost_Full	54.6 ± 3.1	46.4 ± 3.3	24.9 ± 5.6	-6.7 ± 6.2	67.3 ± 3.1	-6.7 ± 6.2
CNN_dualInput	66.0 ± 3.4	62.3 ± 1.7	45.5 ± 2.3	22.8 ± 3.1	75.1 ± 3.5	22.8 ± 3.1
CNN_Spectrogram	63.7 ± 4.1	63.2 ± 1.9	47.2 ± 1.7	23.7 ± 3.7	72.2 ± 4.4	23.7 ± 3.7
CNN_melSpectrogram	64.4 ± 4.0	63.8 ± 1.7	47.8 ± 1.8	24.7 ± 3.3	72.8 ± 4.4	24.7 ± 3.3

Table B.11: Stratification for Females (wheezes vs. others)

Classifiers	Accuracy	AUC Crackle	F1 Crackle	MCC Crackle	F1 Other	MCC Other
LDA_10MRMR	54.9 ± 1.0	68.0 ± 0.8	51.0 ± 0.7	33.2 ± 1.4	58.2 ± 1.3	33.2 ± 1.4
LDA_100MRMR	61.5 ± 1.6	65.1 ± 1.5	48.3 ± 1.5	26.1 ± 2.7	69.3 ± 1.7	26.1 ± 2.7
LDA_Full	58.8 ± 2.2	61.6 ± 3.0	44.4 ± 4.0	20.3 ± 5.1	66.9 ± 3.6	20.3 ± 5.1
SVMrbf_10MRMR	49.1 ± 1.4	64.1 ± 1.3	47.9 ± 1.0	27.4 ± 2.5	50.3 ± 1.9	27.4 ± 2.5
SVMrbf_100MRMR	72.0 ± 5.4	78.3 ± 3.2	62.0 ± 3.7	49.3 ± 5.3	77.7 ± 5.8	49.3 ± 5.3
SVMrbf_Full	72.6 ± 6.7	78.0 ± 3.7	62.2 ± 4.4	49.1 ± 6.1	78.2 ± 7.3	49.1 ± 6.1
RUSBoost_10MRMR	64.7 ± 2.4	69.2 ± 1.3	52.5 ± 1.4	33.4 ± 2.3	71.8 ± 2.7	33.4 ± 2.3
RUSBoost_100MRMR	70.7 ± 2.0	74.1 ± 1.9	57.9 ± 2.1	42.0 ± 3.3	77.5 ± 1.9	42.0 ± 3.3
RUSBoost_Full	72.9 ± 1.7	76.3 ± 2.3	60.4 ± 2.3	45.8 ± 3.8	79.4 ± 1.5	45.8 ± 3.8
CNN_dualInput	80.3 ± 3.1	84.1 ± 1.1	70.0 ± 2.7	60.5 ± 3.0	85.3 ± 2.9	60.5 ± 3.0
CNN_Spectrogram	78.3 ± 3.5	84.2 ± 1.7	69.0 ± 2.9	59.8 ± 3.4	83.2 ± 3.3	59.8 ± 3.4
CNN_melSpectrogram	80.9 ± 2.2	85.9 ± 1.3	71.5 ± 2.2	63.0 ± 2.8	85.6 ± 2.0	63.0 ± 2.8

Table B.12: Stratification for Females (crackles vs. others)

Classifiers	Accuracy	AUC Wheeze	F1 Wheeze	MCC Wheeze	F1 Other	MCC Other
LDA_10MRMR	65.1 ± 0.1	53.5 ± 0.3	76.8 ± 0.1	8.0 ± 0.7	29.5 ± 0.9	8.0 ± 0.7
LDA_100MRMR	57.3 ± 1.9	55.7 ± 1.0	66.1 ± 2.5	10.6 ± 1.9	42.1 ± 1.7	10.6 ± 1.9
LDA_Full	56.1 ± 1.2	54.8 ± 2.0	65.0 ± 1.1	8.8 ± 3.7	41.2 ± 2.7	8.8 ± 3.7
SVMrbf_10MRMR	65.1 ± 1.2	58.1 ± 0.9	75.1 ± 1.2	16.4 ± 1.8	41.2 ± 1.8	16.4 ± 1.8
SVMrbf_100MRMR	68.7 ± 1.0	62.2 ± 1.6	77.8 ± 1.0	24.7 ± 2.7	46.6 ± 3.4	24.7 ± 2.7
SVMrbf_Full	66.8 ± 1.8	63.3 ± 1.2	75.1 ± 2.3	25.5 ± 2.0	49.4 ± 2.3	25.5 ± 2.0
RUSBoost_10MRMR	65.4 ± 1.1	63.2 ± 1.0	73.5 ± 1.4	24.9 ± 1.7	50.0 ± 1.4	24.9 ± 1.7
RUSBoost_100MRMR	66.1 ± 1.2	62.9 ± 1.0	74.5 ± 1.6	24.5 ± 1.6	49.2 ± 1.8	24.5 ± 1.6
RUSBoost_Full	64.4 ± 1.9	62.8 ± 1.1	72.4 ± 2.4	23.9 ± 2.0	49.6 ± 1.7	23.9 ± 2.0
CNN_dualInput	72.4 ± 1.3	70.4 ± 1.9	79.2 ± 1.8	38.9 ± 2.6	58.5 ± 2.5	38.9 ± 2.6
CNN_Spectrogram	73.8 ± 2.1	71.5 ± 1.4	80.4 ± 2.2	41.4 ± 2.6	60.1 ± 1.9	41.4 ± 2.6
CNN_melSpectrogram	73.1 ± 1.7	70.6 ± 0.6	80.0 ± 2.1	39.7 ± 0.9	58.9 ± 0.9	39.7 ± 0.9

Table B.13: Stratification for Males (wheezes vs. others)

Classifiers	Accuracy	AUC Crackle	F1 Crackle	MCC Crackle	F1 Other	MCC Other
LDA_10MRMR	69.5 ± 0.1	61.2 ± 0.4	79.0 ± 0.1	24.6 ± 0.6	44.5 ± 0.8	24.6 ± 0.6
LDA_100MRMR	70.4 ± 0.3	66.1 ± 1.0	78.2 ± 0.3	31.9 ± 1.6	53.6 ± 1.6	31.9 ± 1.6
LDA_Full	69.6 ± 0.6	65.5 ± 1.3	77.5 ± 0.9	30.6 ± 1.9	52.9 ± 2.0	30.6 ± 1.9
SVMrbf_10MRMR	70.7 ± 0.3	58.9 ± 0.7	80.9 ± 0.2	23.2 ± 1.2	36.9 ± 2.0	23.2 ± 1.2
SVMrbf_100MRMR	72.0 ± 0.8	66.6 ± 1.8	79.8 ± 1.0	34.0 ± 2.2	53.7 ± 3.6	34.0 ± 2.2
SVMrbf_Full	70.4 ± 1.6	66.3 ± 2.0	78.0 ± 2.2	32.7 ± 1.9	53.5 ± 4.6	32.7 ± 1.9
RUSBoost_10MRMR	69.6 ± 0.7	66.2 ± 0.7	77.2 ± 1.1	31.6 ± 0.9	54.0 ± 1.3	31.6 ± 0.9
RUSBoost_100MRMR	70.7 ± 0.7	67.7 ± 0.5	78.0 ± 0.9	34.4 ± 0.8	56.1 ± 1.0	34.4 ± 0.8
RUSBoost_Full	70.4 ± 0.7	67.8 ± 0.8	77.6 ± 0.8	34.4 ± 1.4	56.3 ± 1.1	34.4 ± 1.4
CNN_dualInput	87.3 ± 0.7	82.4 ± 1.9	91.2 ± 0.4	69.9 ± 1.6	77.2 ± 2.3	69.9 ± 1.6
CNN_Spectrogram	86.2 ± 1.0	81.5 ± 1.5	90.4 ± 0.9	67.4 ± 2.1	75.8 ± 1.8	67.4 ± 2.1
CNN_melSpectrogram	87.8 ± 0.6	83.2 ± 1.1	91.5 ± 0.4	70.9 ± 1.4	78.4 ± 1.3	70.9 ± 1.4

Table B.14: Stratification for Males (crackles vs. others)

Classifiers	Accuracy	AUC Wheeze	F1 Wheeze	MCC Wheeze	F1 Other	MCC Other
LDA_10MRMR	82.4 ± 0.0	73.9 ± 0.0	88.3 ± 0.0	54.8 ± 0.0	64.0 ± 0.0	54.8 ± 0.0
LDA_100MRMR	81.8 ± 2.6	82.6 ± 1.8	86.0 ± 2.6	61.9 ± 3.6	73.3 ± 2.1	61.9 ± 3.6
LDA_Full	83.7 ± 8.6	81.1 ± 5.4	87.8 ± 7.7	64.0 ± 14.3	74.4 ± 8.6	64.0 ± 14.3
SVMrbf_10MRMR	80.0 ± 4.2	76.7 ± 3.5	85.5 ± 3.7	53.7 ± 7.5	67.0 ± 4.9	53.7 ± 7.5
SVMrbf_100MRMR	82.4 ± 3.7	80.9 ± 4.0	87.0 ± 3.0	60.1 ± 7.4	72.0 ± 5.4	60.1 ± 7.4
SVMrbf_Full	76.7 ± 6.5	74.5 ± 4.9	82.6 ± 5.6	48.0 ± 11.4	64.2 ± 6.7	48.0 ± 11.4
RUSBoost_10MRMR	82.2 ± 4.5	81.9 ± 5.5	86.7 ± 3.4	60.7 ± 10.2	72.8 ± 6.8	60.7 ± 10.2
RUSBoost_100MRMR	84.5 ± 3.7	84.4 ± 3.1	88.4 ± 3.2	66.1 ± 6.4	76.3 ± 4.4	66.1 ± 6.4
RUSBoost_Full	84.1 ± 3.6	82.3 ± 2.5	88.4 ± 3.0	63.5 ± 6.2	74.5 ± 4.0	63.5 ± 6.2
CNN_dualInput	85.5 ± 4.4	83.5 ± 5.1	89.6 ± 3.2	66.0 ± 10.1	76.2 ± 7.0	66.0 ± 10.1
CNN_Spectrogram	84.5 ± 3.3	77.6 ± 5.7	89.6 ± 2.0	61.0 ± 9.6	69.0 ± 10.2	61.0 ± 9.6
CNN_melSpectrogram	86.3 ± 2.9	84.8 ± 3.7	90.1 ± 2.2	68.2 ± 6.8	77.7 ± 4.7	68.2 ± 6.8

Table B.15: Stratification for Non-Chronic (wheezes vs. others)

Classifiers	Accuracy	AUC Crackle	F1 Crackle	MCC Crackle	F1 Other	MCC Other
LDA_10MRMR	77.1 ± 2.2	78.6 ± 2.0	79.0 ± 1.5	59.4 ± 3.3	74.6 ± 3.2	59.4 ± 3.3
LDA_100MRMR	83.9 ± 0.9	84.4 ± 0.8	83.5 ± 0.7	68.6 ± 1.5	84.3 ± 1.2	68.6 ± 1.5
LDA_Full	74.1 ± 5.4	74.3 ± 6.7	71.0 ± 12.7	49.9 ± 11.8	75.4 ± 2.4	49.9 ± 11.8
SVMrbf_10MRMR	74.3 ± 1.8	76.1 ± 1.6	77.1 ± 1.2	55.2 ± 2.7	70.8 ± 2.6	55.2 ± 2.7
SVMrbf_100MRMR	81.6 ± 2.5	82.4 ± 2.5	81.8 ± 2.4	65.1 ± 4.9	81.3 ± 2.7	65.1 ± 4.9
SVMrbf_Full	82.4 ± 3.1	83.3 ± 2.9	82.8 ± 2.5	67.0 ± 5.4	82.0 ± 3.8	67.0 ± 5.4
RUSBoost_10MRMR	79.9 ± 3.3	80.3 ± 3.3	79.3 ± 3.3	60.6 ± 6.5	80.4 ± 3.4	60.6 ± 6.5
RUSBoost_100MRMR	82.5 ± 3.3	82.7 ± 3.2	81.5 ± 3.4	65.2 ± 6.5	83.5 ± 3.2	65.2 ± 6.5
RUSBoost_Full	80.5 ± 2.6	80.7 ± 2.8	79.2 ± 3.2	61.3 ± 5.5	81.6 ± 2.4	61.3 ± 5.5
CNN_dualInput	85.3 ± 3.3	86.0 ± 2.8	85.4 ± 2.6	72.4 ± 5.0	85.1 ± 4.0	72.4 ± 5.0
CNN_Spectrogram	82.9 ± 2.4	84.0 ± 2.3	83.6 ± 2.1	68.9 ± 4.4	82.3 ± 2.9	68.9 ± 4.4
CNN_melSpectrogram	84.1 ± 2.8	85.2 ± 2.7	84.6 ± 2.6	70.9 ± 5.2	83.5 ± 3.2	70.9 ± 5.2

Table B.16: Stratification for Non-Chronic (crackles vs. others)

Classifiers	Accuracy	AUC Wheeze	F1 Wheeze	MCC Wheeze	F1 Other	MCC Other
LDA_10MRMR	62.0 ± 0.1	52.0 ± 0.1	74.1 ± 0.1	4.6 ± 0.2	28.1 ± 0.4	4.6 ± 0.2
LDA_100MRMR	54.2 ± 1.6	53.1 ± 1.1	62.1 ± 2.4	5.9 ± 2.0	41.8 ± 1.8	5.9 ± 2.0
LDA_Full	53.1 ± 1.6	52.5 ± 2.5	60.9 ± 1.0	4.7 ± 4.7	41.5 ± 3.3	4.7 ± 4.7
SVMrbf_10MRMR	62.5 ± 1.0	56.4 ± 0.8	72.6 ± 1.2	13.2 ± 1.6	40.4 ± 1.7	13.2 ± 1.6
SVMrbf_100MRMR	65.6 ± 0.7	60.2 ± 1.6	74.8 ± 0.7	20.8 ± 2.7	45.8 ± 3.4	20.8 ± 2.7
SVMrbf_Full	64.9 ± 1.3	62.0 ± 1.4	72.8 ± 2.1	23.4 ± 2.2	49.9 ± 2.9	23.4 ± 2.2
RUSBoost_10MRMR	63.2 ± 0.8	61.5 ± 0.7	70.8 ± 1.3	22.0 ± 1.2	50.2 ± 1.2	22.0 ± 1.2
RUSBoost_100MRMR	63.0 ± 1.0	60.4 ± 0.8	71.1 ± 1.6	20.2 ± 1.4	48.4 ± 1.7	20.2 ± 1.4
RUSBoost_Full	61.9 ± 1.6	60.7 ± 1.2	69.2 ± 2.5	20.4 ± 2.2	49.7 ± 2.0	20.4 ± 2.2
CNN_dualInput	70.9 ± 1.0	69.3 ± 1.8	77.3 ± 1.6	37.4 ± 2.7	59.2 ± 2.8	37.4 ± 2.7
CNN_Spectrogram	72.7 ± 1.7	71.0 ± 1.4	78.8 ± 2.0	40.9 ± 2.5	61.4 ± 2.0	40.9 ± 2.5
CNN_melSpectrogram	71.4 ± 1.2	69.3 ± 0.7	77.9 ± 1.8	37.7 ± 0.8	59.2 ± 1.3	37.7 ± 0.8

Table B.17: Stratification for Chronic (wheezes vs. others)

Classifiers	Accuracy	AUC Crackle	F1 Crackle	MCC Crackle	F1 Other	MCC Other
LDA_10MRMR	67.9 ± 0.2	60.9 ± 0.4	77.4 ± 0.1	24.4 ± 0.6	45.0 ± 0.8	24.4 ± 0.6
LDA_100MRMR	69.0 ± 0.4	65.3 ± 0.9	76.6 ± 0.2	31.0 ± 1.5	54.3 ± 1.6	31.0 ± 1.5
LDA_Full	68.1 ± 0.6	64.7 ± 1.4	75.7 ± 0.8	29.5 ± 2.2	53.7 ± 2.5	29.5 ± 2.2
SVMrbf_10MRMR	68.5 ± 0.3	58.6 ± 0.6	79.0 ± 0.2	22.9 ± 1.0	36.8 ± 1.9	22.9 ± 1.0
SVMrbf_100MRMR	71.9 ± 0.9	67.8 ± 2.1	79.1 ± 0.8	36.7 ± 3.0	57.0 ± 4.3	36.7 ± 3.0
SVMrbf_Full	70.4 ± 1.0	67.3 ± 2.2	77.4 ± 1.8	35.0 ± 2.1	56.6 ± 5.2	35.0 ± 2.1
RUSBoost_10MRMR	68.7 ± 0.5	65.7 ± 0.7	75.9 ± 1.0	31.3 ± 0.9	55.2 ± 1.6	31.3 ± 0.9
RUSBoost_100MRMR	70.3 ± 0.6	67.8 ± 0.6	77.0 ± 0.8	35.3 ± 0.9	58.1 ± 1.2	35.3 ± 0.9
RUSBoost_Full	70.5 ± 0.6	68.3 ± 0.7	76.9 ± 0.7	36.0 ± 1.2	58.9 ± 1.1	36.0 ± 1.2
CNN_dualInput	86.6 ± 0.9	82.7 ± 1.9	90.4 ± 0.5	70.2 ± 1.8	78.2 ± 2.4	70.2 ± 1.8
CNN_Spectrogram	85.6 ± 0.9	81.8 ± 1.6	89.6 ± 0.7	67.9 ± 1.9	76.9 ± 2.1	67.9 ± 1.9
CNN_melSpectrogram	87.3 ± 0.6	83.6 ± 1.1	90.8 ± 0.4	71.6 ± 1.4	79.6 ± 1.3	71.6 ± 1.4

Table B.18: Stratification for Chronic (crackles vs. others)

Classifiers	Accuracy	AUC Wheeze	F1 Wheeze	MCC Wheeze	F1 Other	MCC Other
LDA_10MRMR	62.7 ± 0.2	61.2 ± 0.2	69.0 ± 0.1	23.4 ± 0.4	53.1 ± 0.4	23.4 ± 0.4
LDA_100MRMR	64.1 ± 2.1	66.0 ± 1.8	58.8 ± 3.9	34.1 ± 3.1	68.1 ± 1.1	34.1 ± 3.1
LDA_Full	61.0 ± 3.6	63.0 ± 3.0	54.6 ± 7.6	27.9 ± 5.0	65.6 ± 1.4	27.9 ± 5.0
SVMrbf_10MRMR	68.8 ± 1.8	68.6 ± 1.7	71.4 ± 2.2	37.3 ± 3.5	65.7 ± 2.2	37.3 ± 3.5
SVMrbf_100MRMR	72.2 ± 3.0	72.6 ± 2.9	73.0 ± 3.8	45.1 ± 5.5	71.2 ± 2.7	45.1 ± 5.5
SVMrbf_Full	74.4 ± 2.5	74.5 ± 2.4	75.7 ± 2.6	48.9 ± 4.8	72.8 ± 2.5	48.9 ± 4.8
RUSBoost_10MRMR	77.5 ± 1.8	77.6 ± 1.9	78.8 ± 1.8	55.1 ± 3.7	75.9 ± 2.1	55.1 ± 3.7
RUSBoost_100MRMR	78.1 ± 2.0	78.8 ± 1.8	78.3 ± 2.4	57.6 ± 3.5	77.8 ± 1.7	57.6 ± 3.5
RUSBoost_Full	77.4 ± 2.4	78.4 ± 2.2	77.0 ± 3.1	57.1 ± 3.9	77.8 ± 1.8	57.1 ± 3.9
CNN_dualInput	79.6 ± 1.8	80.4 ± 1.8	80.3 ± 2.2	60.5 ± 3.5	78.8 ± 1.8	60.5 ± 3.5
CNN_Spectrogram	75.2 ± 3.4	74.5 ± 4.1	78.3 ± 2.0	49.6 ± 7.4	70.6 ± 6.0	49.6 ± 7.4
CNN_melSpectrogram	81.3 ± 2.1	82.2 ± 2.3	81.9 ± 1.7	64.0 ± 4.7	80.6 ± 2.5	64.0 ± 4.7

Table B.19: Stratification for Meditron (wheezes vs. others)

Classifiers	Accuracy	AUC Crackle	F1 Crackle	MCC Crackle	F1 Other	MCC Other
LDA_10MRMR	74.0 ± 1.0	79.4 ± 0.8	72.4 ± 0.8	57.0 ± 1.4	75.5 ± 1.1	57.0 ± 1.4
LDA_100MRMR	86.7 ± 2.2	86.7 ± 1.6	82.1 ± 2.4	71.9 ± 4.0	89.4 ± 2.0	71.9 ± 4.0
LDA_Full	76.4 ± 2.2	74.7 ± 1.1	67.1 ± 1.6	49.5 ± 2.7	81.4 ± 2.7	49.5 ± 2.7
SVMrbf_10MRMR	73.7 ± 1.0	78.6 ± 1.0	71.6 ± 1.0	55.2 ± 1.9	75.5 ± 1.1	55.2 ± 1.9
SVMrbf_100MRMR	86.9 ± 1.0	88.6 ± 0.9	83.5 ± 1.2	74.2 ± 1.9	89.2 ± 0.9	74.2 ± 1.9
SVMrbf_Full	86.4 ± 2.0	88.0 ± 1.6	82.8 ± 2.1	73.1 ± 3.3	88.7 ± 1.8	73.1 ± 3.3
RUSBoost_10MRMR	84.7 ± 1.8	85.2 ± 1.4	79.9 ± 1.9	68.2 ± 3.2	87.6 ± 1.6	68.2 ± 3.2
RUSBoost_100MRMR	89.3 ± 0.9	88.4 ± 1.0	84.8 ± 1.3	76.6 ± 1.9	91.7 ± 0.7	76.6 ± 1.9
RUSBoost_Full	87.7 ± 1.4	86.5 ± 1.8	82.4 ± 2.2	73.0 ± 3.2	90.5 ± 1.0	73.0 ± 3.2
CNN_dualInput	85.2 ± 1.5	86.3 ± 0.8	81.3 ± 1.2	70.4 ± 1.9	87.7 ± 1.6	70.4 ± 1.9
CNN_Spectrogram	84.2 ± 2.4	86.1 ± 1.8	80.8 ± 2.3	69.5 ± 3.7	86.5 ± 2.4	69.5 ± 3.7
CNN_melSpectrogram	85.0 ± 2.3	86.9 ± 2.0	81.7 ± 2.4	71.1 ± 4.0	87.3 ± 2.1	71.1 ± 4.0

Table B.20: Stratification for Meditron (crackles vs. others)

Classifiers	Accuracy	AUC Wheeze	F1 Wheeze	MCC Wheeze	F1 Other	MCC Other
LDA_10MRMR	70.2 ± 0.3	52.3 ± 0.2	81.8 ± 0.2	8.2 ± 0.8	16.9 ± 0.6	8.2 ± 0.8
LDA_100MRMR	43.6 ± 4.3	48.8 ± 2.7	47.8 ± 6.4	-2.4 ± 5.2	38.3 ± 1.5	-2.4 ± 5.2
LDA_Full	46.6 ± 9.9	51.6 ± 7.4	50.5 ± 12.7	2.4 ± 13.8	40.9 ± 5.4	2.4 ± 13.8
SVMrbf_10MRMR	67.8 ± 4.7	59.2 ± 2.6	77.7 ± 4.5	19.6 ± 6.3	40.9 ± 3.2	19.6 ± 6.3
SVMrbf_100MRMR	71.5 ± 3.4	62.4 ± 1.7	80.7 ± 2.9	27.2 ± 5.5	45.4 ± 1.7	27.2 ± 5.5
SVMrbf_Full	68.8 ± 3.9	63.5 ± 4.4	77.5 ± 3.2	26.4 ± 9.0	48.2 ± 6.5	26.4 ± 9.0
RUSBoost_10MRMR	67.8 ± 1.6	64.0 ± 1.6	76.3 ± 1.4	26.5 ± 3.1	49.5 ± 2.1	26.5 ± 3.1
RUSBoost_100MRMR	65.1 ± 4.0	61.5 ± 2.7	74.0 ± 3.9	21.7 ± 5.5	46.6 ± 2.7	21.7 ± 5.5
RUSBoost_Full	65.2 ± 3.0	62.5 ± 2.2	73.7 ± 3.0	23.4 ± 4.3	48.1 ± 2.5	23.4 ± 4.3
CNN_dualInput	51.4 ± 1.8	57.7 ± 3.3	57.1 ± 2.0	14.0 ± 6.0	43.9 ± 3.3	14.0 ± 6.0
CNN_Spectrogram	57.5 ± 3.6	62.8 ± 2.8	63.7 ± 4.5	23.0 ± 5.0	48.4 ± 2.6	23.0 ± 5.0
CNN_melSpectrogram	52.6 ± 3.8	56.3 ± 2.5	59.5 ± 5.5	11.3 ± 4.6	42.0 ± 2.5	11.3 ± 4.6

Table B.21: Stratification for Litt3200 (wheezes vs. others)

Classifiers	Accuracy	AUC Crackle	F1 Crackle	MCC Crackle	F1 Other	MCC Other
LDA_10MRMR	54.2 ± 0.6	62.0 ± 0.4	33.4 ± 0.3	17.5 ± 0.7	65.1 ± 0.7	17.5 ± 0.7
LDA_100MRMR	69.9 ± 0.9	55.1 ± 3.1	25.5 ± 4.0	8.4 ± 4.7	81.1 ± 0.8	8.4 ± 4.7
LDA_Full	63.3 ± 10.1	47.7 ± 5.1	15.0 ± 8.6	-4.1 ± 8.7	75.6 ± 8.5	-4.1 ± 8.7
SVMrbf_10MRMR	52.7 ± 2.5	65.4 ± 0.8	35.7 ± 0.6	22.7 ± 1.1	62.6 ± 3.0	22.7 ± 1.1
SVMrbf_100MRMR	77.3 ± 6.7	71.2 ± 3.1	46.7 ± 4.0	36.1 ± 5.1	85.3 ± 5.5	36.1 ± 5.1
SVMrbf_Full	78.6 ± 6.3	70.7 ± 2.5	47.2 ± 5.0	36.2 ± 6.4	86.4 ± 4.9	36.2 ± 6.4
RUSBoost_10MRMR	63.6 ± 1.7	64.9 ± 1.7	36.4 ± 1.4	21.9 ± 2.4	74.5 ± 1.7	21.9 ± 2.4
RUSBoost_100MRMR	74.5 ± 0.6	64.2 ± 1.8	37.5 ± 2.2	23.7 ± 2.8	84.0 ± 0.3	23.7 ± 2.8
RUSBoost_Full	78.2 ± 1.9	67.0 ± 3.4	42.0 ± 4.8	29.8 ± 5.9	86.6 ± 1.3	29.8 ± 5.9
CNN_dualInput	90.9 ± 1.3	86.3 ± 0.7	80.4 ± 2.1	74.7 ± 3.1	94.1 ± 0.9	74.7 ± 3.1
CNN_Spectrogram	87.6 ± 1.6	84.2 ± 1.0	75.0 ± 2.3	67.0 ± 3.3	91.7 ± 1.2	67.0 ± 3.3
CNN_melSpectrogram	91.7 ± 1.1	86.9 ± 0.8	81.8 ± 1.8	76.8 ± 2.7	94.6 ± 0.8	76.8 ± 2.7

Table B.22: Stratification for Litt3200 (crackles vs. others)

Classifiers	Accuracy	AUC Wheeze	F1 Wheeze	MCC Wheeze	F1 Other	MCC Other
LDA_10MRMR	60.1 ± 0.0	51.7 ± 0.2	72.1 ± 0.1	4.0 ± 0.4	29.7 ± 0.7	4.0 ± 0.4
LDA_100MRMR	58.8 ± 1.4	55.5 ± 0.7	67.6 ± 2.3	10.9 ± 1.2	42.9 ± 2.7	10.9 ± 1.2
LDA_Full	57.1 ± 1.0	54.2 ± 1.0	65.8 ± 2.4	8.2 ± 1.9	41.8 ± 3.0	8.2 ± 1.9
SVMrbf_10MRMR	61.0 ± 0.4	55.3 ± 0.9	71.2 ± 0.4	11.0 ± 1.6	39.5 ± 2.0	11.0 ± 1.6
SVMrbf_100MRMR	64.2 ± 1.3	59.3 ± 1.5	73.3 ± 1.9	19.4 ± 2.3	45.3 ± 4.6	19.4 ± 2.3
SVMrbf_Full	63.2 ± 1.8	60.6 ± 0.4	70.9 ± 2.8	21.0 ± 0.7	49.4 ± 2.1	21.0 ± 0.7
RUSBoost_10MRMR	61.0 ± 1.1	59.6 ± 0.8	68.2 ± 1.8	18.5 ± 1.5	49.4 ± 1.5	18.5 ± 1.5
RUSBoost_100MRMR	61.9 ± 0.8	59.2 ± 1.4	70.0 ± 1.2	18.0 ± 2.5	47.7 ± 2.7	18.0 ± 2.5
RUSBoost_Full	60.7 ± 1.3	59.4 ± 1.5	67.8 ± 2.3	18.1 ± 2.8	49.1 ± 2.8	18.1 ± 2.8
CNN_dualInput	78.2 ± 0.9	74.5 ± 2.2	83.8 ± 1.0	51.2 ± 2.2	66.3 ± 3.4	51.2 ± 2.2
CNN_Spectrogram	77.8 ± 1.7	74.2 ± 1.6	83.4 ± 1.7	50.2 ± 3.0	66.0 ± 2.5	50.2 ± 3.0
CNN_melSpectrogram	78.0 ± 0.8	74.1 ± 1.7	83.7 ± 0.7	50.3 ± 2.3	65.9 ± 2.6	50.3 ± 2.3

Table B.23: Stratification for AKGC417L (wheezes vs. others)

Classifiers	Accuracy	AUC Crackle	F1 Crackle	MCC Crackle	F1 Other	MCC Other
LDA_10MRMR	68.7 ± 0.1	58.2 ± 0.4	79.0 ± 0.1	18.3 ± 0.7	38.2 ± 1.0	18.3 ± 0.7
LDA_100MRMR	68.0 ± 0.2	61.1 ± 1.1	77.5 ± 0.2	22.3 ± 1.8	44.7 ± 1.9	22.3 ± 1.8
LDA_Full	68.2 ± 1.0	61.9 ± 0.5	77.4 ± 1.2	23.8 ± 0.7	46.3 ± 1.2	23.8 ± 0.7
SVMrbf_10MRMR	69.4 ± 0.2	54.5 ± 0.5	80.7 ± 0.2	12.6 ± 1.1	26.0 ± 1.9	12.6 ± 1.1
SVMrbf_100MRMR	70.3 ± 0.5	62.7 ± 2.5	79.4 ± 0.8	26.3 ± 3.6	46.2 ± 5.7	26.3 ± 3.6
SVMrbf_Full	68.7 ± 1.2	62.5 ± 2.9	77.6 ± 2.0	25.0 ± 4.0	46.1 ± 7.8	25.0 ± 4.0
RUSBoost_10MRMR	68.3 ± 0.7	63.5 ± 1.0	77.0 ± 1.1	26.2 ± 1.2	48.9 ± 1.8	26.2 ± 1.2
RUSBoost_100MRMR	68.9 ± 0.6	63.7 ± 1.0	77.6 ± 0.8	26.9 ± 1.4	49.1 ± 2.0	26.9 ± 1.4
RUSBoost_Full	68.7 ± 0.6	64.0 ± 1.0	77.3 ± 0.8	27.2 ± 1.5	49.6 ± 1.5	27.2 ± 1.5
CNN_dualInput	86.4 ± 0.8	79.5 ± 2.3	90.9 ± 0.5	66.2 ± 1.8	72.9 ± 3.0	66.2 ± 1.8
CNN_Spectrogram	85.4 ± 0.9	78.7 ± 2.1	90.2 ± 0.8	63.7 ± 2.0	71.4 ± 2.7	63.7 ± 2.0
CNN_melSpectrogram	87.1 ± 0.5	80.7 ± 1.0	91.4 ± 0.4	67.9 ± 1.3	74.8 ± 1.3	67.9 ± 1.3

Table B.24: Stratification for AKGC417L (crackles vs. others)