POLYTECHNIC OF TURIN

Department of Electronics and Telecommunications

Master's Degree in Biomedical Engineering

Pulmonary Hypertension Detection via Deep Learning of Heart Sounds

Supervisors

Prof. Marco KNAFLITZ

Prof. Francesco RENNA

Candidate

Ilaria PULITO

Academic Year 2024/2025

Table of Contents

List of Tables

List of Figures

Acronyms

PH
Pulmonary Hypertension

ML
Machine Learning

mPAP
Mean Pulmonary Arterial Pressure

PAP
Pulmonary Arterial Pressure

RHC
Right Heart Catheterisation

PCG
Phonocardiogram

ECG
Echocardiogram

MFCC
Mel Frequency Cepstral Coefficients

DNN
Deep Neural Network

DL
Deep Learning

CNN
Convolutional Neural Network

ECHO
Echocardiogram

ROC

Receiver Operating Characteristic curve

auROC

Area Under the Receiver Operating Characteristic curve

TR

Tricuspid Regurgitation

Chapter 1

Introduction

1.1 Context e Motivation

Pulmonary hypertension is a pathological condition characterised by an abnormal increase in blood pressure in the pulmonary circulation. Normally a haemodynamic state characterised by mPAP at rest of 25 mm Hg or above may indicate the presence of PH [1].

PH can be classified into three primary categories: precapillary, postcapillary and mixed.

The precapillary form is characterised by increased pulmonary arterial pressures without the involvement of the left heart and it is often associated with idiopathic pulmonary arterial hypertension, chronic lung disease or chronic thromboembolism.

The postcapillary form, on the other hand, is linked to increased pressures of the left side of the heart, caused by ventricular dysfunction, valvular disease or cardiomyopathy.

Mixed hypertension combines pre-capillary and post-capillary mechanisms, and it often appears in the context of advanced diseases [2].

In general, the main causes of PH include vascular changes, chronic lung disease, cardiac dysfunction, chronic thromboembolism and systemic conditions, requiring accurate diagnosis for effective treatment. [3]

Pulmonary hypertension manifests mainly with progressive and exertional difficulty in breathing (dyspnoea), which is the cardinal symptom, often accompanied by fatigue and exhaustion.

Initial symptoms are generally non-specific, which is why diagnosis may be delayed by many months or even years.

As the disease progresses, symptoms worsen and new manifestations may appear, such as bending dyspnoea (bendopnoea) and syncopes, the latter being particularly frequent during or immediately after physical exertion [1].

Pulmonary hypertension has a higher prevalence than commonly assumed; it is estimated to affect about 1% of the global population, with the prevalence increasing to 10% in people over 65 years of age [1].

Despite significant therapeutic advances in recent decades, PH remains a clinically relevant condition, characterized by persistent symptoms and an elevated risk of mortality.

Notably, the three-year survival rate has markedly improved, rising from approximately 40% in the 1980s to 70-80% today [4].

However, the management of certain PH subtypes, particularly those associated with heart or lung disease, continues to be an area of active research. While clinical investigations have contributed to better outcomes, further studies are needed to optimize treatment strategies for these complex cases [5].

In general, traditional and contemporary diagnostic techniques still play a crucial role in identifying PH, classifying its subtypes, and assessing its severity, thereby suggesting appropriate management strategies.

Echocardiography is the main non-invasive procedure for diagnosing pulmonary hypertension, as it allows the presence of the disease or an overload of the right heart to be suspected. This examination is based on the assessment of cardiac function and the estimation of right ventricular pressure, a parameter that, however, can be inaccurate due to various technical and physiological factors [6].

Despite this limitation, the presence of obvious signs of right heart overload, such as right ventricular dilatation, paradoxical movement of the interventricular septum or dilatation of the inferior vena cava, offers important clinical indications [1][7].

These first elements, integrated with other clinical and instrumental data, guide the physician in the diagnostic pathway and in the possible need for further investigation, as shown in the algorithm present in Fig.1.1.

After the initial clinical evaluation and the performance of an echocardiography to estimate the right ventricular pressure, further investigations are performed based on the findings [6, 8].

As a matter of fact, the definitive diagnosis of PH can only be confirmed by RHC, an invasive diagnostic procedure considered the gold standard for the confirmation of pulmonary hypertension; in fact, it allows direct measurement of pulmonary artery pressure and pulmonary vascular resistance, which represent the key variables in the diagnosis, as they provide a direct assessment of the severity of hypertension and haemodynamic impairment of the pulmonary circulation [9].

However, its clinical usefulness differs depending on the type of pulmonary hypertension suspected.

Indeed, it is essential in cases of suspected pulmonary arterial hypertension or chronic thromboembolic pulmonary hypertension, where an accurate diagnosis is essential to establish the severity of the condition and to assess suitability for specific treatment. Moreover, in these cases, right heart catheterisation allows to make a distinction between pre-capillary and post-capillary forms of the disease, improving diagnostic accuracy compared to non-invasive methods [2].

On the other hand, in some cases RHC appears often not necessary: for patients with chronic left heart failure or lung disease the results of the procedure would not change the treatment already planned for these conditions [1].

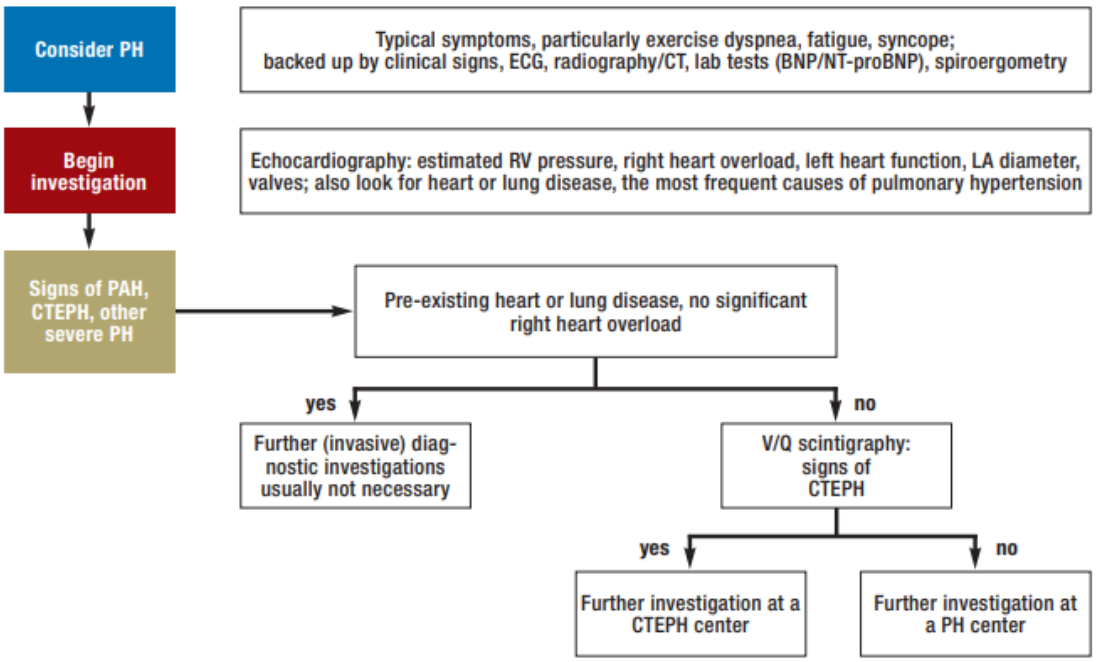


Figure 1.1: Diagnostic algorithm for PH Reproduced from [1]. BNP, Brain natriuretic peptide; CT, computed tomography; CTEPH, chronic thromboembolic pulmonary hypertension; ECG, electrocardi - ography; PAH, pulmonary arterial hypertension; PH, pulmonary hypertension; LA, left atrium; NT-proBNP, N-terminal fragment of pro-brain natriuretic peptide; RV, right ventricle; V/Q, ventilation/perfusion

However, both techniques have significant limitations.

Although RHC represents the gold standard examination for direct measurement of PAP, it is an invasive, expensive procedure and it is only suitable for patients with a high probability of PH [10].

On the other hand, Doppler echocardiography, despite its low cost and minimal risk, shows significant limitations in estimating PAP.

As a matter of fact, it cannot provide reliable measurements in approximately 50% of patients with normal PAP [11] and has an average margin of error of 30% compared to right heart catheterisation [12].

These limitations underline the need to develop innovative, non-invasive diagnostic methods that are more accessible and less costly.

One promising alternative approach is digital cardiac auscultation, which could be used as a screening tool for pulmonary hypertension, especially in clinical settings where the disease is asymptomatic.

In addition to requiring minimal training, this method is less expensive and more widely applicable, providing a practical option for early detection of the disease [13].

In recent years indeed, the widespread of digital cardiac auscultation has increased due to the progresses made in the field of electronic phonendoscope technology and due to the integration of artificial intelligence algorithms, which improve the diagnostic capability and early detection of cardiopulmonary diseases.

Recent studies show that the use of digital stethoscopes combined with sound analysis has made diagnosis more accessible, especially in settings with limited resources [14]. Furthermore, the increasing use of portable devices for monitoring heart sounds has made this technology a practical and economically viable solution for pathology screening [15].

Therefore there is substantial evidence indicating that in the last years, there has been an unmet and significant need for early diagnostic methods for PH that are at the same time non-invasive, reliable and cost-effective [16].

Cardiac auscultation with digital stethoscopes has indeed represented a faster and more timely PH screening method, especially in contexts when an immediate care is necessary and where the disease may be present but asymptomatic.

This method is a low-cost option, requires minimal training and can be easily applied in a variety of clinical settings [17].

1.2 Objectives

The general aim of this project is to develop an innovative algorithm for the diagnosis of PH based on the analysis of phonocardiographic signals captured at the site of the pulmonary valve.

To achieve this goal, several aspects related to the analysis and processing of phonocardiographic signals have been explored. Specifically, the main contributions of this thesis can be summarized as follows:

- Development and application of a pre-processing pipeline for phonocardiographic signals, with the aim of preserving only those components of the signal that are most informative and relevant for diagnostic purposes, while reducing noise and non-significant ones that could compromise the performance of the analysis models.

- Evaluation of the quality of phonocardiographic signals with the aim of providing an objective criterion for selecting the most reliable data, reducing the impact of poor-quality signals on the analysis results.

- Development of several deep learning models with different levels of complexity, with the aim of automated diagnosis of PH from the analysis of heart sounds. This process included experimentation with various neural network architectures, as well as optimisation of hyperparameters to maximise model performance.

- Validation of the models on a real-world dataset of heart sounds acquired under realistic conditions, with the purpose of testing their effectiveness in realistic scenarios.

  The dataset used to train and validate the model was collected in an ambulatory auscultation setting in collaboration with the Cardiac Department of the Unidade Local de Saúde Gaia e Espinho and contains PCG and ECG data from 190 patients, together with clinical and physiological information.

- Comparison between the application of the model on datasets with different sizes and ground truth, considering both RHC, the gold standard method for diagnosing PH, and ECHO, a less invasive but commonly used technique in clinical settings.

  The variables used to determine the ground truth are a level of PH severity derived from the echocardiographic exam, and the mean pulmonary artery pressure value, which allows for the identification of potential pulmonary hypertension in the case of patients who underwent the RHC examination.

  This comparison enabled an evaluation of the differences in model performance depending on the source of the reference truth.

- Comparative analysis between different post-processing techniques in order to determine the most effective strategies to improve the quality of the model output and optimise the interpretability of the results, thus facilitating a possible clinical application of the proposed method.

Together, these contributions enabled the development of an innovative approach for the automated analysis of heart sounds, with the potential to support clinicians in diagnosing pulmonary hypertension by enhancing the reliability and speed of assessment compared to traditional methods.

The proposed method represents a significant step toward a non-invasive screening system for PH, which could be particularly valuable in low-cost clinical settings.
By enabling early detection of at-risk patients and directing them to further examinations only when necessary, it could also reduce reliance on invasive diagnostic tools, improve disease management, and potentially increase patient survival through timely intervention.

1.3 Thesis Outline

The present work is structured to provide a comprehensive review of the methodology, results and discussion on the use of deep learning techniques for the detection of pulmonary hypertension via heart sounds.

Chapter 2 is dedicated to the Background, where the fundamental concepts and theoretical foundations necessary for the development of this work are presented.
This chapter introduces the key principles of the cardiac function, the physiological basis of heart sounds, and the role of artificial intelligence in medical diagnostics.

Chapter 3 presents the Literature Review, providing an overview of previous studies and the current state of the art on the topic. The datasets, approaches, features extracted from the data and results obtained from the different studies are analysed and compared, highlighting the main differences between them, in order to contextualize the problem addressed in this work.

Chapter 4 on the other hand, describes the Materials and Methods, illustrating the dataset composition, pre-processing techniques and deep learning architecture used in this study for classification. The experimental setup, different configurations of the dataset, classification strategies, evaluation metrics and post-processing techniques applied to improve the performance of the model are also discussed.

Chapter 5 presents the Results and the Discussion, evaluating the performance of the proposed algorithm in different dataset configurations. The chapter explores the stability of the model through accuracy and loss curves, confusion matrices and evaluation metrics. It also analyses the impact of different post-processing techniques and the influence of signal quality on classification performance.

The section provides a descriptive review and in-depth interpretation of all the results obtained, in order to facilitate their understanding and improve their interpretability, while ensuring consistency with the objectives estabilished for this work.

Finally, Chapter 6 is dedicated to the Conclusions, in which the main results are critically evaluated in the context of previous research. This chapter reflects on the strengths and limitations of the proposed method, discusses potential clinical applications and suggests possible future developments to improve the non-invasive diagnosis of pulmonary hypertension using deep learning techniques.

Chapter 2

Background

2.1 Cardiac Anatomy Overview

The heart is a fundamental organ that serves as double pump, granting blood flow through the pulmonary and systemic circulation [18].

The pulmonary circulation transports deoxygenated blood to the lungs, where gas exchange takes place: oxygen enters the blood while carbon dioxide is removed. In contrast, the systemic circulation distributes oxygen-rich blood to the whole body, supplying oxygen and nutrients to the tissues before returning the deoxygenated blood back to the heart (Fig. 2.1) [19].

This dual circuit ensures continuous oxygenation and adequate distribution of substances needed to maintain homeostasis and cellular functions.

It is composed of two sections: the right heart, that pumps blood through the lungs, and the left heart that sends it into the systemic circulation, providing blood supply to the other organs of the body [21].
Each one of these two is pulsatile and consists of two chambers: an atrium and a ventricle. The atrium fill the ventricle, which provides the main pumping force pushing blood either through the pulmonary circulation or the systemic circulation [21]. The anatomy of the heart is shown in Figure 2.2.
The heart is enclosed in the pericardium, a fibroserous sac comprising three concentric layers that protects the heart and secure it in the thoracic cavity [18].

Blood flow in the heart is regulated by the four heart valves, which ensure the unidirectional passage of blood between the heart chambers and the great vessels [18]. Cardiac tones are generated with the sudden closure of the valve leaflets and can be auscultated at specific points in the chest:

- mitral valve – cardiac apex
- tricuspid valve – over the lower part of sternum and right sternal edge at the fifth intercostal space
- aortic valve – right sternal edge at second intercostal space
- pulmonary valve – left sternal edge at third intercostal space.

Besides its anatomical structure, the heart is characterised by an intrinsic mechanism of cardiac rhythmicity, which allows the continuous transmission of

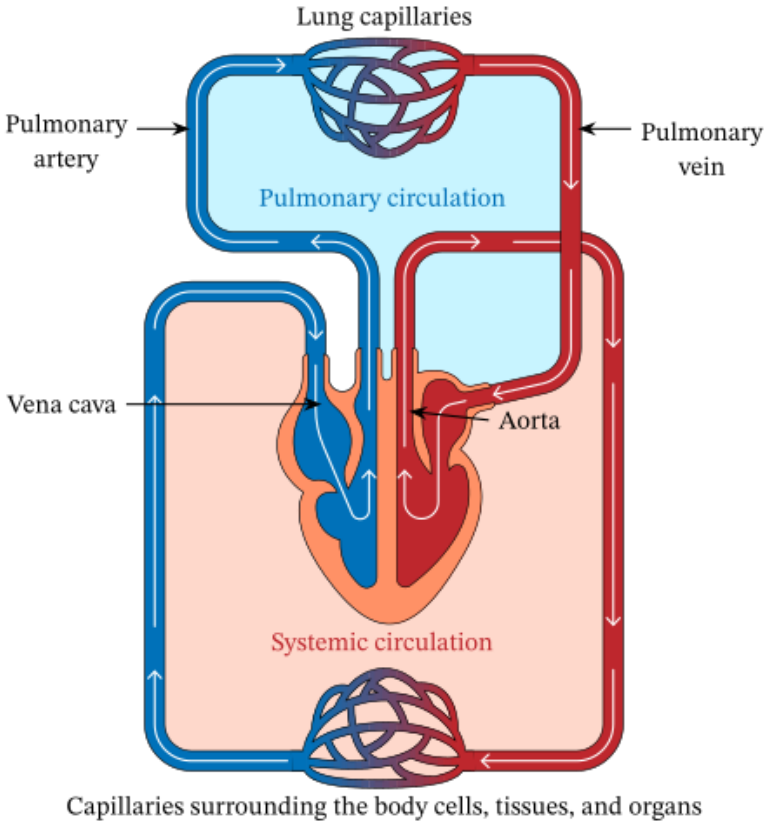


Figure 2.1: Diagram illustrating the human circulatory system, highlighting the pulmonary (blue) and systemic (red) circulation, with the heart and aorta as central components. Reproduced from [20].

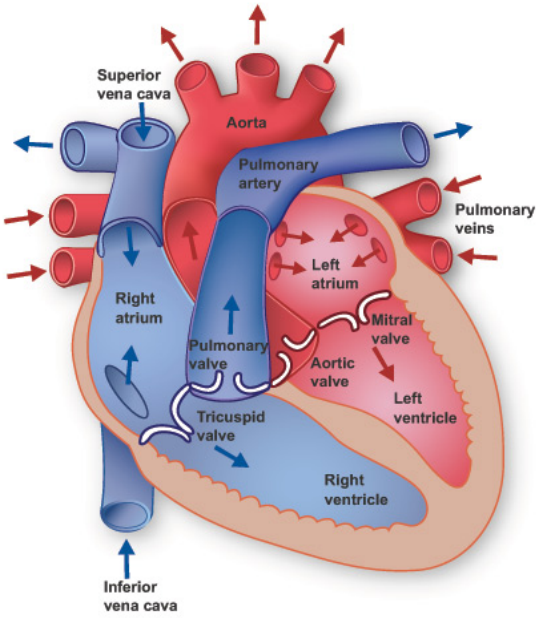


Figure 2.2: Heart Anatomy. Reproduced from [22].

electrical impulses that are necessary to coordinate the contraction of the heart muscle.

This property is essential to ensure an effective and continuous heartbeat, which is fundamental to the body's circulatory function [18].

2.2 Cardiac Cycle

The impulse-conducting system of the heart is ensured by the presence of conductive fibres (Figure 2.3) whose main control points include:

- sinoatrial node
- atrioventricular node
- atrioventricular bundle of His
- right and left branches of the atrioventricular bundle
- subendocardial fibres of Purkinje [18].

The first node that is spontanously triggered to initiate the cardiac contraction is the sinoatrial node; subsequently the impulse reaches the atrioventricular node with a certain delay that ensure proper conduction timing and an adequate filling of the ventricles before the ventricular contraction begins [18]. This phase is known as diastole, and concerns the process in which the heart chambers relax and become filled with blood.

The His bundle originates from the atrioventricular node and carries the electrical impulse through the interventricular septum, where it divides into the right and left branches, where the Purkinje fibres originate [18]. These specialized fibers are distributed over the portion of myocardium that surrounds the ventricles and trigger systole, the phase in which they contract to eject blood into the pulmonary and systemic circulation.

Thus, the entire cardiac surface receives the electrical impulse necessary to stimulate the contraction of the myocardium, the muscolar tissue of the heart. This coordinated cycle of diastole and systole enables the exchange of blood between the chambers and its subsequent ejection into the pulmonary circulation from the right heart and into the systemic circulation from the left heart.

The electrical and mechanical activity of the heart can be monitored using specific diagnostic tools, such as the electrocardiogram (ECG) and phonocardiogram (PCG). These techniques offer valuable information on the relationship between electrical impulses, myocardial contraction and heart sounds during the cardiac cycle.

2.2.1 Electrocardiogram (ECG): the Heart's Electrical Activity

If electrodes are placed on the skin on opposite sides of the heart, electrical potentials generated by the currents can be recorded, giving rise to the electrocardiogram (ECG) [21].

A normal ECG is normally composed of a P-wave, a QRS complex, a T wave, and a series of important intervals, each representing specific phases of cardiac electrical activation.
The P wave represents atria's depolarization, before their contraction begins. It

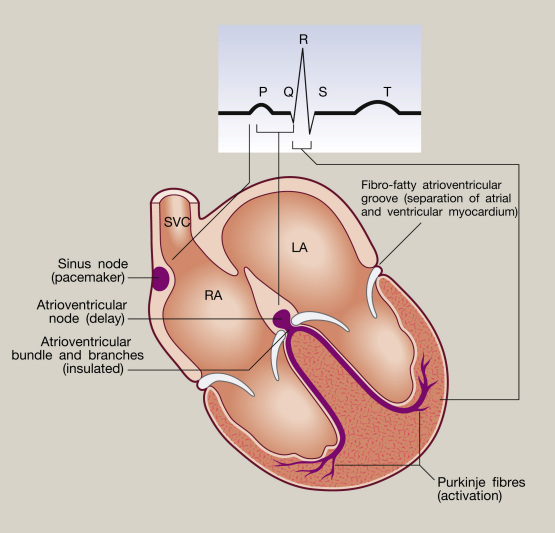


Figure 2.3: Diagram of the cardiac conduction system, and the corresponding electrocardiogram (ECG) waveform. Reproduced from [18].

manifests itself as a small positive deflection, that in normal conditions lasts 0,08-0,10 seconds. It corresponds to the electrical impulse generated by the sinoatrial node, which initiates the cardiac cycle.

QRS complex on the other hand indicates the ventricles' depolarization, which represents the moment when they contract to pump blood out of the heart to the lungs and to the rest of the body.
This phase corresponds to the flow of the electrical impulse through the Purkinje fibers.
The QRS complex normally lasts less then 0.12 seconds and it is characterised by a sharp positive deflection which includes:

- Q-wave: First negative deflection of the complex, if present.
- R wave: First positive deflection.
- S wave: Negative deflection following the R wave [21].

The T wave is known as a repolarization wave because it is caused by potentials generated as the ventricles recover from the initial depolarization. It is characterized by a positive deflection that is subsequent to the QRS complex.

2.2.2 Phonocardiogram (PCG): capturing Heart Sounds

The recording of sounds produced by the heart during a cardiac cycle is called phonocardiogram. It can be obtained by putting a microphone on the chest, recording the audible vibrations and digitizing them for further analysis, or through an acoustic or electronic stethoscope [23].

Listening to a normal cardiac cycle, what one hears is the classic repetition of the 'lub-dub' sounds: these represent respectively the closure of the atrioventricular (mitral and tricuspid) valves at the beginning of the systole and the closure of the semilunar (aortic and pulmonary) valves at the end of systole [21]. These are the fundamental heart sounds and are always audible during ascultation, where the 'lub' corresponds to the S1 and the 'dub', which is normally louder, shorter and sharper, corresponds to the S2 [24].

Among these, additional sounds S3 and S4 can sometimes be captured during the recording of a PCG: both are very low-frequency sounds that can sometimes be indicators of pathologies [24].

The S3 sound, often described as a low-frequency "ventricular gallop", occurs during early diastole due to rapid ventricular filling and is more commonly heard in conditions like heart failure and volume overload.

The S4 sound, or "atrial gallop", appears in late diastole and is usually associated with reduced ventricular compliance, often seen in conditions such as left ventricular hypertrophy or myocardial ischemia [25].

Heart murmurs are abnormal sound vibrations produced by the turbulent flow of blood through the heart and the vessels [21].

They can be physiological and therefore harmless, or pathological and are classified as systolic, diastolic or continuous, depending on when they appear during the cardiac cycle [21].

Pathological murmurs can be indicative of conditions such as stenosis or valvular insufficiency. In the phonocardiogram, heart murmurs appear as higher frequency signals than normal heart tones (S1-S4) and are analysed to distinguish between different clinical conditions [21].

An important correlation is observed between heart murmurs and pulmonary hypertension.

In patients with pulmonary hypertension, increased blood pressure in the lungs can lead to pulmonary valve regurgitation or functional tricuspid valve stenosis, both of which are associated with audible diastolic or systolic murmurs in the PCG [26].

A typical sign is Graham Steell's murmur, a high-frequency diastolic murmur caused by pulmonary insufficiency secondary to pulmonary hypertension [26].

In addition, in advanced cases of PH, the second heart tone (S2) appears accentuated and doubled, due to increased pressure in the pulmonary artery and delayed closure of the pulmonary valve [26].

PCG analysis therefore proves to be a useful tool for monitoring this condition,

facilitating early diagnosis and assessment of the severity of the disease.

An illustrative explanation of a PCG signal and its components is shown in Fig. 2.4.

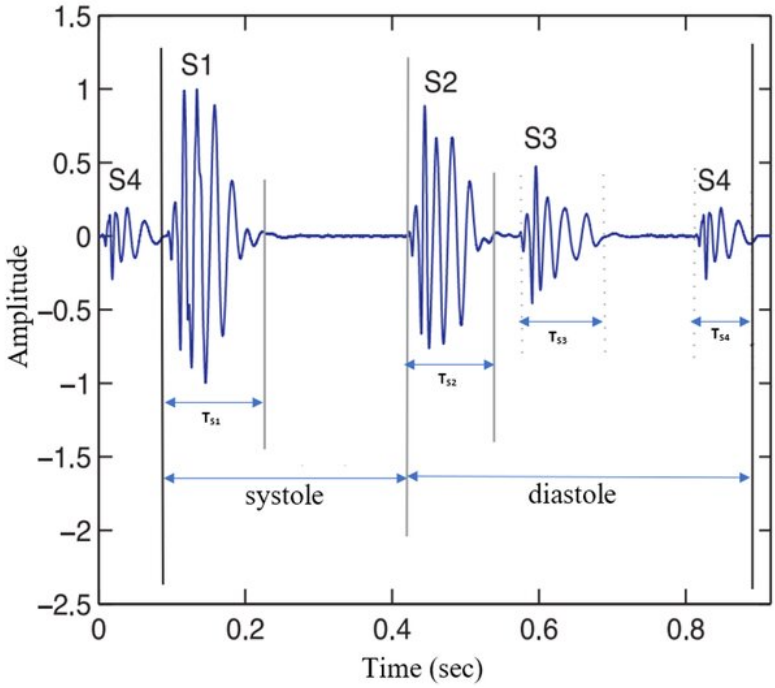


Figure 2.4: Typical waveform of a phonocardiogram (PCG) signal and its components: S1 (T S1 = 70-150 ms), S2 (T S2 = 60-120 ms), S3 (T S3 = 40-100 ms), and S4 (T S4 = 40-80 ms). Reproduced from [27].

It is clear from these explanations that the PCG signal is closely related to the ECG signal, as both are directly linked to the electrical activity of the heart and its consequences on cardiac mechanical function.

The electromechanical phases of the cardiac cycle are properly shown in the Wiggers diagram (Figure 2.5), a graphical representation of the relationship between electrical activity, volume, pressure, and sounds of the heart during every beat.

As a consequence, combining PCG and ECG may allow a simultaneous clinical assessment of the mechanical and electrical condition of the heart, potentially improving the accuracy of an initial diagnosis of a cardiovascular disease [24].

Furthermore, both signals can be acquired using non-invasive and low-complexity devices: although these methods do not replace more advanced techniques, they still offer valuable preliminary information, guiding further diagnostic investigations and facilitating clinical decisions, especially in contexts or situations where advanced imaging technologies are not readily available [28].

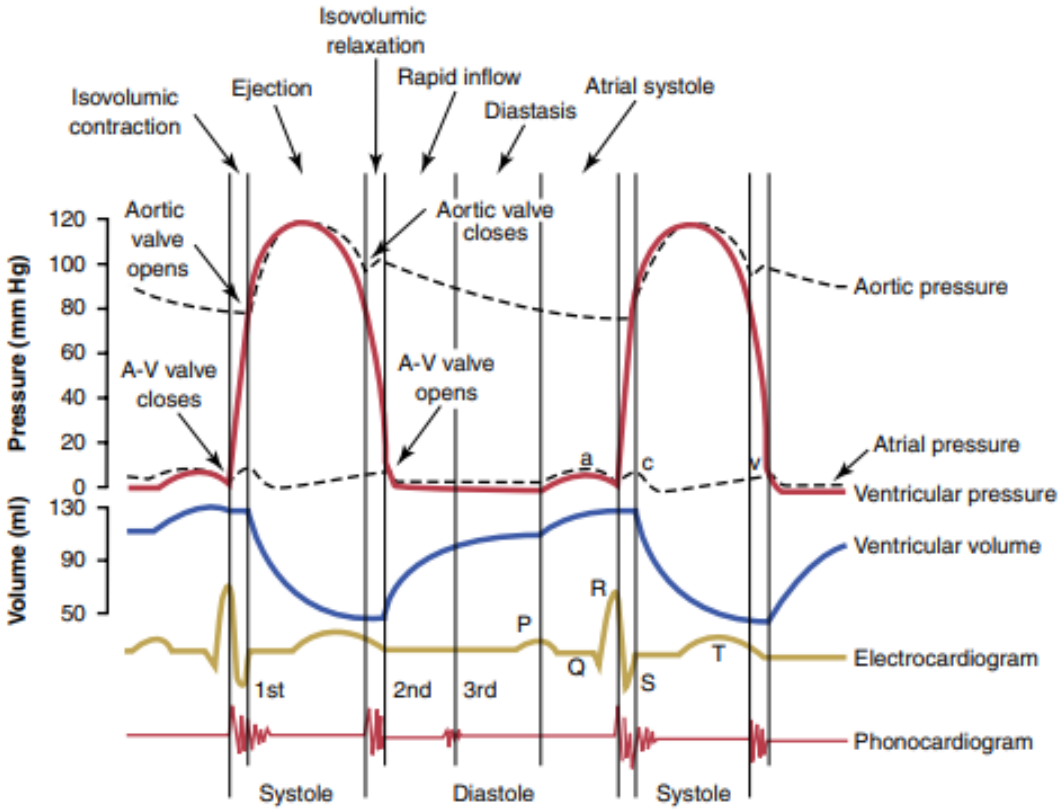


Figure 2.5: Wiggers diagram illustrating the relationship between pressure, volume, and electrical activity in the cardiac cycle. Reproduced from [21].

2.2.3 Heart Sounds Alterations in Pulmonary Hypertension

The analysis of the PCG signal can reveal informative and essential alterations that can suggest the presence of different cardiovascular diseases.

In particular, the second heart tone (S2) often reveals in its pattern key indicative features for diagnosing pulmonary hypertension: abnormal intensity, quality and division in components may serve as auscultatory signals for its detection [29].

Indeed, under normal conditions S2 is composed by the aortic (A2) and the pulmonary (P2) components, which are temporally separated during inspiration due to the increased pulmonary venous return [30], giving rise to a temporal separation between 30 and 80 ms [31].

During expiration on the other hand, the A2-P2 interval narrows to a separation of 15 ms, to the point that only a single sound is normally heard [32].

Figure 2.6 clearly shows this physiological difference: the upper part shows an S2 with a wide split, characteristic of inspiration, while the lower part shows an S2 with a narrow split, typical of exhalation, where the A2 and P2 components are hardly distinguishable.

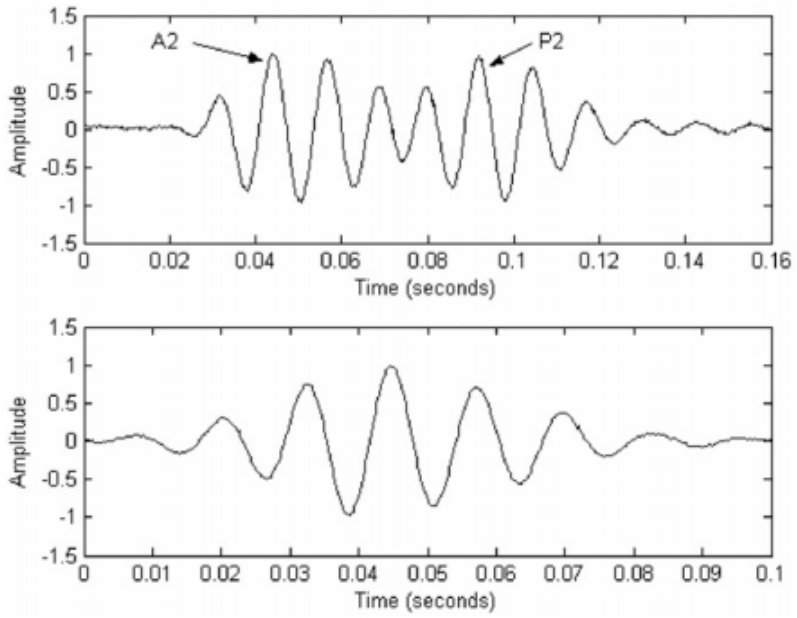


Figure 2.6: An illustrative example of a widely splitted S2 (top) and a narrowly splitted S2 (bottom). Reproduced from [31].

In patients with pulmonary hypertension, the separation time between the two components of the second heart tone tends to increase.

This phenomenon is due to the increase in PAP, which delays the closure of the pulmonary valve compared to the aortic valve.

P2 is therefore further delayed compared to A2, making the separation more apparent even during exhalation. This phenomenon, known as wide S2 splitting, can be used as a non-invasive indicator to monitor pulmonary arterial pressure and identify pathological conditions associated with this haemodynamic alteration at an early stage [31].

More in general, the use of modern diagnostic techniques capable of detecting these kinds of morphological alterations, such as deep learning models trained on datasets of heart sounds, can become a key tool in early diagnosis or screening: these type of algorithms as a matter of fact, can allow the detection of abnormalities with higher sensitivity and specificity compared to traditional auscultation.

2.3 Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCC) are a fundamental technique for the analysis of audio signals and also for the wide application in the processing of cardiac sounds and photoplethysmographic signals.

They are based on the Mel scale, which reflects human perception of sound frequencies, allowing significant characteristics to be extracted from signals [33].

The MFCC calculation process involves several steps. After a pre-emphasis

phase to emphasize the high frequencies, the signal is divided into time windows and transformed into the power spectrum by means of the Fourier transform. Subsequently, a series of triangular filters on Mel scale reduces the complexity of the spectrum, and the logarithm of the energy of the filters is then subjected to the Discrete Cosine Transform, producing a set of numerical coefficients that represent the signal in a compact and effective way [34].

A common extension of MFCCs are Delta and Delta-Delta MFCCs, which are calculated from static MFCCs to capture temporal variations of the signal.

Delta MFCCs represent the first derivative of the original MFCCs in the time domain and provide information on the rate of change of spectral characteristics. This allows significant signal dynamics to be identified, improving the ability to discriminate between different sound classes.

Delta-Delta MFCCs, on the other hand, represent the second derivative, i.e. the variation of the variation of the original MFCCs. This information is useful for capturing more complex variations in the audio signal, such as the acceleration of frequency variations [35, 36].

The inclusion of biologically inspired features, such as MFCCs, in Machine Learning and Deep Learning models can significantly enhance performance in classifying pathological conditions like pulmonary stenosis and pulmonary hypertension, as they provide a more comprehensive representation of the temporal dynamics of heart sounds [37]. Since MFCCs resemble the resolution of the human auditory system, they have been proven effective in differentiating between distinct sound signals [38], thereby improving the performance of ML models and CNNs in biomedical classification tasks [39].

Figure 2.7 illustrates the differences in MFCC representations between a normal PCG signal and a PCG signal containing murmurs. The top illustration shows the MFCCs of a normal heart sound, while the bottom one, associated with pathological heart sounds, displays irregularities and spectral distortions. The distinctive spectral patterns captured by MFCCs enhance their utility in automated classification models, enabling more accurate detection of cardiovascular disorders.

2.4 Deep Learning

In the last years, Deep Learning (DL) has demostrated to be a key technique in the field of artificial intelligence due to its ability to automatically extract complex features from data.

Deep Learning is based on deep artificial neural networks consisting of multiple layers of neurons that process information in a structured manner that is inspired by the way our brain works.

These patterns are particularly effective in applications of image recognition, language analysis and biomedical signal processing, including phonocardiographic signals for heart sound recognition [40].

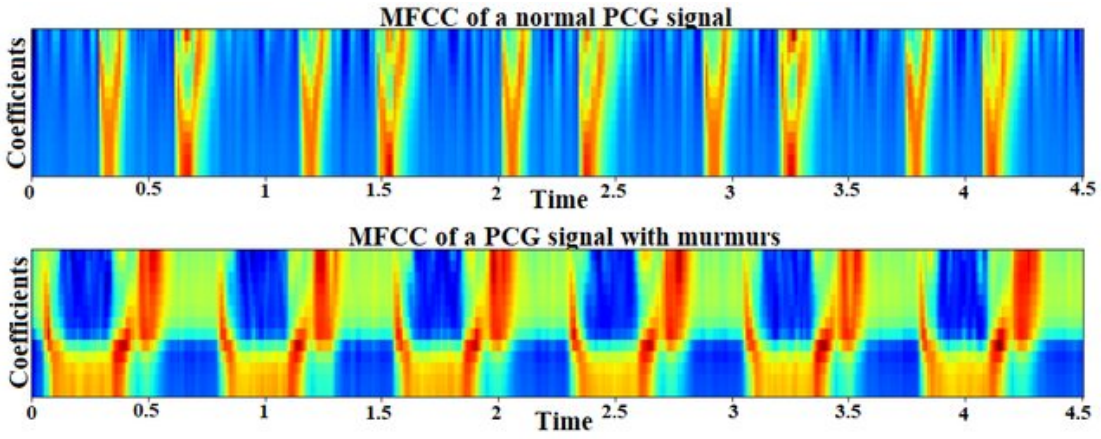


Figure 2.7: Comparison of MFCC representations for a normal PCG signal (top) and a PCG signal with murmurs (bottom). The spectral differences highlight the ability of MFCCs to capture pathological variations in heart sounds. Reproduced from [38].

From Figure 2.8, it can be seen that input data is processed through layers of neurons, starting from the input layer to the final layer, which produces the model prediction. Each layer may hold different neurons, connected to each other by weights, whose role is to adapt to the input data in order to minimise the final error.

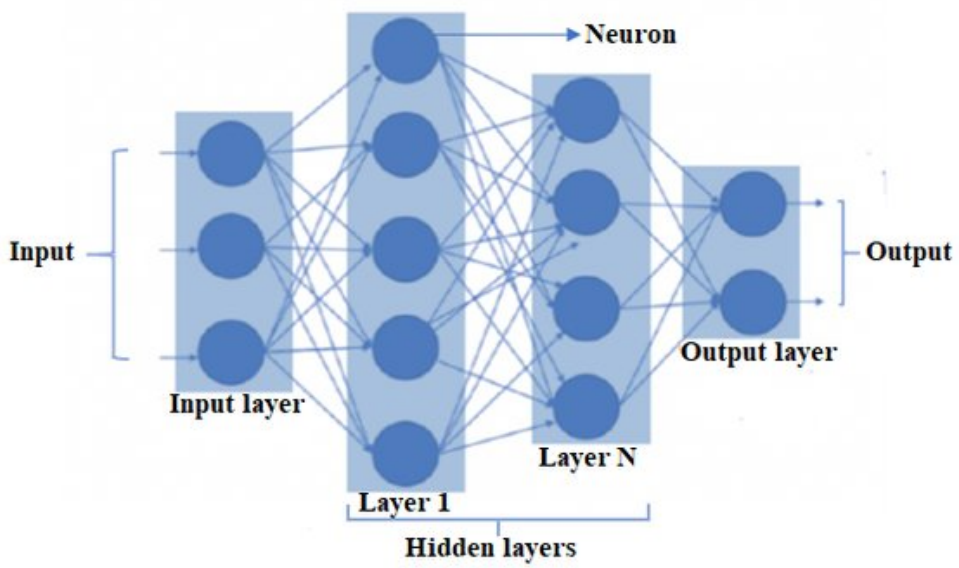


Figure 2.8: A Deep Neural Network with N layers.

During the model training, the DNN uses the backpropagation algorithm, which indeed iteratively updates the weights to minimise error and improve the network's accuracy [41].

A key element in deep neural networks is the activation function, which introduces non-linearity into the model and allows the network to learn complex representations from the data.

Each neuron in a hidden layer applies an activation function to the weighted sum of its inputs, determining whether and how to transmit the signal to the next layer.

Among the most common activation functions is the ReLU (Rectified Linear Unit), which returns zero for negative values and identity for positive values, making the computation efficient [42].

2.4.1 Convolutional Neural Network

One of the most widely used architectures in Deep Learning is the Convolutional Neural Network (CNN), initially designed for image recognition, but now also widely applied to the analysis of biomedical signals [43].

CNNs use convolutional filters to automatically extract features from the data, reducing the need for manual feature extraction. It is been proven to be convenient to use time-frequency representations of the signals in order to effectively capture and convey their information, making them more suitable for feature extraction and pattern recognition in CNNs [44].

The working of a CNN is based on a series of convolutional operations, which apply filters on the input data to identify relevant features. Each filter scans the input, extracting initial information such as frequency variations and recurring temporal patterns.

This process occurs in the first convolutional layers, which operate on small portions of the input, detecting low-level features such as intensity variations and rapid transitions between frequency bands.

As the information passes through successive layers, the network learns increasingly abstract and deeper features, combining the lower-level information to construct a more complex representation of the signal [45].

After the convolutional layers, the CNN uses pooling layers, which reduce the dimensionality of the input while retaining the most meaningful information. This improves the efficiency of the model and makes it more robust to minor variations in the data. Finally, the processed data is transformed into a vector and passes through fully connected layers, which allows the network to make the final prediction [45].

An example of a CNN architecture is shown in Fig. 2.9.

The training of the CNN is done through the backpropagation algorithm already discussed in section 2.4, which updates the weights of the convolutional filters and fully connected layers to minimise the prediction error with respect to the desired output.

As in deep learning models in general, convolutional neural networks also employ activation functions, key elements in these type of algorithms that guarantee

the non-linearity of the model and allow the learning of complex representations [42].

This architecture allows CNNs to adapt effectively to the structure of the input data, identifying complex patterns with a high degree of accuracy. The use of convolutions enables the detection of spatial correlations in feature matrices, while the combination of pooling and fully connected layers ensures effective synthesis of information, optimising model performance.

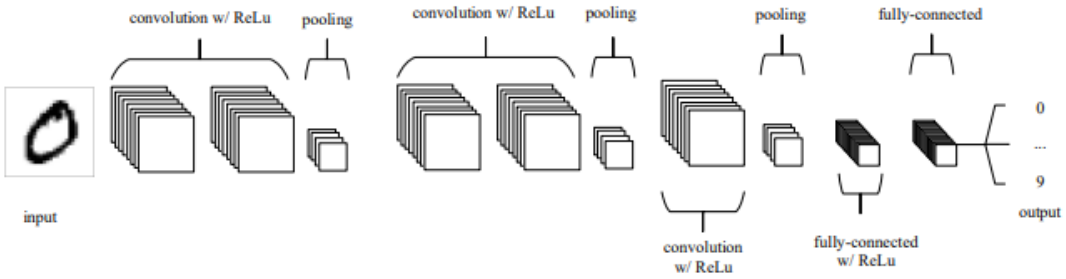


Figure 2.9: A common form of CNN architecture. Reproduced from [46].

Chapter 3

Literature Review

Over the past several years, significant attention has been devoted to the development of a non-invasive method for the detection of pulmonary hypertension: this can involve the analysis of heart sounds, which has proven to be a useful and innovative tool for an early diagnosis [13].

The following section aims to compare and analyse several major studies that have implemented different strategies to produce a consistent and reliable model.

The main aspects that distinguish and allow comparison among the different approaches are:

1. The portion of signal analysed in the study.
2. The type of features extracted from the data.
3. The chosen method, such as statistical, machine learning, or deep learning.
4. The dataset used for the purpose of the investigation.

Table 1 compares a collection of studies that explore different methods with the purpose of diagnosing pulmonary hypertension by analysing heart sounds, focusing on some of the key aspects mentioned above; these are outlined in the following subsections.

3.1 Portion of Signal Analysed

A first relevant distinction is the choice of the signal's portion that is used in the study.

Some authors use the entire phonocardiographic signal [47, 48, 49, 50, 51], while others focus on the component associated with the second heart sound (S2), corresponding to the closure of the aortic and pulmonary valves. Some other studies further divide this second heart sound portion into its aortic and pulmonary subcomponents to obtain a more detailed vision.

Indeed, the analysis of the A2 and P2 components of the second heart tone (S2), can lead to relevant information regarding pulmonary arterial pressure (PAP)

[52], which is a key parameter closely associated with the risk of developing PH [53].

On the contrary, some other works propose the employment of the entire signal, since they argue that erroneous segmentation can strongly influence the quality of the results when large databases are used [49].

However, an approach that focuses on isolating the S2 component, which is considered informative for assessing pulmonary hypertension [31], is the most common method in the analysed studies.

3.2 Features extracted from the data

In order to isolate the aortic-pulmonary component [29, 47], certain studies suggest an energy-based analysis of the phonocardiographic signal [52], coupled with a thresholding technique.

Other research studies exploit time synchronisation between the PCG signal and the T-wave in the electrocardiogram (ECG) when available [16], a technique that enables respectively a more accurate identification of the second heart tone occurrence and a precise windowing of a heart-cycle. This strategy is based on the evidence that since the T wave represents the end of ventricular repolarisation, it also precedes temporally the moment of closure of the semilunar valves, which matches the S2 sound in the PCG signal.

Regardless of the technique adopted to extract the second heart tones (S2), many of these approaches tend to confine them within time windows of fixed duration. The aim is to ensure homogeneous input data to neural networks, in the studies that use a deep learning strategy: this facilitates the better capture of relevant patterns and the extraction of meaningful signal features.

Some of the studies argue that a pre-processing technique is necessary to enhance the quality of the phonocardiograms: the most common one involves filtering the signal in a frequency range generally between 10 Hz and 400 Hz, since most of the relevant information of heart sounds is typically between 50 Hz and 150 Hz [54]. Another frequently adopted technique regards downsampling the signal, which aims to delete unhelpful information, thus simplifying the analysis and reducing the computational load [47, 55].

Once the portion of the signal to be analysed has been extracted and eventually preprocessed, a key step consists in specifically extracting the features that enable the identification of information relevant to the diagnosis of pulmonary hypertension.

Regarding the type of features extracted from the data, a big distinction concerns the choice of hand-crafted or deep features, i.e., learned automatically by neural networks and adapted to the dataset used in the study.

When extracted from the deep learning models, features are often able to automatically capture the most relevant information from the PCG signal for the diagnosis of pulmonary hypertension, without requiring manual selection. Furthermore, since they are adapted to the specific dataset, these features can capture peculiar properties of the signal not otherwise detectable, that may be particularly predictive for the problem under investigation. Besides that, many studies rely on hand-crafted features, including entropy, signal intensity, complexity and strength [50, 29, 56].

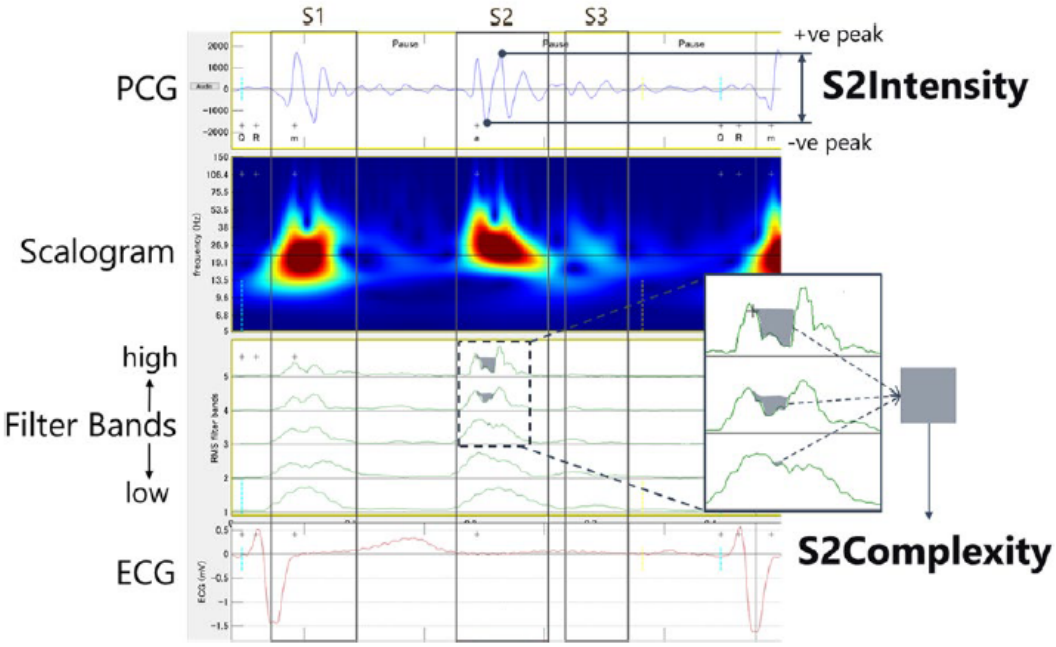


Figure 3.1: Visual Description of Intensity and Complexity as hand-crafted features used in [56]

Signal entropy and complexity are measures of the signal irregularity and they can detect variations in heart sound patterns, which may suggest the presence of pulmonary hypertension [50].

While intensity is a measure related to the signal's amplitude, what in Yamakawa's study is defined as "strength" is a probability score based on acoustic features that reflect the possible presence of an audible S3 and S4 on standard auscultation [56].

These hand-crafted features in most of the studies have been then statistically compared with disease severity levels [11], or with hemodynamic variables [29], to identify associations with the likelihood of having or not pulmonary hypertension and to determine which features are most informative in this context.

In other cases, time-frequency transforms such as the Fourier transform or the continuous wavelet transform (CWT) are used to obtain representations that convey both temporal and frequency-related characteristics [49, 57, 58]: this

can be particularly effective when studying specific frequency bands of the S2 signal that may correspond to pathological changes in pulmonary artery pressure [57] .

One approach used for feature extraction from phonocardiographic signals is based on the wavelet transform, which allows the frequency content of heart sounds to be examined in the time-frequency domain.

Huang *et Al.* employed this technique to extract parameters such as the energy and frequency at the maximum amplitude of the second heart tone (S2), obtaining promising results to distinguish different levels of pulmonary hypertension.

In Figure 3.2, the acoustic cardiographic output of a control patient (A) and a patient with pulmonary hypertension (B) is shown, highlighting how the extracted features can be used to differentiate between the two conditions.

This statistical approach showed encouraging results in detecting severe PH, with an area under the ROC curve (auROC) of 0.882.

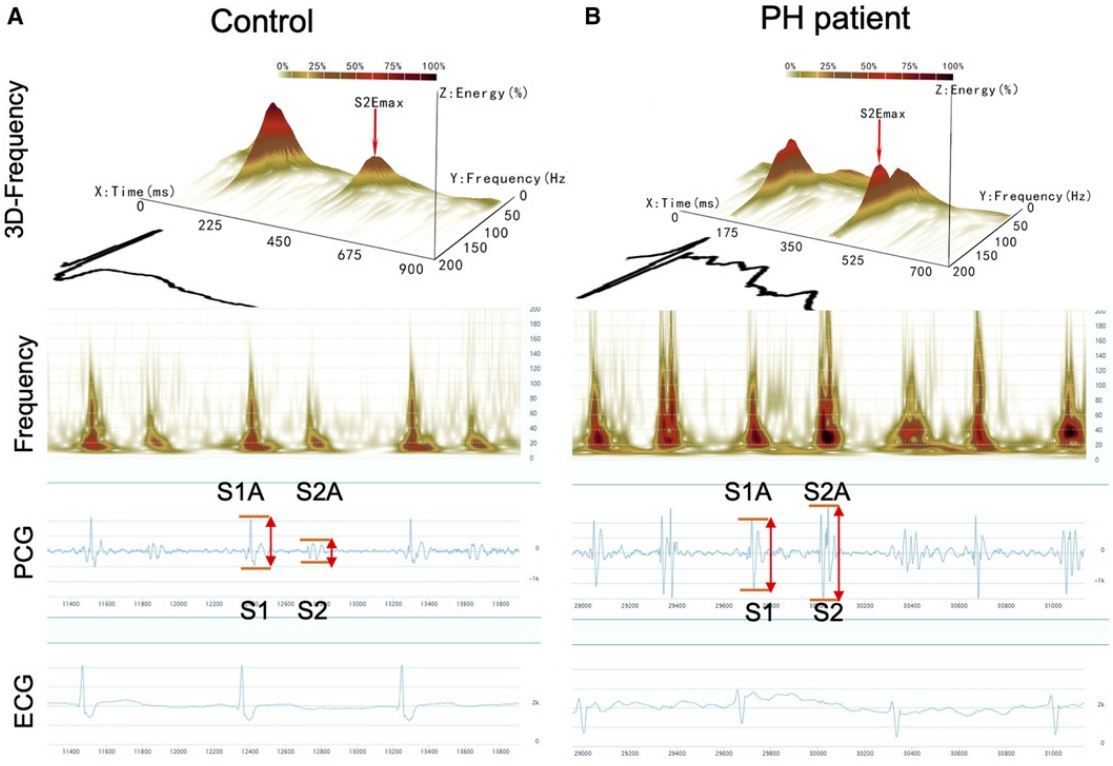


Figure 3.2: Acoustic cardiographic output for a control patient (A) and a pulmonary hypertension (PH) patient (B), showing frequency, phonocardiogram (PCG), and electrocardiogram (ECG). The scalogram time-frequency representation displays frequency on a logarithmic scale (0–200 Hz) and time on the horizontal axis. Wavelet transform energy is color-coded from light yellow to deep red. Key features include S1 (first heart sound), S2 (second heart sound), S1A (S1 amplitude), S1E (S1 energy), S1F (S1 frequency), S2A (S2 amplitude), S2E (S2 energy), and S2F (S2 frequency). Reproduced from [57].

Finally, some more recent research, such as that of Pengyue Ma *et al.*, combines traditional features with deep learning techniques. In their study, they use power-normalised cepstral coefficients (PNCC) to obtain a more robust recognition

feature, as shown in Figure 3.3 [47].

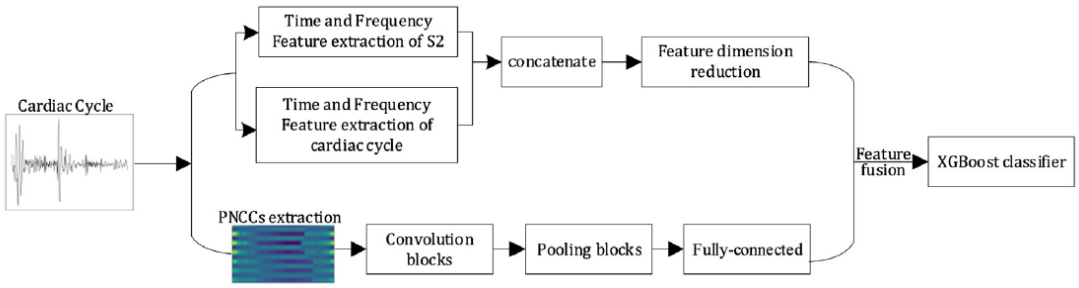


Figure 3.3: Hybrid feature extraction framework from [47], combining time-frequency features of heart sounds with PNCC-based deep learning

The quality of the features extracted from the data is also strongly correlated with the method implemented in the different studies, since each approach, either traditional or deep learning based, shows specific advantages and entails several limitations that affect the quality and interpretability of the selected features.

3.3 Method Adopted

Traditional approaches include statistical techniques or classical machine learning algorithms such as Linear Discriminant Analysis (LDA) [50, 51], linear regression [52, 29] and validation methods such as k-fold [49, 16, 13, 55] or leave-one out [50, 51] cross-validation.

One big advantage of the studies that implement these methods is their interpretability, as they make it possible to understand the physical meaning of the features identified as the most informative for the diagnosis of pulmonary hypertension.

This approach allows indeed more comprehensible results, as demonstrated in studies such as that of Elgendi *et al.* [51], who used measures of signal entropy and complexity to identify patterns in disease-related heart sounds achieving a sensibility ranging from 84% to 93%.

Moreover, traditional approaches are generally less computationally demanding than deep learning methods and also more suitable for smaller datasets. However, these techniques have significant limitations, especially when applied to complex data.

As a matter of fact, traditional approaches rely on manually extracted features, a process that often fails to fully capture the complexity of signals, especially when datasets are characterised by high variability. In these cases, the obtained models can't generalise effectively and the results are not particularly encouraging, like in Yamakawa's and Kaddoura's works [56][16].

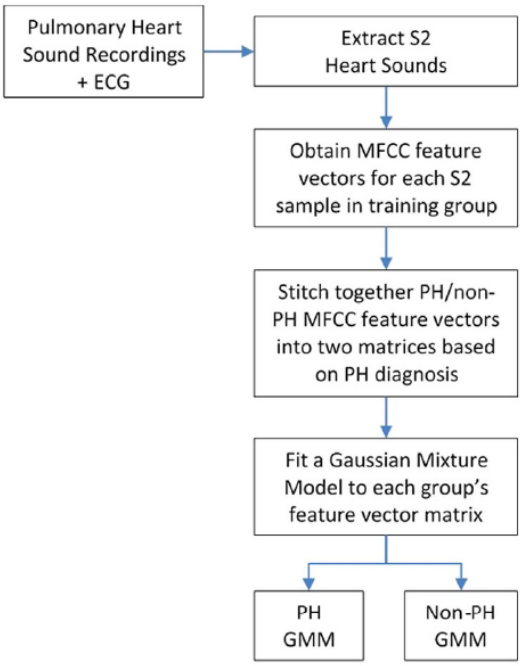


Figure 3.4: Flow diagram employed in [16] illustrating the training of the acoustic models through a classical Machine Learning Algorithm

Figure 3.4 shows the pipeline emplyed by Kaddoura *et Al.* [16] who trained their models using a classical machine learning algorithm based on Gaussian Mixture Models (GMM).

That's the reason why in many cases, more modern approaches such as deep learning networks are preferred. An important advantage of these models lies in the ability of neural networks to automatically learn relevant features from the data, thus avoiding manual intervention and enabling the model to identify eventual complex patterns. Wang *et al.* [**Wanf**], for example, demonstrated that the use of CNNs combined with a spectral representation of heart sounds (Figure 3.5) allows to effectively classify them in five heart diseases, resulting in an auROC value of 0.99.

However, the reliability of the result may be compromised by the fact that the training and validation sets include in some cases the same subjects, which could lead to overfitting. To address this issue, a solution could be employing cross-validation techniques and pre-trained models in studies with a limited amount of data.

This solution is implemented by Gaudio's study [13], which demonstrates with a model gaining an auROC of 95% how deep networks outperform traditional machine learning models.

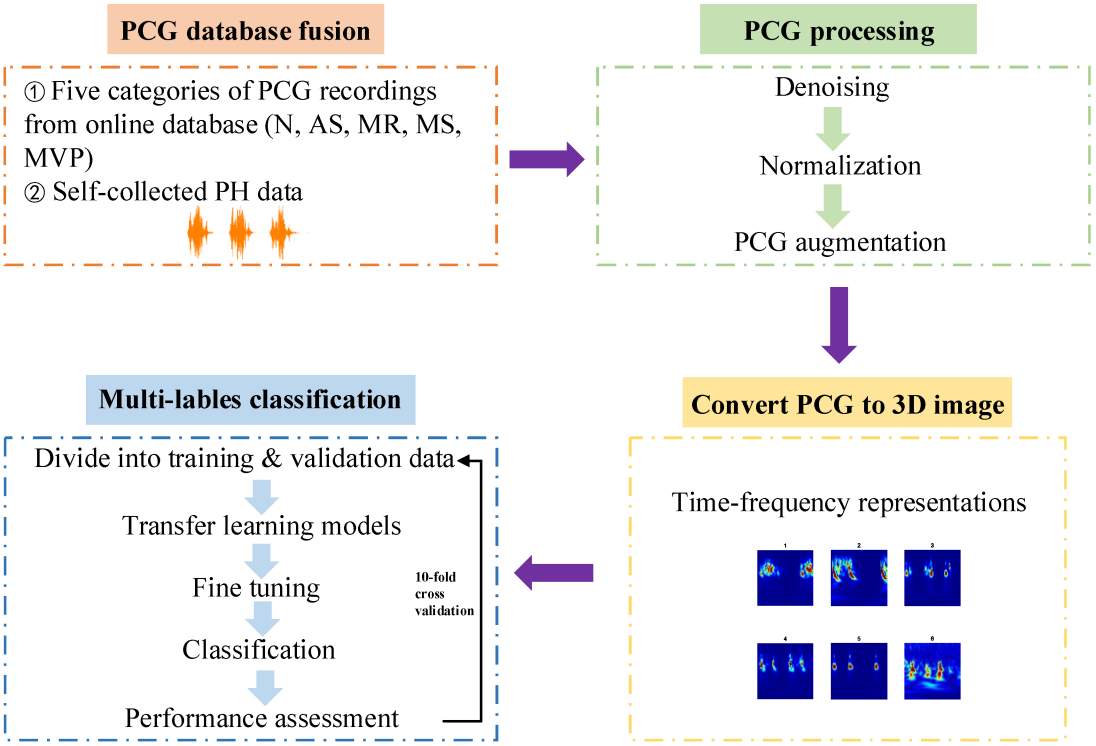


Figure 3.5: Flowchartof PCG classification employed by [49], using deep learning and trasnfer learning models.

3.4 Dataset

Analysing the results obtained, it additionally emerges that the choice of method combined with the dimension of the datasets has a direct impact on the quality of the predictions and the ability of the models to generalise.

Studies that used larger datasets, such as Chan MFCC [29] and Huang MFCC [57] , achieved promising outcomes, confirming that the availability of bigger datasets helps neural network to capture all the possible physiological patterns and trains it the most efficient way.

In contrast, studies based on smaller datasets like Elgendi's work [50], while achieving good results in terms of sensitivity and specificity, must be interpreted with caution due to the high risk of overfitting. To address this issue, recent studies have therefore employed rigorous validation techniques to show more reliable outcomes, such as stratified k-fold cross-validation and bootstrapping.

In conclusion, it can be stated that the choice of method and the size of the dataset play a decisive role in the performance of the model. While deep learning methods emerge as the most promising for large-scale analyses, traditional approaches remain valuable for smaller studies, providing interpretable features, results and limited computational resources. In any case, the reliablity of the results must be verified according to the validation technique implemented, as it can be a key factor in understanding whether the model's outcomes are affected

by overfitting.

Table 3.1: Comparison of Studies on Pulmonary Hypertension Detection

Authors	**Dataset**	**Reference**	**Signal's portion studied**	**Features Extracted**	**Method**	**Results**
Tranulis *et al.* (2007)[58]	N=9 pigs, 15-50 PCGs for each class (baseline, moderate, severe). Accessibility not mentioned.	RHC	S2	Handcrafted features: maximum instantaneous frequency of A2, that of P2, and the splitting interval between A2 and P2	ML: feedforward backpropagation neural network.	Linear regression between NN estimated systolic and mean PAP and cathetered systolic and mean PAP: r coefficient= 0.89 and 0.86 respectively.
Dennis *et al.* (2010) [48]	N=51 patients. Accessibility not mentioned.	RHC.	Both entire cardiac cycle and S2, A2 and P2.	Handcrafted: temporal and frequency related features.	Combination of ML algorithms for detecting the most performative subset of features, location on the chest wall used to record the sounds and ML algorithm	0.77 accuracy and 0.78 auROC
Elgendi *et al.* (2015)[50]	N=27 children's recordings. Accessibility not mentioned.	RHC.	All signal is used	Handcrafted features: relative power, energy and entropy of the heart sound's frequency bands.	Linear Discriminant Analysis (LDA) as a statistical technique to help distinguish the patients based on the heart sound entropy	Sens=93%, Spec=92% for the first sinusoid formant's entropy

Elgendi *et al.* (2018) [51]	N=60 patients (35 PH positive). Accessibility not mentioned.	RHC.	All signal is used	Handcrafted features: entropy of the formants (resonance frequencies) of the heart sound.	Linear Discriminant Analysis (LDA) as a statistical technique to help distinguish the patients based on the heart sound entropy.	Sens=84%, Spec=88.57% for the first sinusoid formant
Cherif *et al.* (2016) [52]	N=17 recordings Accessibility: private.	ECHO.	A2 and P2	Handcrafted: frequency-related features.	Linear regression on spectral parameters of the PCG + kNN	Not shown
Yamakawa *et al.* (2021) [56]	N=40 patients' recordings (18 PH positive). Accessibility: private.	RHC.	S1, S2, S3, and S4	Handcrafted: CABs: Intensity, complexity, and strength of the 4 heart sounds	Statistical analysis between RHC parameters and some cardiac acoustic features for 4 severity levels of PH.	AUC range: Prec-PH: 0.674 to 0.720, Ipc-PH: 0.646 to 0.807, Cpc-PH: 0.742
Chan *et al.* (2013) [29]	N=170 patients (40 PH positive). Accessibility: private.	RHC.	S1 and S2	Handcrafted: Intensity and complexity.	T-test and logistic regression between hemodynamic variables and PCG's features	AUC ROC=0.85 for S2 complexity. AUC ROC=0.89 for S2/S1 complexity ratio
Huang *et al.* (2023) [57]	N=209 patients (121 PH positive). Accessibility: public.	ECHO.	S1 and S2	Handcrafted: frequency-related features.	Statistical comparison between PH's severity (ECHO) and PCG's features	auROC: 0.775 for S2 frequency. Sens=79.34%, Spec=67.05%

Kaddoura *et al.* (2016) [16]	N=164 patients (86 PH positive). Accessibility not mentioned.	RHC.	S2	Handcrafted: mel-frequency cepstral coefficients.	ML Gaussian Mixture Model + Negative log-likelihood	auROC=0.74
Pengyue (2023) [47]	Not mentioned. Accessibility: private.	Not mentioned.	Both entire cardiac cycle and S2	Handcrafted: time-frequency from both the entire cardiac cycle and S2 + Deep features.	Hybrid CNN + XGBoost	Acc=88.61%
Wang *et al.* (2022) [49]	N=1102 patients' recordings (102 PH positive). Accessibility: public.	Not mentioned.	All signal is used	Deep features.	10 pretrained CNN models are employed to classify the converted spectrogram images into six categories.	auROC=0.99
Gaudio *et al.* (2022) [13]	N=42 patients (29 PH positive). Accessibility: private.	RHC.	S2	Deep features.	3 different CNN models initialized with various techniques + ML models to compare the results.	auROC=0.95
Gaudio *et al.* (2023) [55]	N=42 patients' recordings (29 PH positive) + N=110 porcine's recordings (all PH positive). Accessibility: private.	RHC.	S2	Deep features.	DenseNet121 randomly initialized.	auROC=0.92

Chapter 4

Materials and Methods

The following chapter describes the materials and methods used to build the proposed model.

In particular, the characteristics of the dataset, the pre-processing techniques applied to the phonocardiographic and electrocardiographic signals, and the experimental configurations adopted for the training and validation of the model are illustrated. Furthermore, the evaluation metrics employed to analyse the performance of the model in the task of classifying the presence of pulmonary hypertension are detailed.

4.1 Dataset

The dataset includes recordings from 178 individuals, with approximately 60% males and 40% females. The participants' ages range from 23 to 88 years.

To ensure a standardized and efficient data collection process, the Rijuven Cardiosleeve was employed (Figure 4.6). It is an advanced and non-invasive stethoscope capable of recording, analyzing, and displaying both ECG and heart sounds (PCG) signals simultaneously [59].

Figure 4.1: Rijuven Cardiosleeve multimodal stethoscope.

Each patient underwent a single session, during which auscultation was performed. Recordings were acquired at four auscultation sites: the aortic, pulmonary, tricuspid,

and mitral valves, with each session lasting approximately 30 seconds per site.

Patients identified by IDs from 1 to 108 have their PCG and ECG signals recorded in .mp3 and .raw format respectively, with a sampling rate of 8 kHz and 500 Hz. In contrast, data for patients from id 109 onward were acquired at 3 kHz concerning PCG collection, and 500 Hz as regarding ECG, facilitating high-quality multimodal signal acquisition.

The dataset contains also an Excel file that contains physiological and clinical data obtained from echocardiographic exam records for each patient, as well as information extracted from RHC examinations, when present. All patients were subjected to the echocardiogram examination, while only 23 of them underwent RHC.

The variable of interest extracted from the RHC exam and reported in the excel file is the mean pulmonary arterial pressure (mPAP), whose value together with a threshold are used to determine the presence of pulmonary hypertension.

Regarding the echocardiogram variables on the other hand, the relevant value is the maximum velocity of tricuspid regurgitation, on the basis of which doctors estimated a probability of pulmonary hypertension. This probability is described by four values going from 0 (low probability of PH) to 3 (high probability of PH).

Table 4.1 shows the minimum and maximum values of the tricuspid regurgitation velocity (TR) used by physicians and associated with each level of probability of pulmonary hypertension (PH).

PH Probability	Minimum velocity of TR (m/s)	Maximum velocity of TR (m/s)
0	0.00	0.86
1	0.87	2.83
2	2.87	3.43
3	3.50	4.58

Table 4.1: Threshold values used by physicians for the velocity of TR for PH probability estimation

4.1.1 Quality of the data

All signals, both ECG and PCG, acquired from the four heart valves were carefully visualized and organised in a .mat structure to ensure better readability and accessibility of the data.

To assess the overall quality of the dataset, a qualitative analysis of the signals was conducted. Each recording was rated on a scale of 1 to 4, based on the visibility of the main signal characteristics.

In particular for ECG signals, the classification took into account the clarity and definition of QRS complexes, T-waves and other relevant components.

For PCG signals instead, quality was assessed based on the visibility of the S1 and S2 components, which are crucial for the analysis of heart sounds.

Figures 4.2 and 4.3 provide examples of high- and low-quality signals.

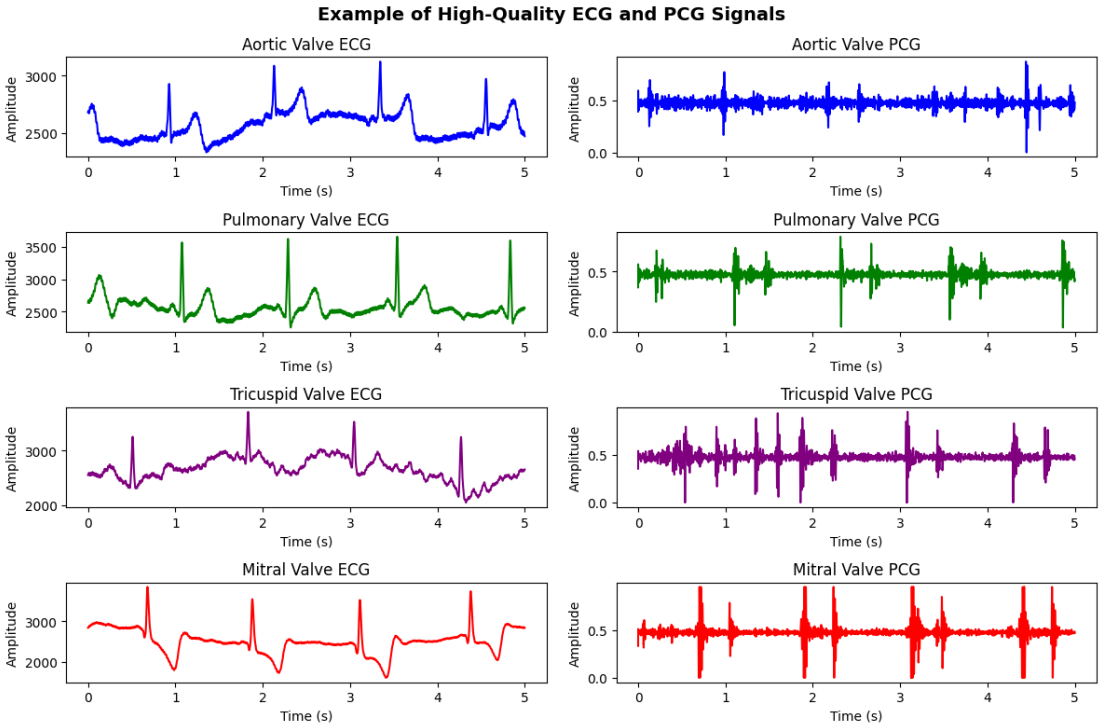


Figure 4.2: Good-quality illustrative recordings from the four heart valves: ECG signals on the left and PCG signals on the right, limited to the first 5 seconds.

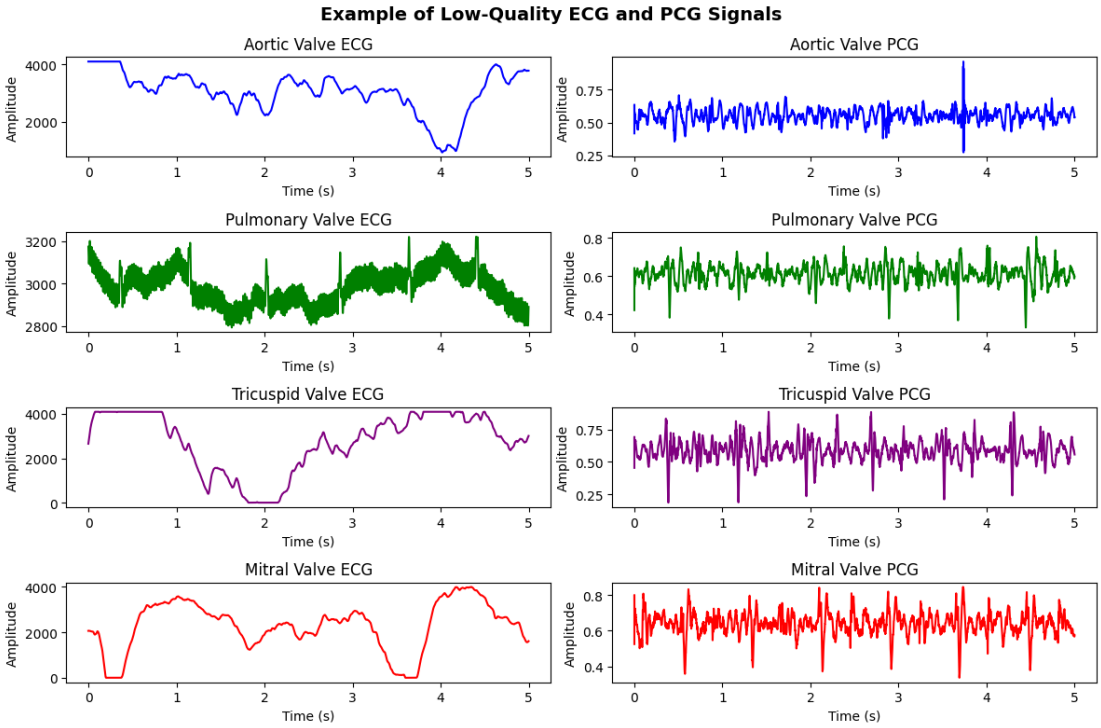


Figure 4.3: Low-quality illustrative recordings from the four heart valves: ECG signals on the left and PCG signals on the right, limited to the first 5 seconds.

This classification made it possible to identify and select the patients who carried highest quality signals, improving the reliability of subsequent analyses and optimizing

the use of the dataset for the next steps.

4.1.2 Selection Strategy and Experimental Configurations

Only the phonocardiographic signals from the pulmonary valve site were selected as useful ones for the development of a model capable of distinguishing patients with pulmonary hypertension from healthy ones.

Indeed, the auscultation of this area is a key site for the diagnosis and assessment of pulmonary hypertension, as the accentuation of the pulmonary component of the second heart tone (P2), one of the most distinctive signs of the condition, is particularly evident when listened from here [60].

Moreover, in order to conduct a comprehensive analysis and assess the impact of the configuration of the dataset on the model, experiments were conducted by successively changing the dataset used, as shown in Table 4.2.

Initially, the analysis was performed considering only the signals from the 23 patients for whom RHC (right heart catheterisation) data was available; this configuration allowed to evaluate the performance of the model against the more reliable ground truth for the diagnosis of pulmonary hypertension [48].

Subsequently, the analysis was extended to include the entire available dataset, using the echocardiography-based pulmonary hypertension severity classification (ECHO) as ground truth.

In this setup, the severity classes were binarised: probabilities of 0 and 1 were regarded as negative, whereas probabilities 2 and 3 were classified as positive. This approach allowed the model to be tested on a larger and more diversified dataset, even with a less precise ground truth source than right heart catheterisation.

Finally, a third configuration was explored in which the dataset was filtered to include only patients with echocardiographic data of better signal quality in order to improve the reliability of the analysis. This selection resulted in a dataset of 60 patients in total, 24 positive and 36 negative. This configuration allowed us to evaluate the performance of the model in a context where the echocardiographic signal quality is more homogeneous, minimising potential bias due to low quality data.

Configuration	Ground truth	Number of patients	Class distribution
1	RHC	23	17 positive and 6 negative
2	Echo	178	34 positive and 142 negative
3	Echo	60	34 positive and 26 negative

Table 4.2: Summary of the dataset configurations used in the study

4.2 Preprocessing

In the preprocessing phase, the first step was downsampling all the PCG signals to 1000 Hz, in order to make all signals homogeneous and reduce the computational load.

Then, a Butterworth bandpass filter of order 5 with a bandwidth between 20 and 200 Hz was applied, to reduce possible noise and non-informative components out of the useful band. The illustration in Fig.4.4 shows a comparison between a normal signal and a filtered one.

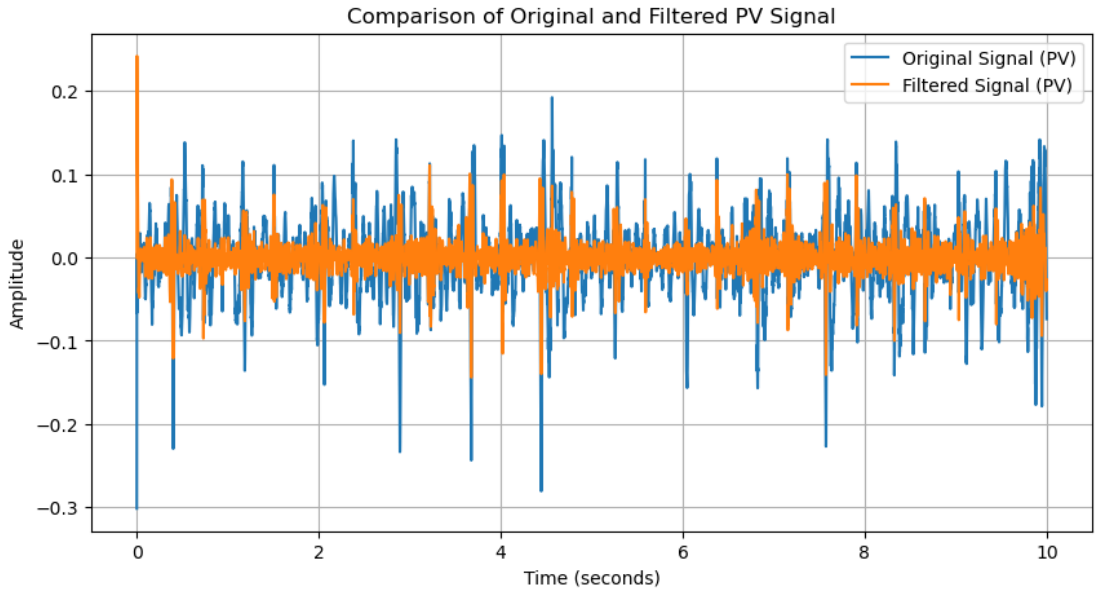


Figure 4.4: Comparison of the original and filtered 10-seconds PV signal over time

The filtered signals were then normalised by subtracting their mean and dividing by the standard deviation as shown in Equation 4.1, in order to equalise the amplitude variations and ensure a more comprehensible and homogeneous representation.

$$x_{\text{norm},i} = \frac{x_i - \mu}{\sigma} \tag{4.1}$$

At this point, the processed signals were used to compute the MFCC, a commonly used technique for representing audio signals described in Section 2.3.

These coefficients map the original signal in a non-linear Mel-Scale that mimics the listening process of the human ear.

For this purpose 13 MFCC coefficients were calculated, selecting a window length of 32 milliseconds and a hop length of 16 milliseconds to achieve an optimal balance between time and frequency resolution.

To capture the temporal variability of the signal as additional information, MFCC delta and delta-delta were also calculated, which are computed from the first and second derivative of the original preprocessed signal, respectively, as shown in Figure 4.5.

These derivatives contribute to give insights into the temporal evolution of the spectral features [61]. The combination of MFCCs, delta, and delta-delta coefficients forms a exhaustive feature set that completely represents both the static and dynamic characteristics of the PCG signals.

The three matrices obtained, corresponding to the MFCCs of the original signal, its first derivative (delta) and second derivative (delta-delta), were divided into 3-second

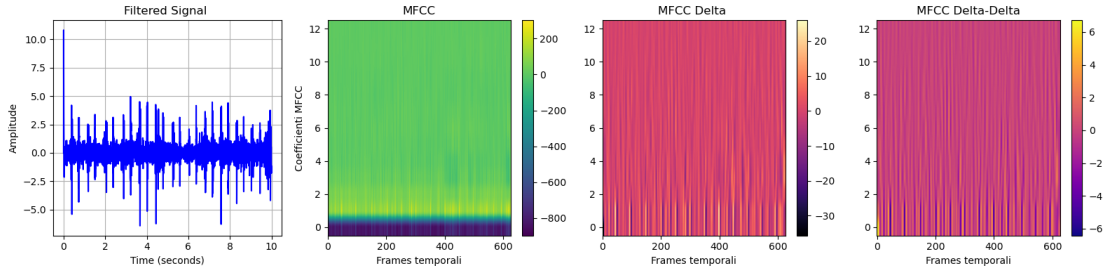


Figure 4.5: Filtered audio signal and MFCC features: time-domain signal, MFCC, MFCC Delta, and MFCC Delta-Delta for a 10-second example

time segments with a 50% overlap, thus ensuring continuous coverage of the information contained in the representation.

For each extracted segment, the three matrices were concatenated and normalized again, obtaining the final representation intended for the input of the neural network.

This type of time-frequency representation of the signal contributed to improve the model's ability to recognize patterns and differences between positive and negative patients.

4.3 Masking

Tests were also conducted using a dataset in which the masking technique was applied to the signals. This consists of selectively preserving certain portions of the original signal, masking the rest with null values.

In particular, masking was applied on the sections corresponding to the S2 components of the phonocardiographic signal, in order to isolate and analyse the effects of these segments alone on classification.

For each patient, S2 intervals were extracted using the technique employed in [62] and were extended by 10 samples before and after to include any relevant transitions. The rest of the signal was zeroed, resulting in a masked version of the original signal, as shown in Figure 4.6.

This masking technique evaluates the model's ability to identify discriminating features based only on segments of the signal that might be most informative, reducing the influence of other components that might introduce noise into the analysis.

4.4 Modelling

For the analysis and classification of phonocardiographic signals, a deep learning model based on a CNN-type convolutional neural network was implemented. Its architecture and characteristics are illustrated in Tables 4.3 and 4.4.

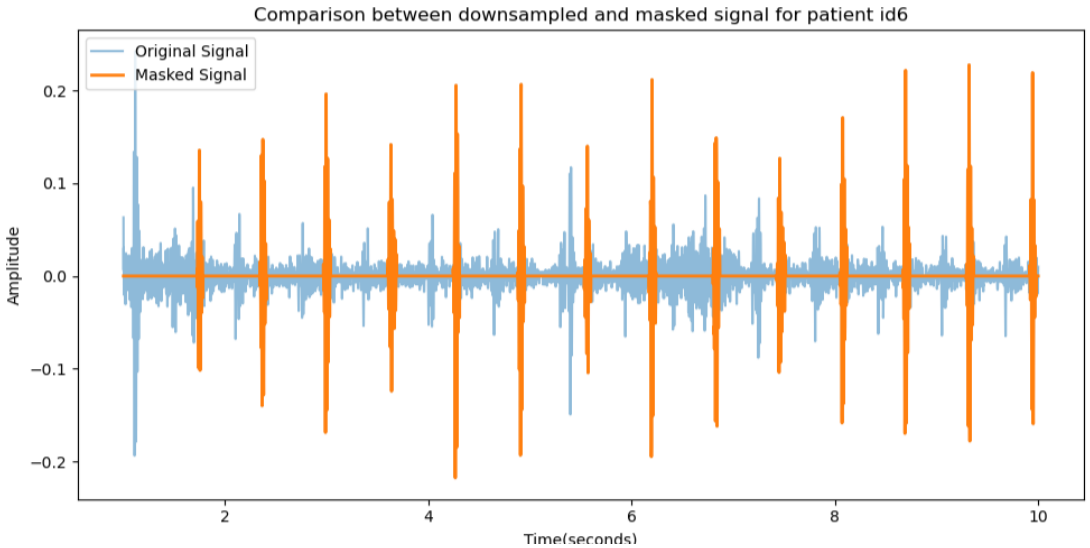


Figure 4.6: Comparison of downsampled and masked phonocardiographic signals for a patient

Parameters	Value
Input Dimension	(39, 187, 1)
Batch Size	32
Epochs	150
Learning Rate	10^{-4}
Optimizer	Adam
Loss Function	Binary Cross Entropy
Activation Function	ReLu

Table 4.3: Characteristics of the Deep Learning Model used

The input of the model has a dimension of (39, 187, 1), where 39 is given by the concatenation of the three matrices of 13 MFCC coefficients and 187 represents the length proportional to the duration of the signal.

The model architecture consists of five convolutional layers with kernels of varying size, each followed by Batch Normalization, MaxPooling and Spatial Dropout layers, with a progressively increasing dropout rate up to 50%, to prevent overfitting.

The activation function used for the convolutional layers is the ReLU (Rectified Linear Unit), which allows effective gradient propagation during training.

After convolutional feature extraction, the model uses three dense layers, each followed by a dropout of 50%, to improve the generalisation capability of the model. Finally, a dense layer with a single neuron and sigmoid activation function is used for binary classification.

For the optimisation of the model, the Adam algorithm was adopted, with a learning rate of 10^{-4}.

The loss function used is the binary crossentropy and an Early Stopping mechanism was implemented to avoid overtraining of the model.

Layer	Dimensions
Convolutional Layer	(32,(5x5))
Convolutional Layer	(64,(3x3))
Convolutional Layer	(128,(3x3))
Convolutional Layer	(256,(3x3))
Convolutional Layer	(512,(2x2))
Dense Layer	512
Dense Layer	256
Dense Layer	128

Table 4.4: Network Layers

In cases of imbalance between classes in the dataset, a weight balance was applied during training, assigning a higher weight to the minority class to avoid the model favouring the most represented class.

The architecture of the model used for this work is shown in figure 4.7.

4.5 Classification Strategies and Evaluation Metrics

In the present work, the k-fold cross-validation method with k=5 was adopted for the evaluation of classification performance in the deep learning model: the dataset was divided into five subgroups of equivalent size, balancing the distribution between positive and negative classes.

At each iteration of the process, three of these groups were used for the training set, another of these was used as the validation set, and the last as the test set.

The division into training, validation and test set was designed in such a way that each subgroup was used as a test set at least once. This approach ensures that, at the end of the cross-validation procedure, all patients were included in the test set, thus allowing an evaluation of the model on completely new data during each iteration.

In order to evaluate the performance of the different trials performed, different evaluation metrics were employed.

The confusion matrix is a table used to evaluate the performance of a classification model. It shows the number of correct and incorrect predictions divided into four categories: TP (True Positives) and TN (True Negatives) refer to correct prediction instances, while FP (False Positives) and FN (False Negatives) represent incorrect predictions.
The confusion matrix has been used to evaluate the model's performance both at the level of individual segment classification and at the patient level, where the assessment holds greater diagnostic significance.

The accuracy measures the proportion of correct predictions in relation to the total

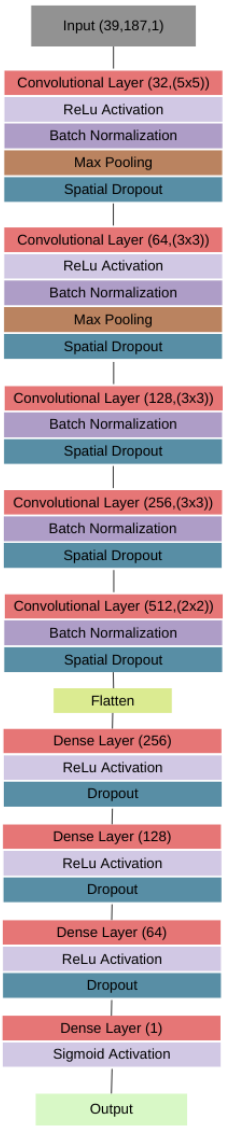


Figure 4.7: Architecture of the Convolutional Neural Network (CNN) for feature extraction and classification.

number of observations (4.2). It represents an overall measure of model performance, but can be misleading in the case of unbalanced classes.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

Sensitivity, also called recall or True Positive Rate (TPR), measures the model's ability to correctly identify positive cases (4.3). A high value indicates that the model detects well positives cases, which can be a crucial demand in medical contexts, but it is a metric that does not take into account the number of false positives.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4.3}$$

Specificity on the other hand, measures the model's ability to correctly identify negative cases (4.4). It puts weight on the number of false positive but in cases of imbalanced datasets, it may not be sufficient without considering performance on the minority class.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4.4}$$

Precision measures the proportion of samples classified as positive that are actually positive (4.5). For example, in a medical test, a high value of precision indicates that if the model says that the disease is present, the probability that the patient is actually ill is high.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.5}$$

A useful evaluation metric when the classes are inbalanced is the F1-score, which is the armonic mean between precision and sensitivity (4.6).

$$F_1 = 2 \times \frac{\text{precision} \times \text{sensitivity}}{\text{precision} + \text{sensitivity}} \tag{4.6}$$

An important tool for evaluating the performance of the model as the threshold value changes is the ROC curve: this is a graph showing the relationship between the True Positive Rate (Sensitivity) and the False Positive Rate (1 - Specificity) for threshold values that go from 0 to 1. A perfect model will have a curve approaching the upper left vertex of the graph.

Finally, the AUROC measures the area under the ROC curve and provides a numerical value to assess the model's ability to distinguish between classes.

4.6 PostProcessing

In the context of performance evaluation of the classification model, predictions are generated at the segment level. However, to obtain a clinically meaningful evaluation, it is necessary to translate these predictions to the patient level. This process made it possible to aggregate the outputs per segment into a single output for each patient. This chapter will illustrate three different post-processing approaches and describe their principles.

4.6.1 Probability Average with Fixed Threshold

The first method used to aggregate predictions at the patient level concerns calculating the mean probability with a fixed threshold of 0.5.

In this approach, the classification probabilities associated with its segments are collected for each patient and the average probability is calculated. If the average probability exceeds the threshold of 0.5, the patient is classified as positive, otherwise as negative.

4.6.2 Optimized Threshold Based on F1-Score

A second, more advanced approach involves searching for an optimal threshold for classification, instead of using the fixed value of 0.5.

In this strategy, a process is performed in which a series of thresholds are systematically tested, ranging from the minimum to the maximum value of the average probability of the patients within the validation data of each fold.

For each tested threshold, the F1-score is computed to evaluate classification performance, and the threshold that achieves the highest F1-score is selected as the optimal one.

Once the best threshold is determined during the validation phase, it is subsequently applied to the test data of the corresponding fold to classify patients.

This approach ensures that the classification decision is based on an empirically optimized threshold, verifying the model's ability to generalize effectively to unseen data.

4.6.3 Segment-Based Probability Aggregation with Optimal Threshold

The third method adopted is based on counting positive segments within each patient. Instead of calculating the average probability of the segments, the segment-by-segment probabilities are binarised using the fixed threshold of 0.5. Then, for each patient, the fraction of segments classified as positive out of the total number of segments belonging to that patient is calculated.

This value represents the final probability for each subject and it is used to find an optimal threshold through the same strategy as described in the previous method, i.e. by selecting the threshold that maximises the F1-score on the validation data.

The optimal threshold thus obtained is then applied to the test data for final classification.

Once the patient-level predictions have been obtained with each of the three described methods, the performances of the model are evaluated by calculating the standard metrics listed in 4.5.

Chapter 5

Results and Discussion

The purpose of this section is to show and evaluate the impact of the different configuration of the dataset on the performance of the classification model.

Indeed, as the available dataset was organised into three distinct configurations as reported in section 4.1.2, each characterised by a different level of ground truth reliability and a different distribution of patients, the performance analysis present in this chapter provide insight into how these aspects affect the model's ability to distinguish patients with pulmonary hypertension from healthy subjects.

From a methodological point of view, the model is evaluated through a 5-fold cross-fold validation method, as discussed in Section 4.5, to analyze the stability of performances.

However, as the classification is initially performed at the phonocardiographic segment level, an aggregation process of the performances was necessary to obtain a diagnosis at the patient level, according to the methods reported in Section 4.6.

The analysis of the results is structured by initially presenting the trend of the accuracy and loss curves in the training and validation for each configuration.

Next, the confusion matrices and evaluation metrics used for classification (Section 4.5) at segment level and at patient level are reported.

Finally, a comparison is made between the three configurations to identify the most reliable for diagnostic application.

5.1 Model Performance with the RHC Dataset

The first results presented are those obtained using the RHC dataset, which represents the configuration that carries the most reliable ground truth for the diagnosis of pulmonary hypertension.

As a matter of fact, this dataset is mainly composed of positive patients (16 positive and 7 negative) since right cardiac catheterization is a procedure that is typically performed on patients with suspected advanced cardiopulmonary dysfunction, thus at higher risk of pulmonary hypertension, as shown in Figure 1.1 [63].

In order to monitor the stability of the net's training and detect any overfitting or underfitting of the model, the accuracy and loss curves were studied for each of the five

folds of the cross-fold validation (Figure 5.1).

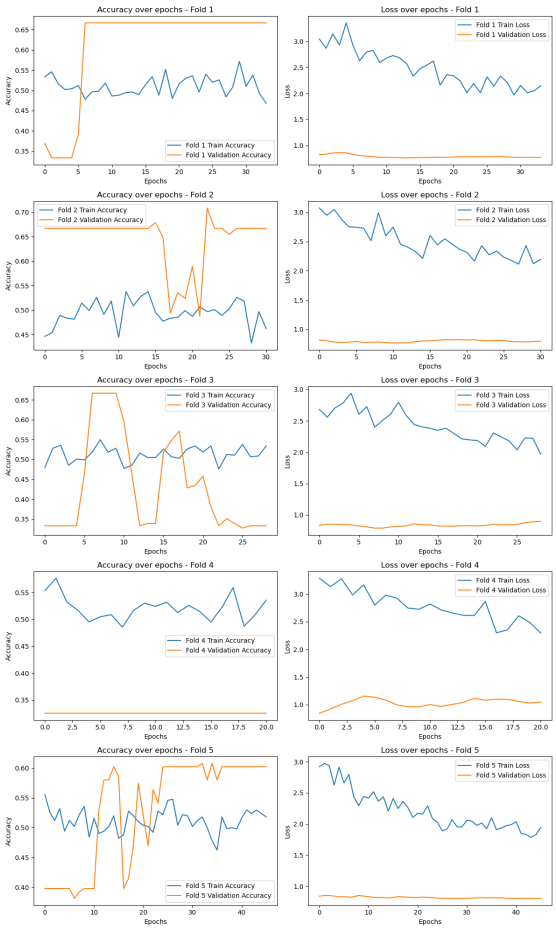


Figure 5.1: Accuracy and Loss Trends for the model trained on the RHC dataset

Next, the model's ability to discriminate between positive and negative patients at the phonocardiographic segment level was analysed.

Figure 5.2 reports the aggregated confusion matrix on segment-level predictions.

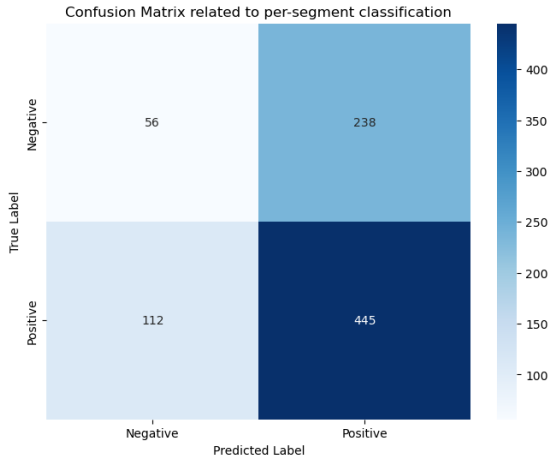


Figure 5.2: Confusion matrix for per-segment classification

Analysing the trends in loss and accuracy for the different folds, it was generally possible to highlight overfitting and a great instability in the validation performances.

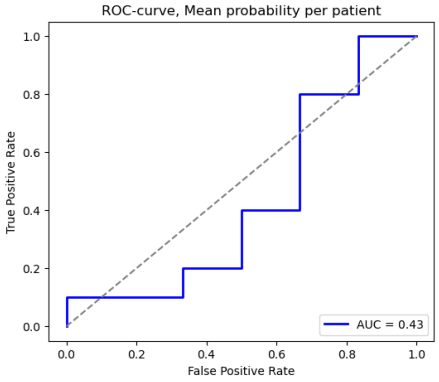
Continuous oscillation within the validation accuracy trend emphasizes a poor ability of the model to generalise, and reveal a first algorithm that is not able to find a stable pattern within the data.

The small size of the dataset and the strong imbalance between the two classes lead the model to favour the dominant one, as it is shown in Figure 5.2.

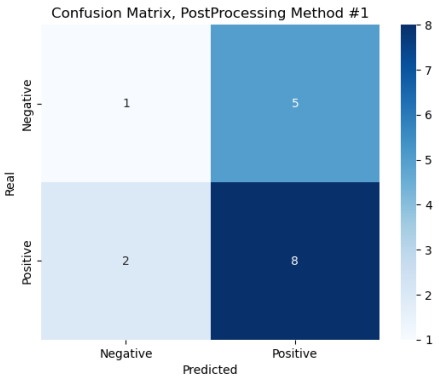
Balancing strategies were also tested, including assigning weights to classes to favor the least represented one and using focal loss to give greater emphasis to the samples that are most difficult to classify, but without significant improvements in performance.

To obtain clinically relevant performance, aggregated results per patient are also shown: Figure 5.3 shows the general ROC curve related to the classification of patients comparing their mean probability with different thresholds and the confusion matrices for the three aggregation strategies.

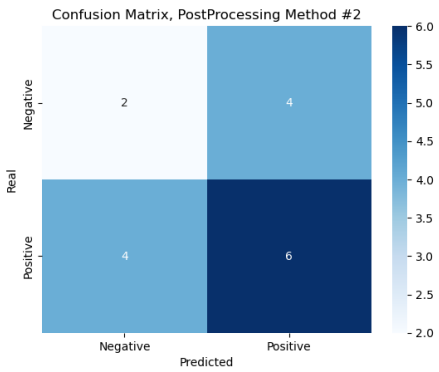
In addition, Table 5.1 shows the accuracy, F1-score, sensitivity and specificity values for each method.



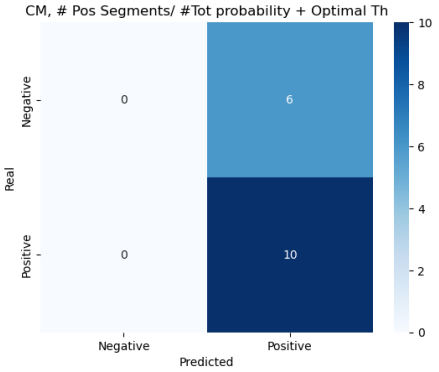
(a) ROC curve obtained using mean probability per-patient



(b) Confusion Matrix using 4.6.1



(c) Confusion Matrix using 4.6.2



(d) Confusion Matrix using 4.6.3

Figure 5.3: Per-patient performances using RHC dataset configuration

The analysis of the different post-processing strategies for aggregating segmental predictions shows how the choice of method influences the performance of the model at

the patient level. Through these techniques, an attempt was made to mitigate the effect of bias towards one of the two classes by using different classification methods, such as using non-fixed thresholds, or using the fraction of positive segments as the probability value per patient.

The method based on the average probability with a fixed threshold of 0.5, although the simplest, shows some obvious limitations, in particular a tendency to generate false positives, which indicates that this post-processing technique follows the trend and the performances of the model in general, and does not bring any room for improvement in classification.

This indicates that a rigid threshold is not necessarily the best choice, as it does not take into account the fact that, due to class imbalance within the dataset, the average probability values may be skewed toward higher or lower values depending on which class is more represented (in this case the positive one).

As a consequence, the imbalance results in a higher number of false positives compared to false negatives. However, it can be stated that, given the initial model performances, the algorithm in general struggles to effectively learn from the available data, and this also affects the performances per patient.

The next approach, which involves optimising the threshold based on the F1-score and should allow the decision criterion to be dynamically adapted according to the characteristics of the dataset, does not show a significant improvement in performance. Although this method increases the number of negative predictions, thereby increasing specificity as shown in 5.1, it still proves to be a rather random classifying approach.

In fact, using the optimal thresholds of each validation set on the test set, the model continues to show significant variability in performance, suggesting that the threshold optimisation strategy is not sufficient to correct the inherent instability of the model.

Finally, the method based on positive segment count with optimal threshold shows even more extreme behaviour, classifying all patients as positive. Thus, although with this last method an attempt was made to mitigate the effect of the imbalance between classes in situations where the average probability per patient can be skewed towards higher values for truly negative patients, the performances show how in practice the model completely loses the ability to distinguish between the two classes and even ends up amplifying the problem.

Per Patient PostProcessing	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F1-Score (%)**
Method 1	56.25	80.00	16.67	69.57
Method 2	50.00	60.00	33.33	60.00
Method 3	62.50	100.00	0.00	76.92

Table 5.1: Comparison of Performances Obtained using the RHC Dataset

5.2 ECHO Dataset

In the initial setup, in which the analysis was only conducted on the few patients for whom RHC data was available, the results showed a strong bias towards one of the two classes, due to the limited number of patients and the nature of the dataset itself.

The second step involved extending the analysis to all patients for whom echocardiographic (ECHO) data was available, bringing the total to 178 patients, of whom 34 were positive and 142 negative.

Although less accurate than catheterisation, echocardiography is currently the most widely used method for screening pulmonary hypertension and is widely adopted in clinical practice for an initial assessment of the condition [48].

5.2.1 Correlation between the two Ground Truths

In order to demonstrate the correlation between the two methods used as ground truth, in the specific case of the patients and the available data, an illustrative graph of the correlation between mPAP values measured by RHC and the corresponding severity classes obtained by ECHO was made for patients with both examinations available (5.4).

The graph also highlights via a horizontal line the mPAP value used as a threshold for diagnosing pulmonary hypertension.

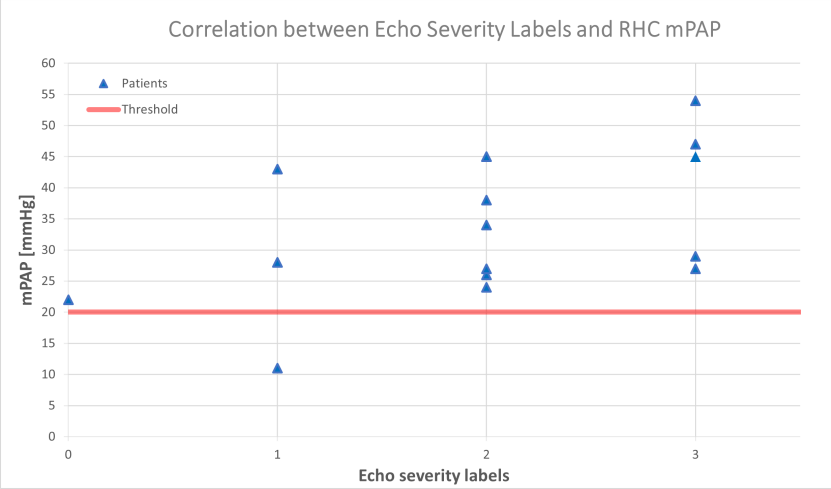


Figure 5.4: Correlation between mPAP values obtained via RHC and ECHO severity labels obtained via ECHO; a line marking the mPAP threshold value used to discriminate PH patients is also shown

Although the examples shown are limited to the 23 patients with catheterism data and are therefore mostly positive patients, a correlation between the two ground truth sources can generally be observed.

However, the variability through mPAP values within each ECHO class indicates some discrepancies between the two diagnostic methodologies.

Since the aim is to split the patients into two classes (positive and negative) and use this as ground truth for the model, it can be deduced from this graph that the ECHO examination can also be used as an indicator of the presence of pulmonary hypertension (PH).

As a matter of fact, although there is some variability in mPAP values within each severity class assigned by the ECHO, there is a general trend that follows that of ground truth verified by cardiac catheterisation.

In this regard, in the entire set of 178 patients, those with severity labels 0 and 1 were assigned to the negative class, while those with labels 2 and 3 were assigned to the

positive class.

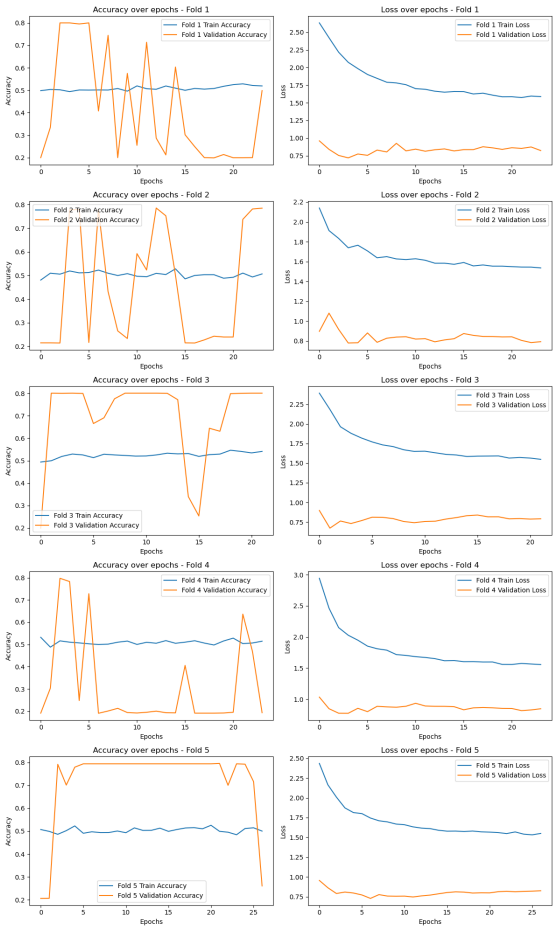


Figure 5.5: Accuracy and Loss Trends for the model trained on the full ECHO dataset

Similarly, using this second dataset configuration, accuracy and loss curves for each fold of the cross-fold validation were inspected to evaluate the stability of the trainings: the results are shown in figure 5.5.

Validation accuracy shows sudden fluctuations between consecutive epochs, suggesting a poor capability of the model to detect stable patterns in the data.

The training loss, on the other hand, moderately decreases, while the validation loss stays pretty much constant, indicating overfitting.

Therefore, the model seems to learn from the training data, but struggles to transfer this knowledge to the validation data, confirming that the use of the full dataset did not solve the instability problems observed in the RHC dataset.

Again, the aggregated confusion matrix on segment-level predictions (Fig. 5.6) shows the extreme bias towards the most represented class, signaling the low sensitivity in detecting patients with pulmonary hypertension.

Since the clinical evaluation of the classification is most relevant at the patient level, the three methods of aggregating the predictions were applied also in this case to obtain an overall diagnosis for each patient.

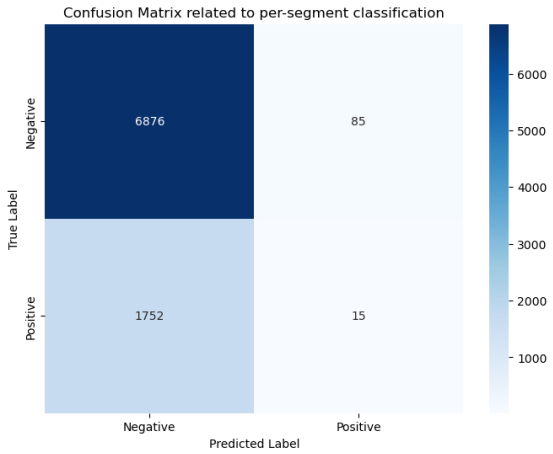


Figure 5.6: Confusion matrix for per-segment classification

The ROC curve obtained using the average probability per patient and the confusion matrices for the three aggregation methods (Fig. 5.11) show that performance remains random and unsatisfactory.

Indeed, the first method with a fixed threshold of 0.5, reports the performance of a model that classifies all patients as negative, following the performance trend by segment with no room for improvement: this could be due to a distribution of the average probability of patients tending towards values less than 0.5 even for positive patients.

An attempt was made to mitigate this problem by using the optimized threshold method based on F1-score. Indeed, in this case the sensitivity shows an improvement, proving that the model manages to classify, although erroneously, some patients as positive. The lack of ability to generalise from patients in the validation set to those in the test set is evidence that also the quality of the signals strongly influences the performance of the model.

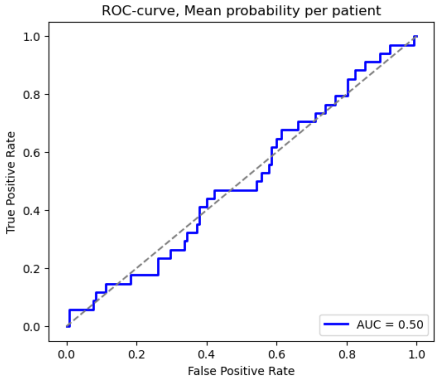
The analysis of the last post-processing method, based on counting the fraction of positive segments per patient, led to the classification of all patients as positive. This behaviour, which is anomalous compared to previous results, confirms that the probability distribution of the model is strongly skewed towards very low values.

As a matter of fact, in this approach, the final value assigned to each patient is determined by the proportion of segments classified as positive in relation to the total. However, in the specific case of this setup, the fraction of positive segments is extremely small for most patients. Consequently, during the threshold optimization process to maximize the F1 score in the validation data, the algorithm identified very low threshold values, almost close to zero, as the most effective for improving performance.

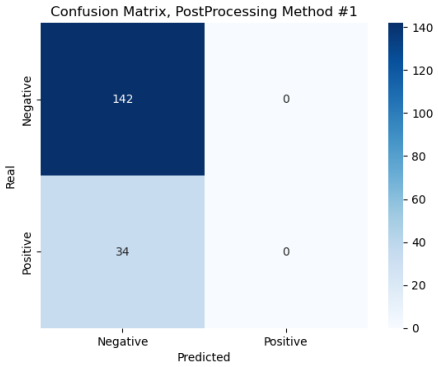
When these thresholds are applied to the test set, the result is that any patient is automatically classified as positive, leading the model to identify all patients as having pulmonary hypertension. This behaviour shows that the algorithm is not really learning a separation between classes, but is rather adapting its decisions to an unbalanced probability distribution.

In this context, threshold optimization is not sufficient to correct the problem of imbalance between classes and ends up amplifying it, confirming that the aggregation method based on the fraction of positive segments may not be suitable for this scenario.

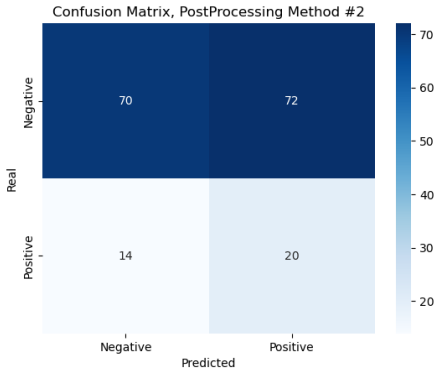
All these results suggest that, in order to increase the robustness of the classification,



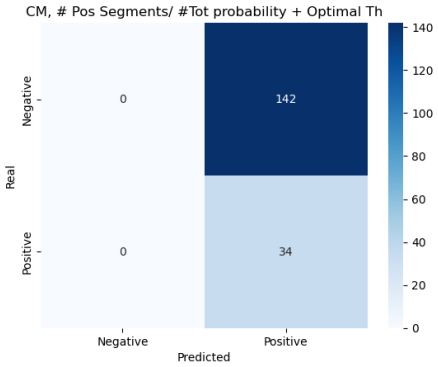
(a) ROC curve obtained using mean probability per-patient



(b) Confusion Matrix using 4.6.1



(c) Confusion Matrix using 4.6.2



(d) Confusion Matrix using 4.6.3

Figure 5.7: Per-patient performances using full ECHO dataset configuration

Per Patient PostProcessing	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
Method 1	80.68	0.00	100.00	0.00
Method 2	51.14	58.82	49.30	31.75
Method 3	19.32	100.00	0.00	32.38

Table 5.2: Comparison of Performances Obtained using the full ECHO Dataset

it would be crucial to intervene on the configuration of the dataset, ensuring a balance between the number of positive and negative patients and giving more importance to quality. This would create more favorable conditions for learning the model, reducing the impact of imbalance and improving its ability to generalize.

5.3 Dataset with Higher-Quality Signals

Following the analysis of the model's performance on the RHC and ECHO datasets, characterised respectively by a more reliable ground truth but with a smaller number of patients, and a larger dataset but with great imbalance, a third configuration was explored.

In this configuration, the dataset was filtered to include only patients with higher quality asclultation data in order to improve the reliability of the predictions and reduce noise in the data. This selection resulted in a dataset containing 60 patients in total, of

which 24 were positive and 36 negative.

The objective of this set-up was to investigate whether a more homogeneous dataset in terms of signal quality and balance between the two classes can improve the performance of the model enhancing its reliability.

The analysis of the accuracy and loss curves (Fig. 5.8) for each fold shows an improvement in stability compared to the configuration with the complete ECHO dataset. The training curves show a progressive increase in accuracy and a progressive reduction in loss, suggesting that the model is learning more effectively. However, the validation curves still show fluctuations, suggesting that the model may suffer from overfitting on some folds.

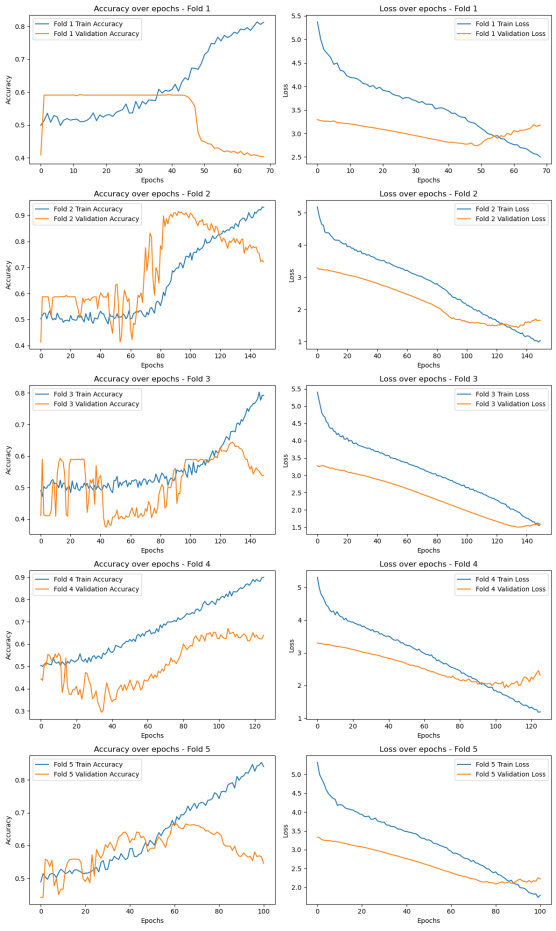


Figure 5.8: Accuracy and Loss Trends for the model trained on the high-quality dataset

The confusion matrix analysis for segment-level classification presented in Fig. 5.9 shows a reduction in errors compared to the previous configurations. However, the model continues to generate a significant number of false negatives and false positives, which can possibly impact on overall performance in classification at patient level.

However, both representations suggest that the model has a higher learning capacity if the dataset is composed of a comparable number of positive and negative patients, but above all good quality signals. This indeed facilitates the process of extraction of intrinsic and more informative features by the network and thus improves the ability to distinguish between segments of patients with pulmonary hypertension and healthy.

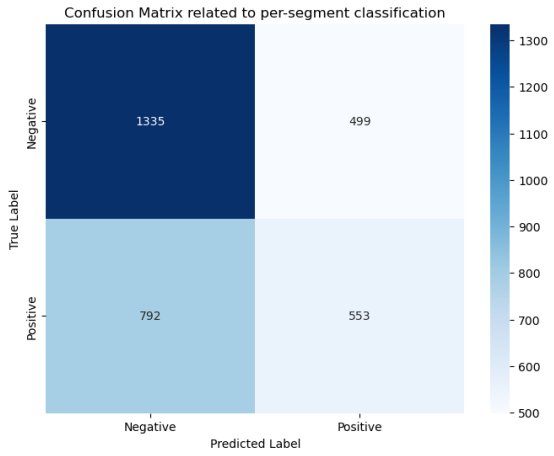


Figure 5.9: Confusion matrix for per-segment classification

The confounding matrices for the three aggregation methods shown in figure 5.10, confirm an improvement in performance compared to previous configurations, but still show margins of error, especially in the classification of positive patients.

The first method of aggregation, based on the mean of the probabilities with fixed threshold, obtained an accuracy of 73.33%, with a sensitivity of 84.62% and a specificity of 64.71%, proving to be the best for this configuration. The confusion matrix shows that the model correctly identifies most positive and negative patients, although some false positives persist.

The second method, with optimized threshold based on F1-score, increased sensitivity up to 92.31%, as the number of false negatives decreased but compromised specificity, which dropped to 32.35%, with a higher number of false positives. This behavior suggests that this second method of post-processing based on the best threshold according to the F1 score, tends to misclassify negative patients, therefore prioritizing sensitivity while compromising specificity.

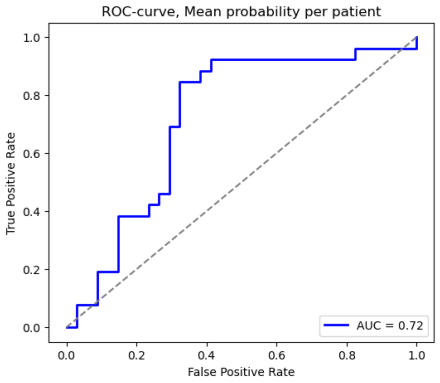
Similarly, the third method suffers from the same problem, in fact the number of patients correctly classified as positive almost equals the number of false positives.

In summary, this configuration improved the model's ability to discriminate between positive and negative patients compared to previous models. However, specificity still needs to be improved as the model continues to favour the positive class.

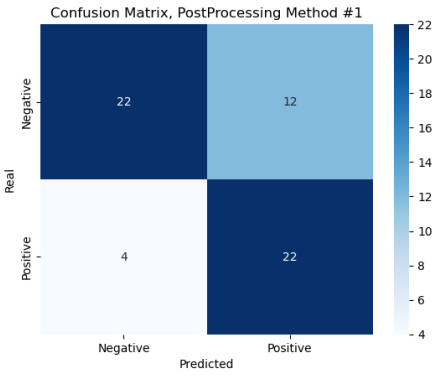
This behaviour may be considered acceptable in clinical contexts where the priority is the detection of true positives, even at the cost of reduced specificity.

In other words, priority is given to the model's ability to correctly detect patients with the condition of interest, accepting a certain margin of false positives, i.e. misclassification of healthy patients as sick.

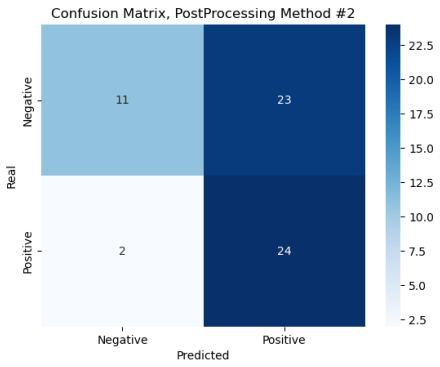
This approach is justified by the fact that, from a diagnostic point of view, it may be preferable to overestimate the presence of the condition rather than underestimate it, thus avoiding the risk of failing to recognise a patient who is actually sick and excluding him or her from possible early treatment.



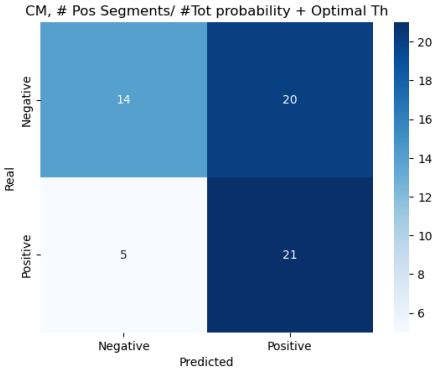
(a) ROC curve obtained using mean probability per-patient



(b) Confusion Matrix using 4.6.1



(c) Confusion Matrix using 4.6.2



(d) Confusion Matrix using 4.6.3

Figure 5.10: Per-patient performances using high-quality ECHO dataset configuration

Per Patient PostProcessing	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
Method 1	73.33	84.62	64.71	73.33
Method 2	58.33	92.31	32.35	65.75
Method 3	58.33	80.77	41.18	62.69

Table 5.3: Comparison of Performances Obtained using the high-quality ECHO Dataset

5.4 Masking technique

After identifying a dataset configuration that showed promising performance in the context of classification, it was decided to examine in depth the analysis by applying the same model to the same high-quality dataset modified through the masking technique described in Section 4.3.

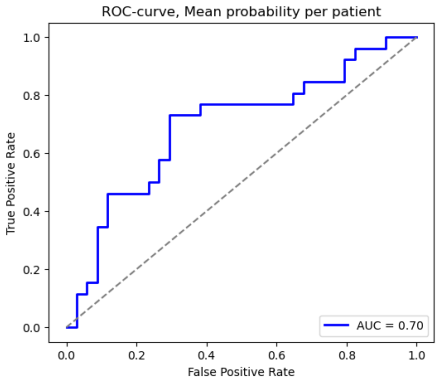
The objective of this trial is to verify whether the model is able to learn and generalise better when the signal only includes the components that should be more informative, thus reducing the influence of potentially irrelevant or noisy portions.

This segmentation technique was feasible and reliable since it was known that the signals from the selected patients were of high quality; contrarily, since the masking technique requires precise annotations to accurately isolate the S2 components, it would have resulted as ineffective, leading to the selection of signal portions that were not truly informative.

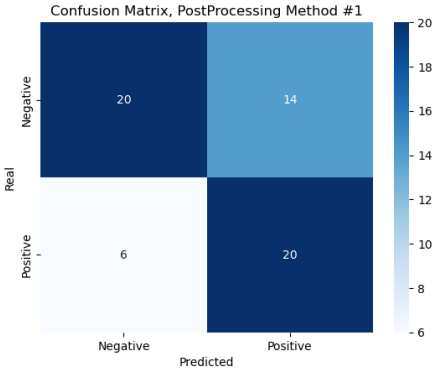
In the following paragraphs, the results obtained with the masked dataset are shown and compared to those of the last configuration (Section 5.3).

According to the results shown in Table 5.4, masking has in most of the cases worsened model performance, leading to a decrease in accuracy and specificity in almost all three post-processing methods.

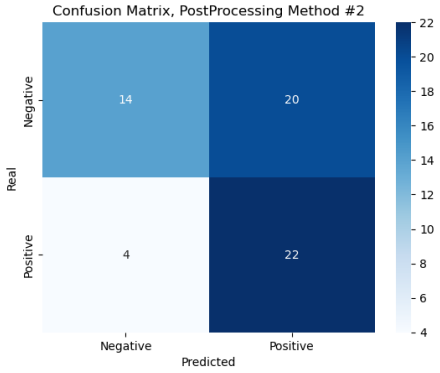
In particular the application of the technique has worsened the ability to distinguish negative patients (reduction in specificity). This suggests that although the S2 component of the phonocardiographic signal is considered a key element in the diagnosis of pulmonary hypertension, the rest of the signal may contain useful information that the model loses when masking is applied.



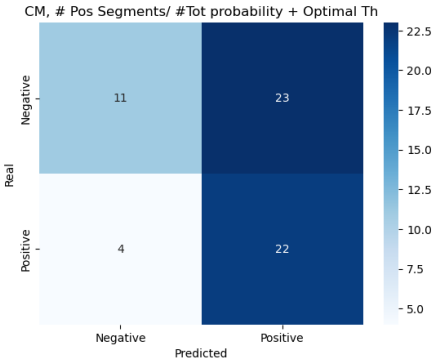
(a) ROC curve obtained using mean probability per-patient



(b) Confusion Matrix using 4.6.1



(c) Confusion Matrix using 4.6.2



(d) Confusion Matrix using 4.6.3

Figure 5.11: Per-patient performances using high-quality ECHO dataset configuration and Masking technique

Per Patient PostProcessing	Without Masking				With Masking			
	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
Method 1	73.33	84.62	64.71	73.33	66.67	76.92	58.82	66.67
Method 2	58.33	92.31	32.35	65.75	60.00	84.62	41.18	64.71
Method 3	58.33	80.77	41.18	62.69	55.00	84.62	32.35	61.97

Table 5.4: Comparison of evaluation metrics for different methods with and without masking. The higher value for each metric between the two conditions is highlighted for each method

Analysing the performances per patient reported in the confusion matrix in fig. 5.11,

it is possible to state that these indicatively follow those for the model trained on the complete signals of its entire morphology (Section 5.3).

This highlights the fact that applying the masking technique did not lead to the expected results, and in general it is better to preserve the entire signal, rather than restricting the analysis to a specific part of it.

Chapter 6

Conclusion

This study demonstrated how it is possible to develop a non-invasive, universally accessible, cost-effective and easily available method for the early diagnosis of pulmonary hypertension. This was achieved by developing a processing and deep learning algorithm for heart sound analysis, enabling the automatic detection of patterns associated with the disease.

The results obtained show that, while it is possible to develop a reliable classification system, the key to success depends more on the quality and configuration of the dataset used for training rather than on the deep learning model itself.

Indeed, for the neural network to be able to learn information that is truly useful for diagnosis, it is essential to start from high quality input data that contains relevant and well distinguishable features. A balanced dataset, with a sufficient number of examples for both classes (patients with and without pulmonary hypertension), proved to be a decisive element in the progressive improvement of the model's performance.

The review made on the literature review confirms that the choice of method and dataset size play a crucial role in verifying model performance.

While deep learning methods emerge as the most promising for large-scale analyses, traditional approaches remain valuable for smaller studies, as they provide interpretable features and require fewer computational resources.

Additionally, the reliability of results must always be verified according to the validation technique employed, as improper validation can lead to overfitting and misleading conclusions.

This perspective aligns with the findings of this study, where an initial dataset with limited size and not opitmal quality led to significant performance fluctuations since it was used to train a deep learning network, while a more structured and balanced dataset improved the robustness of the model.

Througout the study, different configurations of the dataset were tested, starting with the one consisting of the 23 patients with the mPAP data from the RHC examination as ground truth, moving on to the one based on the ECHO data, and finally to the final configuration comprising only high-quality echocardiographic signals.

This process showed a gradual improvement in the performance of the model, with a final auROC of 0.72, confirming that, with a larger dataset of adequate quality, more reliable results can be obtained.

In particular, the performance obtained with the first dataset (RHC) showed that this was highly sensitive to signal quality, demonstrating that even a small group of patients with non-optimal signals, out of a total of 23, could drastically alter performance and prevent the neural network from learning effectively.

This was also evident from the loss and accuracy curves, where those for the validation set showed large fluctuations in all folds, which demostrates a lack capacity of learning from the net. This made it non possible to use a leave-one-out validation strategy, as the inclusion of a single patient in the validation set would have compromise the overall performance too much, making it strictly dependent on the quality of the single patient's signals and thus making the model evaluation unreliable.

When the dataset was expanded to 178 patients with ECHO data, the problem of signal quality was less relevant as the number of signals was markedly increased, but the main obstacle remained the imbalance of classes. With a much higher number of negative than positive subjects, the model showed a strong propensity to misclassify positive patients, and none of the post-processing methods tested could effectively compensate for this problem.

Only with the latest configuration of the dataset, in which high-quality data and a balanced number of positive and negative patients were selected, did the model begin to provide more reliable results. In this configuration, the model achieved an accuracy of 73.33%, a sensitivity of 84.62%, a specificity of 64.71% and an F1-score of 73.33%, with an AUROC of 72%. This demonstrates that, when provided with adequate data, the model is able to learn effectively from the available data, allowing it to also evaluate which post-processing method is most suitable.

Regarding the performance obtained on the final configuration of the dataset, it can be stated that among the various aggregation methods tested to obtain a patient-level classification, the simplest one, based on the average probability per patient and the use of a fixed threshold of 0.5, proved to be the most effective in the final configuration of the dataset.

However, the other post-processing methods may be useful in different scenarios, where for example there is a heterogeneous distribution of patients within the different folds, and therefore it may be convenient to use optimised thresholds for each of them. Based on the results obtained from our model, methods that optimise the threshold for each individual fold have been shown to improve the sensitivity of the model, albeit at the expense of specificity. This means that these methods could be advantageous in contexts where the main objective is to identify as many patients with the disease as possible, even at the cost of including some false positives.

In the clinical setting, this approach could be justified in situations where it is preferable to overestimate the presence of the disease rather than risk not detecting it, especially if early diagnosis can lead to timely and potentially life-saving treatment. For example, in the case of serious diseases such as pulmonary hypertension, a high-sensitivity model may be preferable because it ensures that the majority of affected patients are identified and undergo further diagnostic investigations.

However, it is important to consider that an excessive number of false positives could lead to an overload of the healthcare system, increasing the number of unnecessary invasive examinations and the associated costs.

Therefore, the choice of the most appropriate post-processing method will depend on the balance between sensitivity and specificity required by the specific clinical context.

6.1 Limitations of the Study and Future Developments

While the study achieved promising results, several limitations must be acknowledged.

One of the primary constraints was the limited number of patients with the most reliable ground truth, obtained through RHC.

The small size of this subset made it insufficient for training a deep learning model effectively, leading to considerable fluctuations in performance and limiting the network's ability to generalize.

Additionally, the study was conducted under real-world conditions, meaning that the model had to deal with variability in signal quality. This significantly impacted its ability to learn, as low-quality signals introduced noise and potential misclassifications, hindering performance.

To address these limitations and improve future models, several potential developments can be considered:

- Larger and More Diverse Datasets: Expanding the dataset to include a larger number of patients, especially those with ground truth obtained through RHC, would enable the model to be trained on a dataset with fully reliable ground truth.
- Integration of Multi-Modal Data: Combining heart sound analysis with other diagnostic signals, such as ECG, could provide additional features for the model to learn from, improving classification accuracy.
- Using Data Augmentation to artificially increase the number of high-quality training examples could help compensate for data scarcity.

In conclusion, the study demonstrated that, through the use of a quality dataset and appropriate post-processing techniques, a non-invasive and cost-effective method for the diagnosis of pulmonary hypertension can be achieved. Although the deep learning model was a key element, the quality of the dataset and its balance between the two classes were decisive factors in improving performances.

Bibliography

[1] Marius M. Hoeper, Hossein A. Ghofrani, Ekkehard Grünig, Hans Klose, Horst Olschewski, and Stephan Rosenkranz. «Pulmonary Hypertension». In: *Deutsches Ärzteblatt International* 114.5 (Feb. 2017), pp. 73–84. DOI: 10.3238/arztebl.2017.0073 (cit. on pp. 1–3).

[2] Christian F. Opitz et al. «Pre-Capillary, Combined, and Post-Capillary Pulmonary Hypertension». In: *JACC* 68.4 (2016), pp. 368–378. DOI: 10.1016/j.jacc.2016.05.047. eprint: https://www.jacc.org/doi/pdf/10.1016/j.jacc.2016.05.047. URL: https://www.jacc.org/doi/abs/10.1016/j.jacc.2016.05.047 (cit. on pp. 1, 2).

[3] Shelsey Johnson, Natascha Sommer, Katherine Cox-Flaherty, Norbert Weissmann, Corey E. Ventetuolo, and Bradley A. Maron. «Pulmonary Hypertension: A Contemporary Review». In: *American Journal of Respiratory and Critical Care Medicine* 208.5 (Sept. 2023), pp. 528–548. ISSN: 1535-4970. DOI: 10.1164/rccm.202302-0327SO. URL: https://doi.org/10.1164/rccm.202302-0327SO (cit. on p. 1).

[4] Marius M. Hoeper et al. «Elderly patients diagnosed with idiopathic pulmonary arterial hypertension: results from the COMPERA registry». In: *International Journal of Cardiology* 168.2 (Sept. 2013). Epub 2012 Nov 17, pp. 871–880. DOI: 10.1016/j.ijcard.2012.10.026 (cit. on p. 2).

[5] Richard A. Krasuski et al. «Outcomes in Patients Receiving Treatment for Pulmonary Arterial Hypertension Associated With Repaired Congenital Heart Disease». In: *JACC: Advances* 4.3 (2025), p. 101626. DOI: 10.1016/j.jacadv.2025.101626. eprint: https://www.jacc.org/doi/pdf/10.1016/j.jacadv.2025.101626. URL: https://www.jacc.org/doi/abs/10.1016/j.jacadv.2025.101626 (cit. on p. 2).

[6] Dalma Horvat, Rares Ilie Orzan, and Lucia Agoston-Coldea. «A Non-Invasive Approach to Pulmonary Hypertension». In: *Journal of Clinical Medicine* 14.5 (2025). ISSN: 2077-0383. DOI: 10.3390/jcm14051473. URL: https://www.mdpi.com/2077-0383/14/5/1473 (cit. on p. 2).

[7] Surinder Janda, Neal Shahidi, Kenneth Gin, and John Swiston. «Diagnostic accuracy of echocardiography for pulmonary hypertension: a systematic review and meta-analysis». In: *Heart* 97.8 (2011), pp. 612–622. ISSN: 1355-6037. DOI: 10.1136/hrt.2010.212084. eprint: https://heart.bmj.com/content/97/8/612.full.pdf. URL: https://heart.bmj.com/content/97/8/612 (cit. on p. 2).

[8] Monica Mukherjee et al. «Guidelines for the Echocardiographic Assessment of the Right Heart in Adults and Special Considerations in Pulmonary Hypertension: Recommendations from the American Society of Echocardiography». In: *Journal of the American Society of Echocardiography* 38.3 (2025), pp. 141–186. ISSN: 0894-7317. DOI: https://doi.org/10.1016/j.echo.2025.01.006. URL: https://www.sciencedirect.com/science/article/pii/S0894731725000379 (cit. on p. 2).

[9] Ria Patel, Jay Pescatore, and Shameek Gayen. «The FEV1/DLCO Ratio as an Effective Predictor of Severity and Survival in COPD-Associated Pulmonary Hypertension: A Retrospective Analysis». In: *Journal of Clinical Medicine* 14.5 (2025). ISSN: 2077-0383. DOI: 10.3390/jcm14051606. URL: https://www.mdpi.com/2077-0383/14/5/1606 (cit. on p. 2).

[10] Michele D'Alto et al. «Accuracy and precision of echocardiography versus right heart catheterization for the assessment of pulmonary hypertension». In: *International Journal of Cardiology* 168.4 (2013), pp. 4058–4062. ISSN: 0167-5273. DOI: https://doi.org/10.1016/j.ijcard.2013.07.005. URL: https://www.sciencedirect.com/science/article/pii/S016752731301190X (cit. on p. 3).

[11] Jingping Xu, Louis G. Durand, and Philippe Pibarot. «A new, simple, and accurate method for non-invasive estimation of pulmonary arterial pressure». In: *Heart* 88.1 (July 2002), pp. 76–80. DOI: 10.1136/heart.88.1.76 (cit. on p. 3).

[12] Robert Smith and Daniel Ventura. «A general model for continuous non-invasive pulmonary artery pressure estimation». In: *Computers in Biology and Medicine* 43.7 (Aug. 2013). Epub 2013 Apr 22, pp. 904–913. DOI: 10.1016/j.compbiomed.2013.04.010 (cit. on p. 3).

[13] Alex Gaudio, Miguel Coimbra, Aurélio Campilho, Asim Smailagic, Samuel Schmidt, and Francesco Renna. «Explainable Deep Learning for Non-Invasive Detection of Pulmonary Artery Hypertension from Heart Sounds». In: (Dec. 2022). DOI: 10.22489/CinC.2022.295 (cit. on pp. 3, 19, 23, 24, 28).

[14] Ling Guo et al. «Development and Evaluation of a Deep Learning–Based Pulmonary Hypertension Screening Algorithm Using a Digital Stethoscope». In: *Journal of the American Heart Association* 14.3 (2025), e036882. DOI: 10.1161/JAHA.124.036882. eprint: https://www.ahajournals.org/doi/pdf/10.1161/JAHA.124.036882. URL: https://www.ahajournals.org/doi/abs/10.1161/JAHA.124.036882 (cit. on p. 4).

[15] A Katende et al. «Use of a Handheld Ultrasonographic Device to Identify Heart Failure and Pulmonary Disease in Rural Africa». In: *JAMA Netw Open* 7.2 (2024), e240577. DOI: 10.1001/jamanetworkopen.2024.0577 (cit. on p. 4).

[16] Tarek Kaddoura et al. «Acoustic diagnosis of pulmonary hypertension: automated speech-recognition-inspired classification algorithm outperforms physicians». In: *Scientific Reports* 6 (Sept. 2016), p. 33182. DOI: 10.1038/srep33182 (cit. on pp. 4, 20, 23, 24, 28).

[17] Francesco Renna, Alex Gaudio, Sandra Mattos, Mark D. Plumbley, and Miguel Tavares Coimbra. «Separation of the Aortic and Pulmonary Components of the Second Heart Sound via Alternating Optimization». In: *IEEE Access* 12 (2024), pp. 34632–34643. DOI: 10.1109/ACCESS.2024.3371510 (cit. on p. 4).

[18] Vishy Mahadevan. «Anatomy of the heart». In: *Surgery (Oxford)* 36.2 (2018), pp. 43–47. ISSN: 0263-9319. DOI: https://doi.org/10.1016/j.mpsur.2017.11.010. URL: https://www.sciencedirect.com/science/article/pii/S0263931917302648 (cit. on pp. 7, 9, 10).

[19] Cindy Stanfield. *Principles of Human Physiology*. 5th. Pearson, 2012. ISBN: 9780321819345 (cit. on p. 7).

[20] Nagwa. *Understanding the Human Circulatory System*. Accessed: 2025-03-04. 2025. URL: https://www.nagwa.com/en/explainers/912123271719/ (cit. on p. 8).

[21] John E. Hall. *Guyton and Hall Textbook of Medical Physiology*. Ed. by Michael E. Hall and Arthur C. Guyton. 14th. Enhanced digital version included. Philadelphia: Elsevier, 2021, p. 1132. ISBN: 9780323672801 (cit. on pp. 7, 10, 11, 13).

[22] Texas Heart Institute. «Heart Anatomy». In: (2025). Accessed: 2025-02-07. URL: https://www.texasheart.org/heart-health/heart-information-center/topics/heart-anatomy/ (cit. on p. 8).

[23] S. Leng, R. S. Tan, K. T. C. Chai, C. Wang, D. Ghista, and L. Zhong. «The electronic stethoscope». In: *Biomed. Eng. OnLine* 14.1 (Dec. 2015), pp. 1–37. DOI: 10.1186/s12938-015-0056-y (cit. on p. 11).

[24] Sofia M. Monteiro, Pedro M. Rodrigues, and João P. S. Cunha. «A Novel Approach to Simultaneous Phonocardiography and Electrocardiography Using a 3D-Printed Electronic Stethoscope». In: *IEEE Access* 11 (2023), pp. 1–10. DOI: 10.1109/ACCESS.2023.10190595 (cit. on pp. 11, 12).

[25] Bjorn Watsjold, Jonathan Ilgen, Sandra Monteiro, Matthew Sibbald, Zachary D. Goldberger, W. Reid Thompson, and Geoff Norman. «Do you hear what you see? Utilizing phonocardiography to enhance proficiency in cardiac auscultation». In: *Perspectives on Medical Education* 10.3 (June 2021). Epub 2021 Jan 12, pp. 148–154. DOI: 10.1007/s40037-020-00646-5. URL: https://doi.org/10.1007/s40037-020-00646-5 (cit. on p. 11).

[26] Joseph K. Perloff. «Auscultatory and phonocardiographic manifestations of pulmonary hypertension». In: *The Medical Clinics of North America* 51 (6 1967), pp. 1513–1527. DOI: 10.1016/S0033-0620(67)80011-2. URL: https://www.sciencedirect.com/science/article/abs/pii/S0033062067800112 (cit. on p. 11).

[27] Ali Harimi, Yahya Majd, Abdorreza Gharabagh, Vahid Hajihashemi, Zeynab Esmaileyan, José Machado, and Joao Tavares. «Classification of Heart Sounds Using Chaogram Transform and Deep Convolutional Neural Network Transfer Learning». In: *Sensors* 24 (Dec. 2022), p. 9569. DOI: 10.3390/s22249569 (cit. on p. 12).

[28] Shuenn-Yuh Lee et al. «Electrocardiogram and Phonocardiogram Monitoring System for Cardiac Auscultation». In: *IEEE Transactions on Biomedical Circuits and Systems* 13.6 (Dec. 2019), pp. 1471–1482. ISSN: 1940-9990. DOI: 10.1109/tbcas.2019.2947694 (cit. on p. 12).

[29] Warren Chan, Meron Woldeyohannes, Rachel Colman, Peter Arand, Alan D. Michaels, John D. Parker, John T. Granton, and Steve Mak. «Haemodynamic and structural correlates of the first and second heart sounds in pulmonary arterial hypertension: an acoustic cardiography cohort study». In: *BMJ Open* 3.4 (Apr. 2013), e002660. DOI: 10.1136/bmjopen-2013-002660 (cit. on pp. 13, 20, 21, 23, 25, 27).

[30] Fabio de Lima Hedayioglu. «Looking at the second heart sound: A multi-facet study». PhD Thesis. Departamento de Ciência de Computadores, Universidade do Porto, 2014. URL: https://hdl.handle.net/10216/79585 (cit. on p. 13).

[31] Vishal Nigam and Raymond Priemer. «A dynamic method to estimate the time split between the A2 and P2 components of the S2 heart sound». In: *Physiological Measurement* 27.7 (July 2006). Epub 2006 Apr 27, pp. 553–567. DOI: 10.1088/0967-3334/27/7/001 (cit. on pp. 13, 14, 20).

[32] M. E. Tavel. *Clinical Phonocardiography and External Pulse Recording*. Year Book Medical Publishers, 1972 (cit. on p. 13).

[33] Siddhant Joshi and A.N. Cheeran. «MATLAB Based Feature Extraction Using MFCC for ASR». In: June 2014 (cit. on p. 14).

[34] Devi Lokesh S. «Speech recognition system using enhanced mel frequency cepstral coefficient with windowing and framing method». In: *Cluster Comput* 22 (2017), pp. 11669–11679. DOI: 10.1007/s10586-017-1447-6. URL: https://link.springer.com/article/10.1007/s10586-017-1447-6 (cit. on p. 15).

[35] S. Davis and P. Mermelstein. «Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences». In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4 (1980), pp. 357–366. DOI: 10.1109/TASSP.1980.1163420 (cit. on p. 15).

[36] Lawrence R. Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, 1993 (cit. on p. 15).

[37] Vibha Tiwari Tiwari. «MFCC and its applications in speaker recognition». In: *Int. J. Emerg. Technol.* 1 (Jan. 2010) (cit. on p. 15).

[38] Md Tanzil Hoque Chowdhury, Khem Poudel, and Yating Hu. «Time-Frequency Analysis, Denoising, Compression, Segmentation, and Classification of PCG Signals». In: *IEEE Access* PP (Sept. 2020), pp. 1–1. DOI: 10.1109/ACCESS.2020.3020806 (cit. on pp. 15, 16).

[39] Khushbakht Iqtidar. «Classification of Cardiac Disorders using PCG Signal Analysis». PhD thesis. Dec. 2021. DOI: 10.13140/RG.2.2.18162.11204 (cit. on p. 15).

[40] Hassaan Malik, Umair Bashir, and Adnan Ahmad. «Multi-classification neural network model for detection of abnormal heartbeat audio signals». In: *Biomedical Engineering Advances* 4 (2022), p. 100048. ISSN: 2667-0992. DOI: https://doi.org/10.1016/j.bea.2022.100048. URL: https://www.sciencedirect.com/science/article/pii/S266709922200024X (cit. on p. 15).

[41] B. Benuwa, Y. Z. Zhan, B. Ghansah, D. K. Wornyo, and F. K. Banaseka. «A Review of Deep Machine Learning». In: *International Journal of Engineering Research in Africa* 24 (2016), pp. 124–136 (cit. on p. 16).

[42] Abien Fred Agarap. «Deep Learning using Rectified Linear Units (ReLU)». In: *CoRR* abs/1803.08375 (2018). arXiv: 1803.08375. URL: http://arxiv.org/abs/1803.08375 (cit. on pp. 17, 18).

[43] Pratima Upretee and Mehmet Emin Yüksel. «13 - Accurate classification of heart sounds for disease diagnosis by using spectral analysis and deep learning methods». In: *Data Analytics in Biomedical Engineering and Healthcare*. Ed. by Kun Chang Lee, Sanjiban Sekhar Roy, Pijush Samui, and Vijay Kumar. Academic Press, 2021, pp. 215–232. ISBN: 978-0-12-819314-3. DOI: https://doi.org/10.1016/B978-0-12-819314-3.00014-8. URL: https://www.sciencedirect.com/science/article/pii/B9780128193143000148 (cit. on p. 17).

[44] Peyman Afshari Bijarbaneh, Fatemeh Noori, Shima Faramarzi, and Seyed Amirhossein Mousavi. «Enhancing Hand Movement Recognition: A Hybrid Fuzzy Deep Neural Network Approach with Time-Frequency EMG Representations». In: *2025 Fifth National and the First International Conference on Applied Research in Electrical Engineering (AREE)*. 2025, pp. 1–4. DOI: 10.1109/AREE63378.2025.10880263 (cit. on p. 17).

[45] Suresh Dara and Priyanka Tumma. «Feature Extraction By Using Deep Learning: A Survey». In: *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. 2018, pp. 1795–1801. DOI: 10.1109/ICECA.2018.8474912 (cit. on p. 17).

[46] Keiron O'Shea and Ryan Nash. «An Introduction to Convolutional Neural Networks». In: *arXiv* 1511.08458 (2015). arXiv preprint arXiv:1511.08458. URL: https://doi.org/10.48550/arXiv.1511.08458 (cit. on p. 18).

[47] Peng Ma, Bin Ge, Haiqing Yang, Tong Guo, Jie Pan, and Wei Wang. «Application of time-frequency domain and deep learning fusion feature in non-invasive diagnosis of congenital heart disease-related pulmonary arterial hypertension». In: *MethodsX* 10 (Jan. 2023), p. 102032. DOI: 10.1016/j.mex.2023.102032 (cit. on pp. 19, 20, 23, 28).

[48] Anne Dennis, Alan D. Michaels, Peter Arand, and Daniel Ventura. «Non-invasive diagnosis of pulmonary hypertension using heart sound analysis». In: *Computers in Biology and Medicine* 40.9 (Sept. 2010). Epub 2010 Aug 6, pp. 758–764. DOI: 10.1016/j.compbiomed.2010.07.003 (cit. on pp. 19, 26, 32, 44).

[49] Meng Wang, Bo Guo, Yu Hu, Zihang Zhao, Chunfeng Liu, and Haiyan Tang. «Transfer Learning Models for Detecting Six Categories of Phonocardiogram Recordings». In: *Journal of Cardiovascular Development and Disease* 9.3 (Mar. 2022), p. 86. DOI: 10.3390/jcdd9030086 (cit. on pp. 19–21, 23, 25, 28).

[50] Mohamed Elgendi et al. «The unique heart sound signature of children with pulmonary artery hypertension». In: *Pulmonary Circulation* 5.4 (Dec. 2015), pp. 631–639. DOI: 10.1086/683694 (cit. on pp. 19, 21, 23, 25, 26).

[51] Mohamed Elgendi, Prerna Bobhate, Sanjay Jain, Lisa Guo, John Rutledge, Julie Y. Coe, Roger Zemp, Dale Schuurmans, and Ian Adatia. «The Voice of the Heart: Vowel-Like Sound in Pulmonary Artery Hypertension». In: *Diseases* 6.2 (Apr. 2018), p. 26. DOI: 10.3390/diseases6020026 (cit. on pp. 19, 23, 27).

[52] Lotfi Hamza Cherif. «Algorithm for the estimation of pulmonary hypertension by the heart sounds using a Hilbert transform». In: *International Journal of Biomedical Engineering and Technology* 20 (May 2016), p. 356 (cit. on pp. 20, 23, 27).

[53] Yiqing Li, Jia Qian, Xiaoyan Dong, Jun Zhao, Qi Wang, Yajun Wang, Xiaofeng Zeng, Zhenlin Tian, and Mengtao Li. «The prognosis and management of reclassified systemic lupus erythematosus associated pulmonary arterial hypertension according to 2022 ESC/ERS guidelines». In: *Arthritis Research Therapy* 26.1 (May 2024), p. 109. DOI: 10.1186/s13075-024-03338-1 (cit. on p. 20).

[54] V. K. Iyer, P. A. Ramamoorthy, Hua Fan, and Y. Ploysongsang. «Reduction of heart sounds from lung sounds by adaptive filtering». In: *IEEE Transactions on Biomedical Engineering* 33.12 (Dec. 1986). Erratum in: IEEE Trans Biomed Eng 1988 Jan;35(1):76, pp. 1141–1148. DOI: 10.1109/TBME.1986.325693 (cit. on p. 20).

[55] Alex Gaudio, Noemi Giordano, Miguel Coimbra, Benedict Kjaergaard, Samuel Schmidt, and Francesco Renna. «Cross-Domain Detection of Pulmonary Hypertension in Human and Porcine Heart Sounds». English. In: *Computing in Cardiology, CinC 2023*. Computing in Cardiology. 2023 Computing in Cardiology (CinC) ; Conference date: 01-10-2023 Through 04-10-2023. United States: IEEE (Institute of Electrical and Electronics Engineers), Oct. 2023, pp. 1–4. ISBN: 979-8-3503-5903-9. DOI: 10.22489/CinC.2023.071 (cit. on pp. 20, 23, 28).

[56] Norihiro Yamakawa, Naoki Kotooka, Takashi Kato, Takafumi Kuroda, and Koichi Node. «Cardiac acoustic biomarkers as surrogate markers to diagnose the phenotypes of pulmonary hypertension: an exploratory study». In: *Heart and Vessels* 37.4 (Apr. 2022). Epub 2021 Oct 1, pp. 593–600. DOI: 10.1007/s00380-021-01943-7 (cit. on pp. 21, 23, 27).

[57] Jun Huang, Wei Zhang, Wenjie Fu, Jia Le, Yan Qi, Xiaojing Hou, Xiaojie Pan, Rong Li, and Bin He. «Noninvasive evaluation of pulmonary hypertension using the second heart sound parameters collected by a mobile cardiac acoustic monitoring system». In: *Frontiers in Cardiovascular Medicine* 10 (Dec. 2023), p. 1292647. DOI: 10.3389/fcvm.2023.1292647 (cit. on pp. 21, 22, 25, 27).

[58] Constantin Tranulis, Louis G. Durand, Lotfi Senhadji, and Philippe Pibarot. «Estimation of pulmonary arterial pressure by a neural network analysis using features based on time-frequency representations of the second heart sound». In: *Medical Biological Engineering Computing* 40.2 (Mar. 2002), pp. 205–212. DOI: 10.1007/BF02348126 (cit. on pp. 21, 26).

[59] Rijuven. *CardioSleeve - The First FDA-Cleared ECG and Digital Auscultation Device*. Accessed: 2025-02-12. 2025. URL: https://www.rijuven.com/cardiosleeve (cit. on p. 29).

[60] Ali Ataya, Sheylan Patel, Jessica Cope, and Hassan Alnuaimat. «Pulmonary arterial hypertension and associated conditions». In: *Disease-a-Month* 62.10 (2016), pp. 382–405. DOI: 10.1016/j.disamonth.2016.03.006. URL: https://pulmonary.medicine.ufl.edu/files/2017/11/Ataya19.pdf (cit. on p. 32).

[61] Md. A. Hossan, S. Memon, and M. A. Gregory. «A novel approach for MFCC feature extraction». In: *2010 4th International Conference on Signal Processing and Communication Systems*. IEEE, Dec. 2010, pp. 1–5. DOI: 10.1109/ICSPCS.2010.5709752 (cit. on p. 33).

[62] Noemi Giordano and Marco Knaflitz. «A Novel Method for Measuring the Timing of Heart Sound Components through Digital Phonocardiography». In: *Sensors* 19.8 (2019). ISSN: 1424-8220. DOI: 10.3390/s19081868. URL: https://www.mdpi.com/1424-8220/19/8/1868 (cit. on p. 34).

[63] Stacy A. Mandras, Hirsch S. Mehta, and Anjali Vaidya. «Pulmonary Hypertension: A Brief Guide for Clinicians». In: *Mayo Clinic Proceedings* 95.9 (2020), pp. 1978–1988. ISSN: 0025-6196. DOI: https://doi.org/10.1016/j.mayocp.2020.04.039. URL: https://www.sciencedirect.com/science/article/pii/S0025619620306121 (cit. on p. 40).