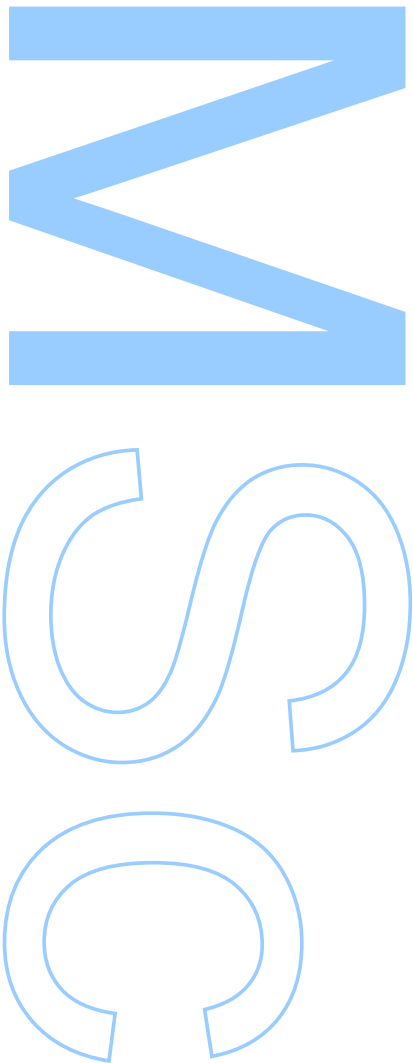
Multimodal deep learning for heart sound and electrocardiogram classification

Hélder Vieira
Master Degree in Data Science
Departamento de Ciência dos Computadores
2023

Orientador
Dr. Francesco Renna, Professor Auxiliar, Faculdade de Ciências da Universidade do Porto

Coorientador
Dr. Miguel Coimbra, Professor Catedrático, Faculdade de Ciências da Universidade do Porto




U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Acknowledgements

First of all, I would like to thank my supervisor Prof. Francesco Renna for the guidance and assistance provided, continuously giving feedback, insightful comments and suggestions, providing a huge support in this journey and the opportunity to work on such an interesting topic. I would also like to thank co-supervisor Prof. Miguel Coimbra for the guidance provided. I would also like to present a special thanks to the Physionet team and involved community for publicly providing the data that allowed the development of this project. Finally, I thank my family for the constant unending support.

Resumo

As doenças cardiovasculares são a principal causa de morte em todo o mundo, afetando particularmente os países mais carenciados. O diagnóstico precoce e a terapia adequada são cruciais no combate a este problema de saúde global.

Esta tese apresenta uma nova abordagem para a avaliação da saúde cardiovascular através da análise multimodal de sinais de eletrocardiograma e fonocardiograma capturados simultaneamente. Estas ferramentas de diagnóstico não invasivas e de custo eficaz fornecem informações complementares sobre as atividades elétricas e mecânicas do coração. Foram desenvolvidos e implementados métodos de aprendizagem profunda, especificamente Redes Neurais Convolucionais (CNN), para a análise destes sinais.

Os escalogramas extraídos de segmentos curtos de sinal são usados como base para o processo de aprendizagem.

Para enfrentar o desafio da limitada disponibilidade de dados publicamente acessíveis, foram desenvolvidas abordagens de *transfer learning* para tirar partido do conhecimento de outros domínios.

A abordagem mais bem-sucedida alcançou uma *accuracy* de 82,79%, um ROC AUC de 91,26% e um F1-score de 88,52%. Comparativamente, o melhor modelo individual teve uma *accuracy* de 81,15%, um ROC AUC de 88,51% e F1-score de 87,01%, demonstrando o potencial da análise multimodal e do *transfer learning* na deteção de doenças cardiovasculares. Este estudo fornece uma base para futuras pesquisas destinadas a melhorar o desempenho dos sistemas multimodais de deteção de anomalias cardíacas.

Palavras-chave: Eletrocardiograma, Fonocardiograma, Sons cardíacos, Multimodal, Aprendizagem profunda, Diagnóstico

Abstract

Cardiovascular diseases are the leading cause of death worldwide, particularly affecting low and middle-income countries. Early diagnosis and appropriate therapy are crucial in combating this global health issue.

This thesis presents a novel approach to the assessment of cardiovascular health through the multimodal analysis of simultaneously recorded electrocardiogram and phonocardiogram signals. These cost-effective, noninvasive diagnostic tools provide complementary information about the heart's electrical and mechanical activities. Deep learning methods, specifically Convolutional Neural Networks (CNN), were developed and implemented for the analysis of these signals.

Scalograms extracted from short signal segments are used as the foundation for the learning process.

To tackle the challenge of limited publicly available data, transfer learning approaches were employed to leverage knowledge from other domains.

The most successful multimodal approach achieved an accuracy of 82.79%, a ROC AUC score of 91.26%, and a F1-score of 88.52%. Comparatively, the best performing single input model had an accuracy of 81.15%, a ROC AUC score of 88.51%, and a F1-score of 87.01%, demonstrating the potential of multimodal analysis and transfer learning in the detection of CVDs. This study provides a foundation for future research aimed at enhancing the performance of multimodal cardiac abnormality detection systems.

Keywords: Electrocardiogram, Phonocardiogram, Heart sounds, Multimodal, Deep learning, Diagnosis

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cardiovascular diseases (CVDs) are the number one cause of death globally, taking approximately 17.9 million lives each year corresponding to an estimated 32% of all deaths worldwide, according to the World Health Organization (WHO) [1, 2]. Among these deaths, 85% were due to heart attack and stroke.

More 75% of CVDs related deaths occur in low and medium income countries [1]. Early diagnosis and appropriate therapy are often effective in eliminating or delaying the disease progression [3].

Access to appropriate healthcare systems (including efficient disease monitoring and diagnosis) is not always possible and affordable, especially on low-income countries. The poorest people in low and middle-income countries are most affected. At the household level, evidence is emerging that CVDs and other noncommunicable diseases contribute to poverty due to catastrophic health spending and high out-of-pocket expenditure. At the macro-economic level, CVDs place a heavy burden on the economies of low and middle-income countries [1].

Electrocardiogram (ECG) and phonocardiogram (PCG) are cost effective, painless, noninvasive and complementary diagnosis tools that provide relevant information in the assessment of the cardiovascular health. Separate acquisition and automatic analysis of this signals is already a valuable resource in the diagnosis of CVDs and for that several systems have been developed, such as the DigiScope2 [4] for PCG and the BITalino for ECG [5].

Information provided by the ECG is mostly related with the heart's eletrical activity, while PCG reflects the cardiac mechanical activity, complementing each other. The multimodal analysis of simultaneously recorded PCG and ECG has a substantial potential for

the assessment of the cardiac health [6].

Over the years, researchers have extensively analyzed PCG or ECG separately using deep learning methods to detect cardiovascular diseases.

1.1 Objectives

The main aim of this thesis is to study and implement deep learning based methods to classify simultaneously obtained PCG and ECG signals. The development of highly accurate, automatic, inexpensive CVDs detection systems can contribute to an early diagnosis and treatment and therefore possibly allow a reduction on the number of CVDs related deaths. Associated with the main objective the following sub-objectives are set:

- Characterize the current solutions for multimodal PCG and ECG analysis.
- Demonstrate the effectiveness of multimodal analysis for anomaly detection, when compared to single input analysis.
- Exploration of different approaches to combine the information from PCG and ECG.

1.2 Contributions

The main contribution of this work is the development of the preprocessing and machine learning framework. The developed framework is responsible for transforming the raw data and extracting meaningful features from it, which are then used as input for the machine learning models.

In addition, this study provides a comprehensive review and analysis of existing literature in the field of multimodal classification of PCG and ECG signals and a description of the available databases.

Another significant contribution is the application of transfer learning techniques to overcome the challenge of limited publicly available data. By leveraging knowledge acquired in other domains, the developed models are able to learn more effectively even with limited training data.

Furthermore, this work contributes to the field by demonstrating the potential of multimodal analysis in the detection of cardiovascular diseases. By combining ECG and PCG signals, the developed models are able to capture a more holistic view of the heart's activities, leading to improved performance in disease detection.

Finally, the results of this work provide valuable insights into the performance of various machine learning techniques in the context of cardiovascular disease detection. These insights could guide future research in this field and contribute to the development of more effective diagnostic tools.

1.3 Outline

The thesis is organized as follows:

- **Chapter 2:** Provides a brief description of the anatomy of physiology of the heart. The PCG and the ECG are described. An introduction to machine learning, focusing specifically on deep learning, is also presented in this chapter.
- **Chapter 3:** Presents the current state-of-the-art of multimodal PCG and ECG machine learning approaches and the publicly available databases that contain such signals.
- **Chapter 4:** Describes the methodology used, namely the models implemented and the development process.
- **Chapter 5:** Presents the results obtained with the implemented models.
- **Chapter 6:** Presents the main conclusions of the work developed indicating possible future research directions.

Chapter 2

Background

2.1 Heart physiology

PCG and ECG are two of the most important biomedical signals to assess the condition of the heart. Briefly, PCG presents the timing, duration and amplitude of different heart sounds and allows the detection of structural defects of the heart valves by analyzing heart sound signals. ECG represents the electrical activity of a heart. ECG is used to monitor the cyclical contraction and relaxation of the human heart muscles. In this section, the structure of the human heart and the characteristics of PCG and ECG signals are discussed.

2.1.1 Cardiac anatomy and physiology

The heart is one of the most important organs of the human body. Its main function is to pump adequate blood to the entire body through a network of veins and arteries. The surface of the heart is surrounded by the coronary arteries that provide oxygen and nutrients rich blood to the heart muscles and take way the waste products. The heart has two main functions:

- Collect oxygen-rich blood from the lungs and send it to all the tissues of the body, then oxygen and carbon dioxide are exchanged from the blood to the tissue cells.
- Collect the blood rich in carbon dioxide from the tissues and send it to the lungs, then carbon dioxide and oxygen are exchanged from the alveolar blood to the alveolar air.

As seen in Figure 2.1, the heart is composed by four chambers: the right atrium, the right ventricle, the left atrium, and the left ventricle. The right atrium receives blood from the veins and pumps it to the right ventricle through the tricuspid valve. The right ventricle receives blood from the right atrium and pumps it to the lungs. The left atrium receives oxygenated blood from the lungs and pumps it to the left ventricle through the mitral valve. The left ventricle pumps oxygen-rich blood to the rest of the body [7].

The internal heart structure and components compel the blood to flow in one-way only. The atrioventricular valves allow blood to flow only from the atriums to the ventricles. The semilunar valves allow blood to flow out of the heart from the ventricles to the great arteries, as shown in Figure 2.1.

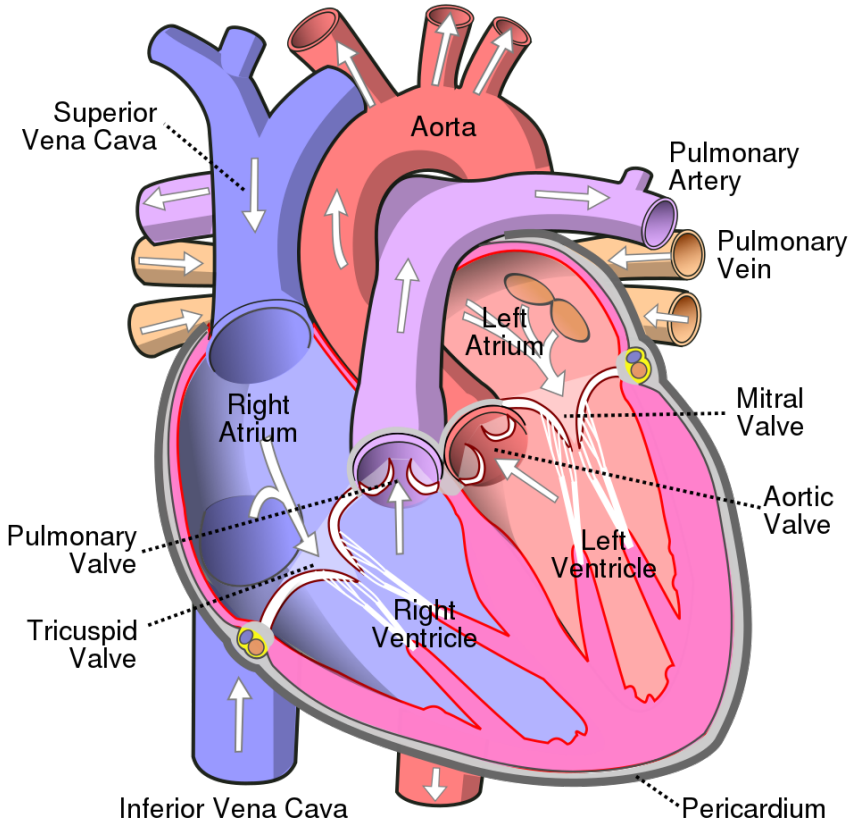


FIGURE 2.1: The heart, showing valves, arteries and veins. The white arrows show the normal direction of blood flow [8].

2.1.2 Cardiac cycle

A cardiac cycle is defined as a complete heartbeat. It consists of a complete relaxation and contraction of both the atria and ventricles. It defines the electrical and mechanical

activities of the heart throughout the systole and diastole interval. Systole interval is the duration of the cardiac contraction and the diastole interval is the duration of the cardiac relaxation. The average duration of a cardiac cycle is around 0.8 seconds.

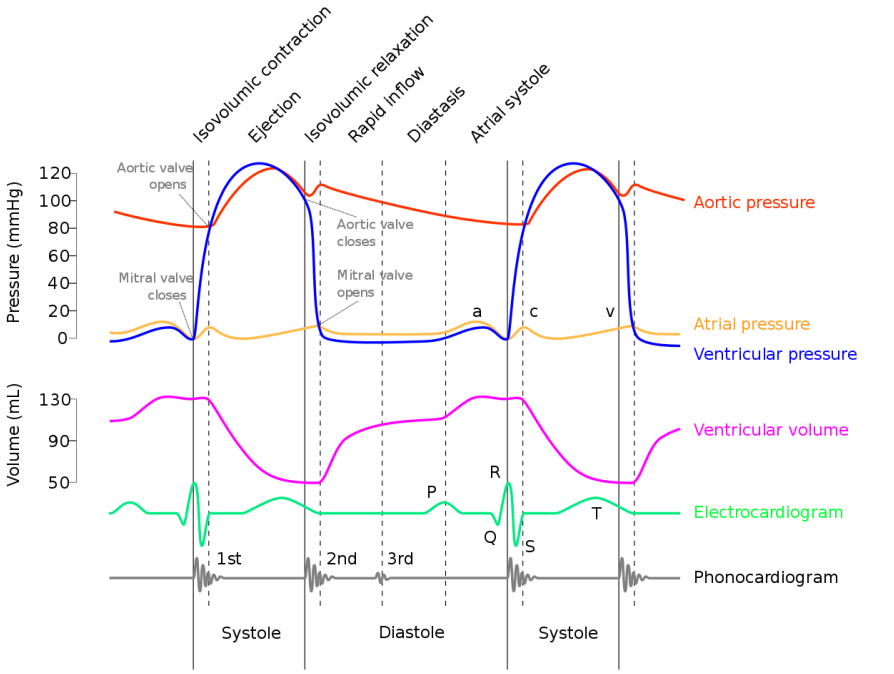


FIGURE 2.2: The Wiggers diagram representing the behavior of typical pattern of certain signals (including PCG and ECG) in the cardiac cycle [9].

During the systole, large amounts of blood are stored in the atriums since the atrioventricular valves are closed. At the end of the systole, the atrioventricular valves open suddenly due to the increasing pressure in the atriums and the decreasing pressure in the ventricles, a period known as the rapid filling of the ventricles. This period of rapid filling corresponds to 2/3 of the diastolic time and the last 1/3 corresponds to the atrial contraction. This rapid filling results in a rising pressure in the ventricles, causing at the end of the diastole the closing once again of the atrioventricular valves, the resulting vibration is low in pitch and relatively long-lasting and it is known as the first heart sound (S1). On the other hand, the semilunar valves do not open immediately and it takes around 0.02 to 0.03 seconds to do it, during this period, contraction is occurring in the ventricles, but there is no emptying a period known as isometric contraction. When the left ventricular pressure rises slightly above 80 mm Hg (and the right ventricular pressure slightly above 8 mm Hg), the semilunar valves open and blood is ejected outside of the ventricles, this period of ejection corresponds to the systole. At the end of this period, ventricular relaxation begins suddenly, allowing both the right and left intraventricular

pressures to decrease rapidly, in contrast the pressure in large arteries are very high since they have been just filled with blood from the contracted ventricles, at the end of this period, some expected blood flows back to the ventricles, forcing the aortic and pulmonary valves to close resulting in a rapid snap sound called the second heart sound (S2). The ventricle muscle continues to relax (isometric relaxation) and the intraventricular pressures decrease rapidly compelling once again atrioventricular valves to open, therefore marking the beginning of a new heart cycle [10].

2.1.3 PCG signal

PCG is a diagnosis tool providing the graphical depiction of heart sounds and murmurs. It helps to monitor various components of the heart sounds through the heart cycle. Heart sounds are produced due to the flow of blood across the heart valves, the opening and the closure of the heart valves, and from the mechanical actions of heart muscles. These heart sounds are primary monitoring technique for diagnosing different cardiac diseases. Doctors and cardiologists usually use stethoscope to hear the heart sounds before any clinical diagnosis. Heart sounds are similar in all healthy hearts. Abnormal heart sounds are related to cardiovascular diseases. The interval from the starting point of S1 to the starting point of S2 is called the systole interval (S1-S2 interval) and the interval from the starting point of S2 to the starting point of S1 is called the diastole interval (S2-S1 interval). Diastole interval is usually longer than the systole interval. Beside S1 and S2, two extra heart sounds known as third and fourth heart sound (S3 and S4) can appear in both normal and pathological conditions. S3 appears just after S2, and S4 appears just before S1. Besides these heart sounds different kinds of heart murmurs may also present in the signal which are produced because of the turbulent flow of blood across the valves and related to the cardiac diseases. Murmurs may present in systole or diastole or in both intervals. Figure 2.3 shows a pathological PCG recording with the presence of murmurs and the S3 sound. Murmurs usually have higher frequency compared to the heart sounds. When the blood circulates through the heart valves and chamber, sometimes it produces innocent murmurs which is not related to any cardiac diseases. There are mainly three kinds of heart murmurs:

- Systolic murmurs: Start after S1 and ends before S2.
- Diastolic murmurs: Start after S2 and end before S1.

- Continuous murmurs: Usually occur throughout or some parts of the cardiac cycle.

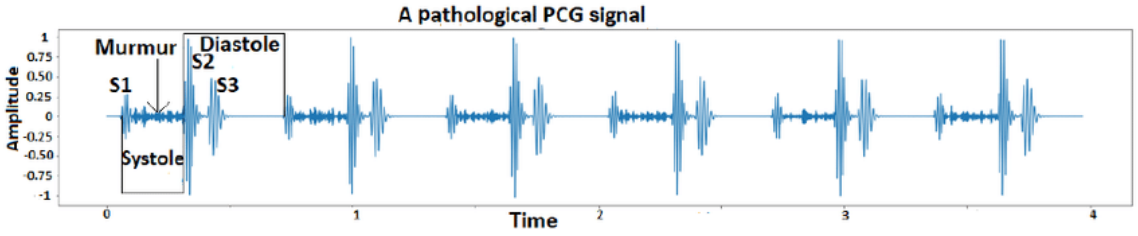


FIGURE 2.3: A pathological PCG recording [11].

Cardiac auscultation consists on an approach performed by physicians while listening to heart sounds using a stethoscope. It involves listening the heart on specific points, each of them near a cardiac valve (with the exception of the Erb's point located at the third intercostal space and the left lower sternal border), enabling the detection of murmurs associated with valvular abnormalities.

There is not a standard methodology to collect heart sounds, although two systematic procedures are well accepted: the physician should first auscultate the right upper sternal border, followed by the left upper sternal border. Afterwards the down left sternal and finally the apex is also auscultated. The other way around is also acceptable as long the sequence is mantained (see Figure 2.4). In each spot, the frequencies listened are dominated by a unique heart valve, allowing to uniquely assess the mechanical properties of a specific heart valve [10].

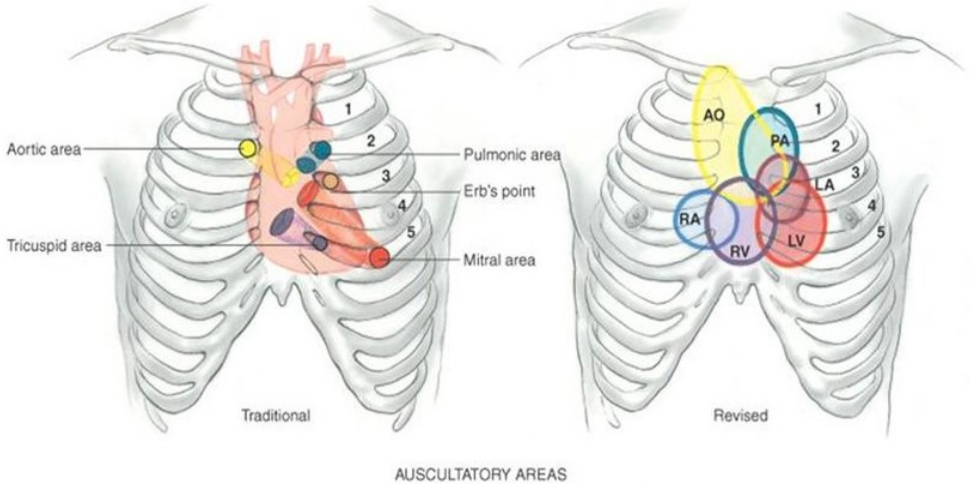


FIGURE 2.4: Auscultation spots (adapted from [12]).

2.1.4 ECG signal

The ECG is an essential tool in the evaluation of the cardiac function. The detailed study and methodological analysis of the ECG components (waves, intervals and segments) form the basis of its interpretation. It has many applications in the clinical diagnosis and prognosis of CVDs, as well as in health assessment, biomedical recognition, fatigue studies, and other areas [3].

When the cardiac impulse spreads through the eletrical conduction system of the heart (see Figure 2.5), electromagnetic waves also spread from the heart into the adjacent tissues surrounding the heart. These are detected and recorded by placing electrodes on opposite sides of the heart creating the ECG signal.

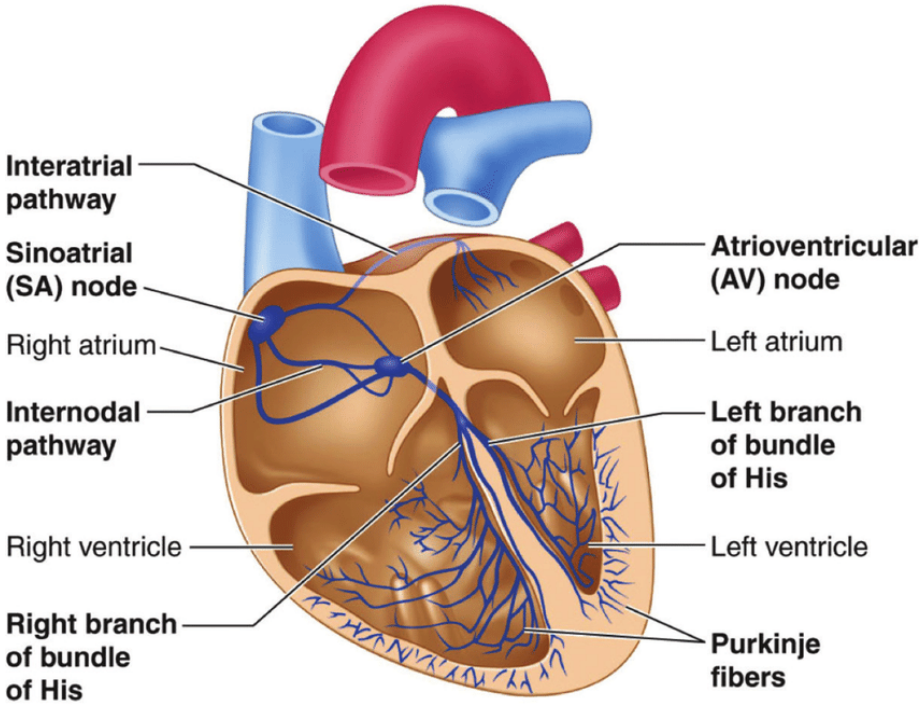


FIGURE 2.5: Electrical conduction system of the heart [13].

ECG is the graphical record of changes in the magnitude and direction of the electrical activity of the heart. More specifically, the electric current that is generated by the depolarization and repolarization of the atria and ventricles can be monitored through the ECG signal. The ECG signal is captured through an array of electrode sensors known as leads. Leads are attached to the skin to detect the electrical activity of the heart. This information is recorded on a graph. As the electrical signal traverses through the heart, the graph shows each phase of the signal. Under normal condition, the ECG signal has a very predictable direction, duration, and amplitude. Any change in the ECG signal

is usually related to cardiac abnormalities. Therefore, by analyzing ECG signal continuously, it is possible to detect any abnormal heart function in the primary stage which helps cardiologists for the proper clinical diagnosis. However, reliable and efficient clinical applications are highly dependent on the accuracy of information extracted from the ECG recording. Usually ECG signals are subjected to contamination by various noises. The sources of noise may be either cardiac or extra-cardiac. Reduction or disappearance of the isoelectric interval, prolonged repolarization, and atrial flutter are responsible for cardiac noise. Respiration, changes of electrode position, muscle contraction, and power line interference can cause extra-cardiac noise.

The ECG signal is composed of 5 waves - P, Q, R, S and T. This signal can be measured by electrodes from human body in typical engagement. Signals from these electrodes are brought to simple electrical circuits with amplifiers and analogue to digital converters. The muscle mass of the atria is small compared with the ventricles, and the electrical change of the atria is very small. Contraction of atria associated with the ECG wave is called P. For the large mass of ventricular, it has large deflection which is called QRS complex. The T wave of the ECG is associated with the return of the ventricular mass to its resting electrical state. Figure 2.6 shows the basic shape of a normal ECG signal. Different ECG waves and their properties are given bellow:

- P-wave: It occurs due to the depolarization of atrial muscle. The amplitude of P wave is around 0.25 mV.
- QRS complex: It occurs due to the repolarization of atria and depolarization of ventricles. The amplitude of R wave is around 1.60 mV. The amplitude of Q wave is around 25% of R wave. The time duration QRS interval is around 0.09 seconds.
- T-wave: It happens due to the ventricular repolarization. The amplitude of T wave is around 0.1 to 0.5 mV.
- U-wave: If present, it comes after potential in the ventricular muscle and represents repolarization of the purkinje fibers.

In a normal cardiac cycle, the P wave occurs first, followed by the QRS complex and the T wave. The section of the ECG between the waves and complexes are called segments and interval such as the PR segment, the ST segment, the TP segment, the PR interval, the QT interval, and the R-R interval. When the electrical activity of the heart is not being detected the ECG is a straight, flat line.

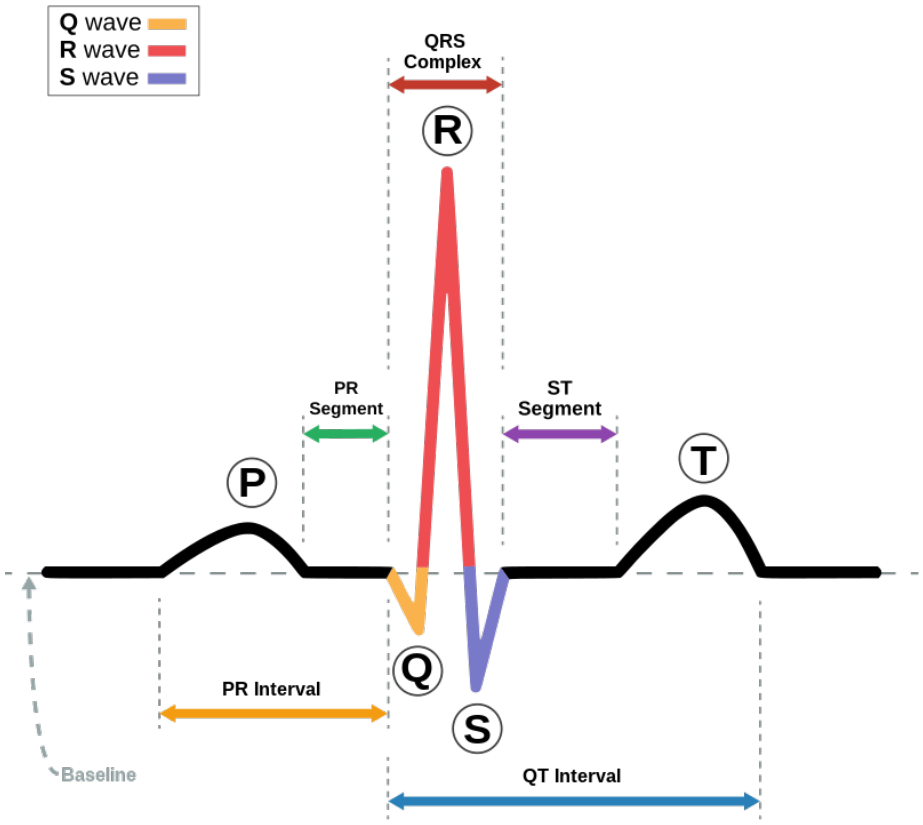


FIGURE 2.6: Schematic diagram of the basic shape of a normal ECG [14].

The most common technique to record an ECG signal is the 12-lead ECG method. The standard ECG has 12 leads. Six of the leads are considered as limb leads, as they are placed on the arms and/or legs of the individual. The other six leads are considered as chest leads because they are placed on the chest. The six limb leads are called lead I, II, III, aVL, aVR and aVF. The letter "a" stands for "augmented", as these leads are calculated as a combination of leads I, II and III. The six chest leads are called leads V1, V2, V3, V4, V5 and V6. Leads I, II and III are each making use of a pair of electrodes (bipolar), with one electrode measuring between itself and the other. Leads aVR, aVL, and aVF make use of all the connections to the patient. Each of the six pericardial or chest electrodes (V1-V6) represent six different views and unique information that cannot be derived from other leads.

2.1.5 Relationship between PCG and ECG signals

The PCG defines the mechanical activity of the heart and the ECG defines electrical activity of the heart. Mechanical activity of the heart is known as the opening and closure of heart valves and the sound they produce during the cardiac cycle. This mechanical function relies on the electrical operation of the heart. So, if there is any defect in the electrical action of the heart, the mechanical function of the heart will also be affected. Therefore, ECG and PCG signals are correlated with each other. From Figure 2.7 we can see that, in healthy subjects, the 1st heart sound (S1) appears 0.04 second to 0.06 second after the beginning of the QRS complex. The second heart sound (S2) starts at the end of the T wave. The third heart sound occurs after the T wave and before the P wave. The fourth heart sound (S4) occurs after the P wave and before the QRS complex. S3 and S4 both occur during the diastolic period [15].

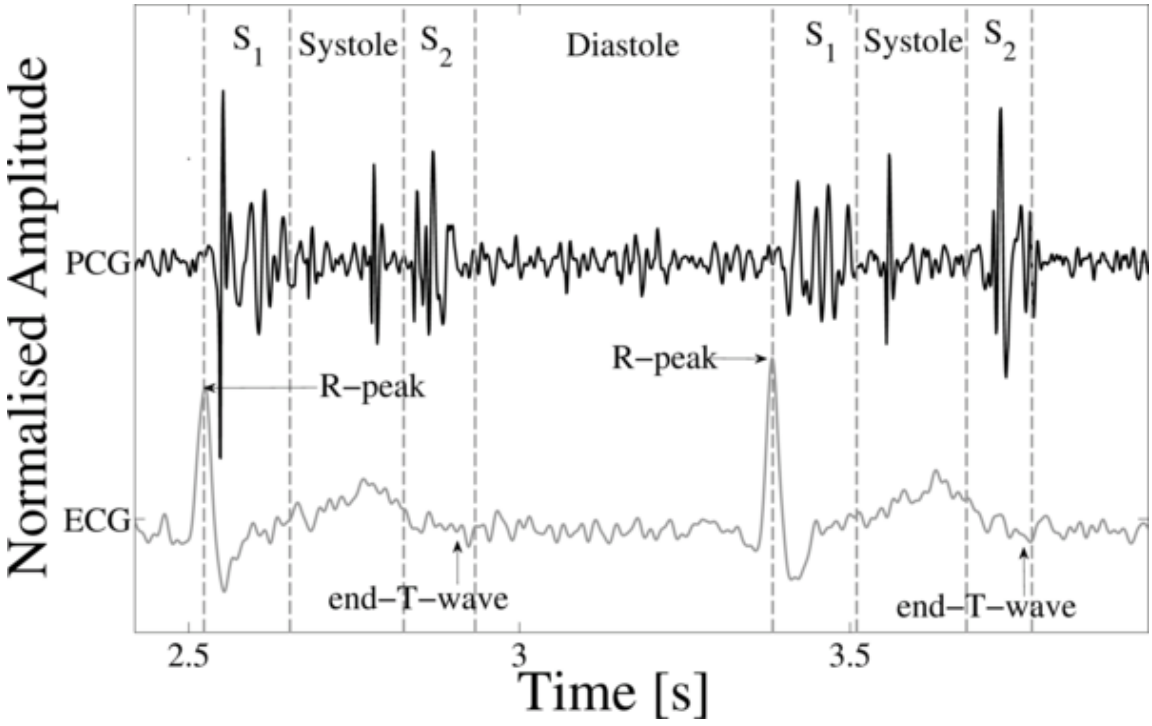


FIGURE 2.7: A PCG (top tracing), with simultaneously recorded ECG (lower tracing) and the four states of the PCG recording; S1, Systole, S2 and Diastole [16].

2.2 Machine Learning

Machine Learning (ML) can be identified as a branch of Artificial Intelligence (AI), a big field that generally can be described as an approach for adopting human cognitive thinking abilities, developing the ability to learn without being explicitly programmed. The hierarchical relationship between data science, ML, AI and deep learning (DL) is shown in Figure 2.8. The concept of AI was first developed in the middle of the 20th century and started its development with various stages; the adoption of AI as a scientific method started to increase from the beginning of the recent 21st century. It was also triggered by the availability of an enormous amount of data and increasing computational power, which made the use of AI possible in practical areas, namely healthcare [17].

Basic principles of those subfields of Data Science will be explored further in this chapter.

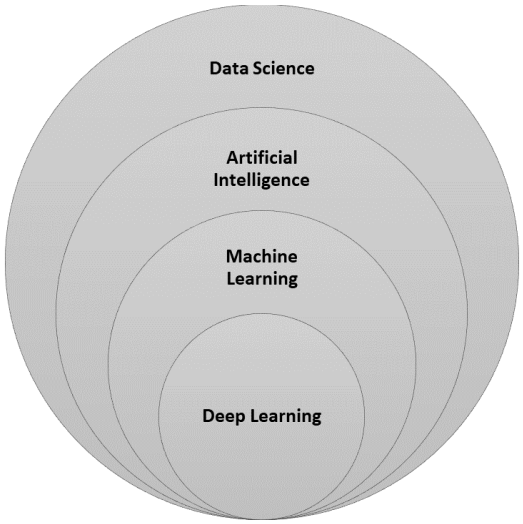


FIGURE 2.8: Hierarchical relationship between data science, artificial intelligence, machine learning and deep learning

2.2.1 ML algorithms

ML algorithms can be divided into three main categories: supervised, unsupervised and reinforcement learning [18].

Supervised learning systems use labeled datasets, essentially being taught by example. The training set of input-output pairs is used to find a deterministic function that maps inputs to the respective output value. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately. This training dataset includes

inputs and outputs, which allow the model to learn over time. The algorithm measures its performance through the loss function, adjusting until the error has been sufficiently minimized.

Supervised learning can be mainly separated into two types of problems:

- Classification: the learning algorithm is used to accurately assign test data into specific categories. It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labeled or defined. Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbors (kNN), and random forest.

- Regression: aims to model the relationship between the features and a continuos target variable. Examples of popular regression models include linear regression and polynomial regression.

Supervised learning models can be used to build and advance a number of business applications, including image and object recognition, customer sentiment analysis and spam detection [19].

Unlike supervised learning, unsupervised learning uses unlabeled data. From that data, it discovers patterns that help solve for clustering or association problems. This is particularly useful when subject matter experts are unsure of common properties within a data set. Common clustering algorithms are hierarchical, k-means, and Gaussian mixture models [19]. Costumer segmentation and recommender systems are popular applications of this learning strategy.

Reinforcement learning can be broadly described as an action-reward system. The training and test phases are interconnected in the reinforcement learning process. The learning process goes through user feedback for each guess or action that improves the learning. The learning system calls an agent, after each action performed, the system receives rewards or penalties (in the form of negative rewards). The algorithm aims to learn by itself to receive a maximum reward as a result of its cases of action. Popular algorithms include Markov decision process and Q-learning [17].

2.2.2 Feature extraction

Feature extraction is a process of deriving a compact and useful representation of the signal information. Feature extraction and selection is an important step in several classification systems.

Regarding signals, such as the ECG and the PCG, features can be extracted with respect to different domains, namely: time domain (for example the duration of certain morphologically defined segments), frequency domain and time-frequency domain.

The Fourier transform is a widely used technique for providing resolution on the frequency domain, it can show precisely the frequency content of a signal, but it does not provide temporal resolution, meaning that there is no knowledge of when said frequencies occurred in time. The Short-Time Fourier Transform (STFT) improves on this by computing a sequence of Fourier transforms throughout the time series [20]. STFT provides the time-localized frequency information to create what is called a spectogram, whereas the standard Fourier transform provides the frequency information averaged over the entire signal time interval. The spectrogram is obtained by windowing the input signal with a window of constant length (duration) that is shifted in time and frequency [21]. The application of a constant window, leads to a fixed time-frequency resolution (as shown in 2.9).

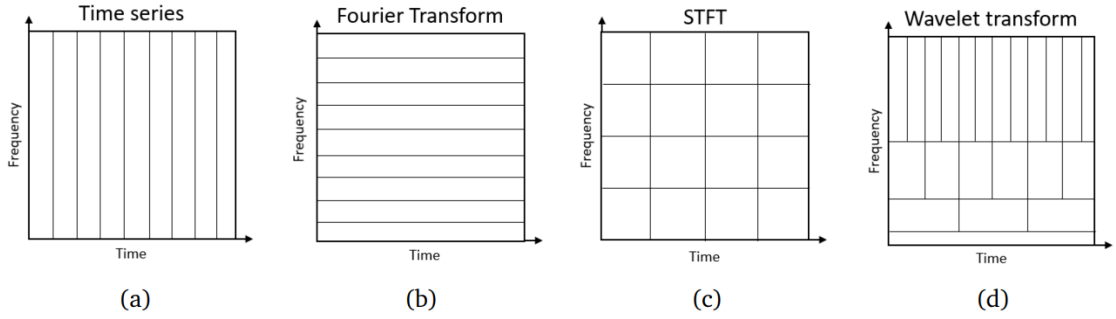


FIGURE 2.9: Difference in time and frequency resolution in time series (a), Fourier Transform (b), STFT (c) and a Wavelet Transform (d) (adapted from [22]).

Another technique that inherits and improves on the ideia of time localization is the Continuous Wavelet Transform (CWT), but unlike STFT, CWT is capable of providing high time resolution and low frequency resolution in the high frequencies, and high frequency resolution and low time resolution in the low frequencies by changing its parameters of scale and translation [22].

The CWT for the signal $f(t)$ is defined as the integration of the $f(t)$ with the shifted or scaled shapes from a mother wavelet $\psi_{a,b}(t)$:

$$CWT(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi^* \left(\frac{t-b}{a}\right) dt \tag{2.1}$$
$$a \in \mathbb{R}^+ \setminus \{0\}, b \in \mathbb{R}$$

The value a is a scale factor for scaling the function $\psi(t)$, while the b is a shift factor for translating the function $\psi(t)$. The CWT is the sum of the signal multiplied by the shifted and scaled shapes from a mother wavelet $\psi_{a,b}(t)$, resulting in a matrix with wavelet coefficients located by scale and position. Choosing the scale parameters and mother wavelet in CWT is very important for analyzing signals [23].

In wavelet analysis, the way to relate scale (a) to an aproximate frequency, usually referred as "pseudo frequency" (F_a), is to determine the center frequency of the wavelet, F_c, and use the following relationship [24]:

$$F_a = \frac{F_c}{a} \tag{2.2}$$

Figure 2.10 shows the graphical representation of several wavelet families.

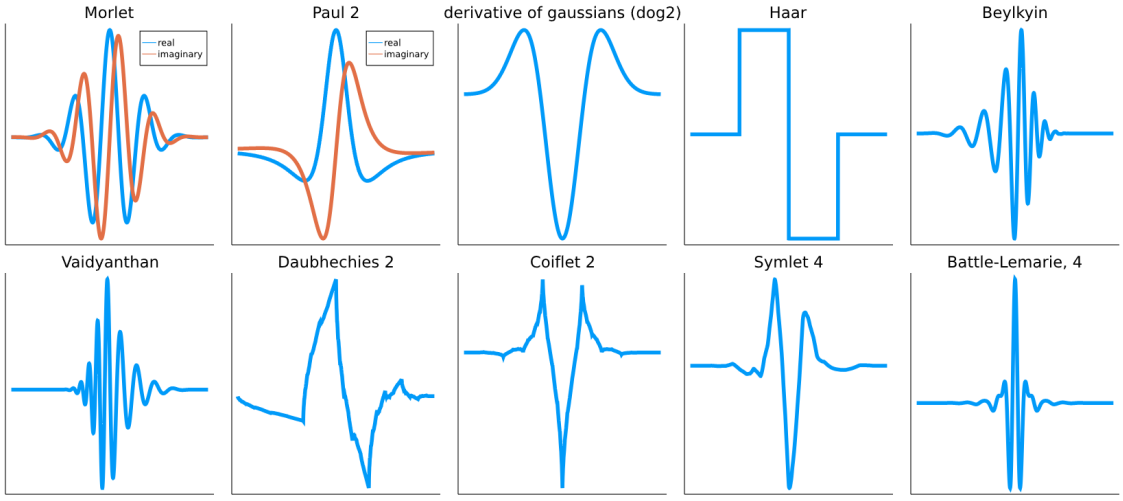


FIGURE 2.10: Graphical representation of commonly used wavelet families [25].

The scalogram is the absolute value of the CWT coefficients of a signal, plotted as a function of time and scale (or frequency) [21]. Figure 2.11 shows an example of a short segment of normal sinus rhythm ECG (a) and its scalogram (b) obtained by a CWT using the complex morlet.

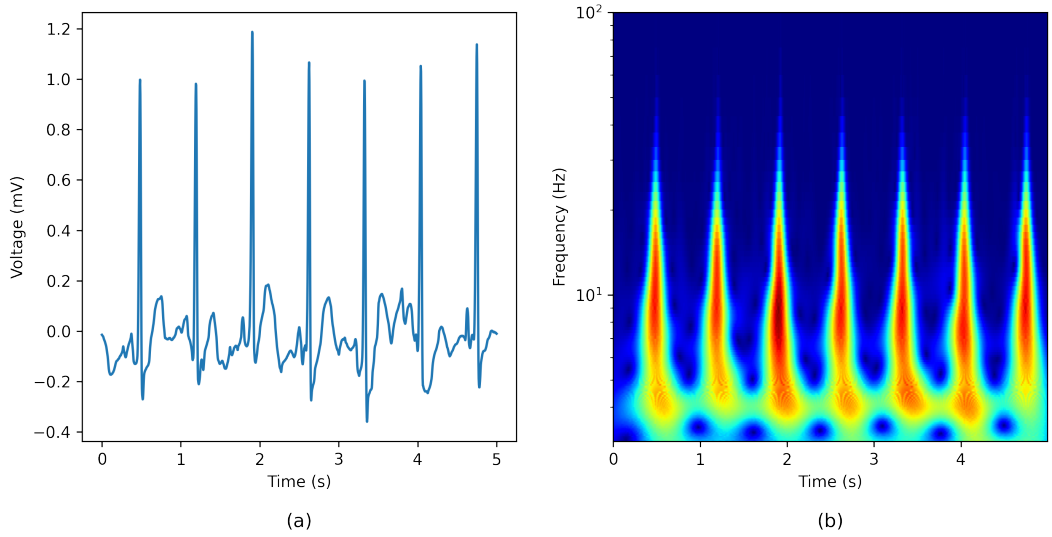


FIGURE 2.11: Short segment of a Normal ECG (a) and its CWT scalogram (b) obtained using the complex morlet wavelet.

Another popular choice for feature extraction of accoustic signals (such as the PCG) is the Mel Frequency Cepstral Coefficients (MFCC) which maps the signal onto a non-linear Mel-Scale that mimics the human hearing [26].

2.2.3 DL algorithms

Deep learning is a branch of ML that focuses on building deep neural networks to perform tasks in different fields such as object detection, speech recognition and autonomous vehicles [27].

Inspired by the human brain with its billions of connections, deep neural networks have multiple layers of interconnected units called neurons. Depending on the signals it receives as inputs, the neuron can be activated, producing another signal sent to another neuron. The set of input signals are propagated through the middle layers, called hidden layers, and then to the output layer.

The basic structure of a neuron is represented in Figure 2.12.

A set of input values are weighted and summed. This result is used as input for an activation function that determines how much the neuron will be activated by the received input signal. In Figure 2.12 $x_1, \ldots, x_n$ represents the input values and $w_1, \ldots, w_n$ represents the weights. An activation function is a mathematical function that determines whether a particular input should be activated or not. There are several types of activation functions that can be chosen for a given neural network, such as the Rectified Linear Unit

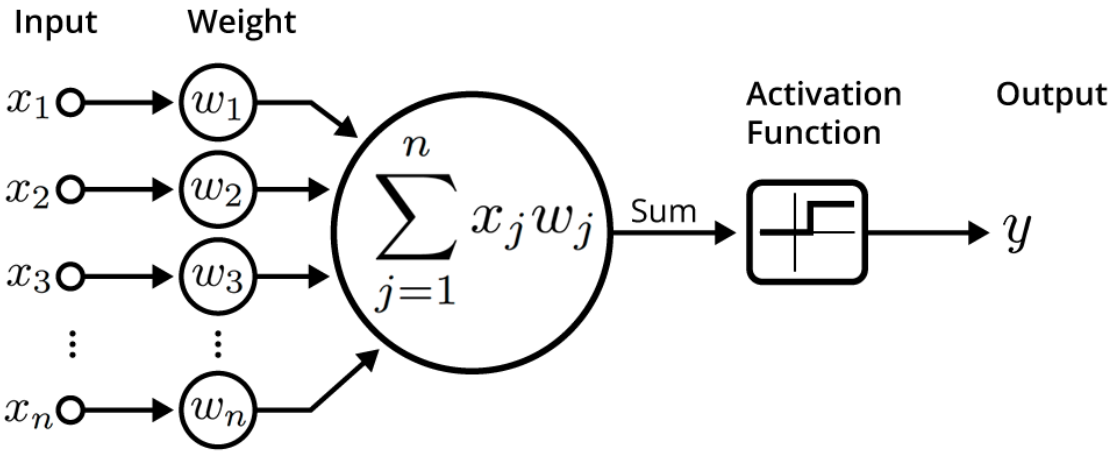


FIGURE 2.12: Basic neuron structure (adapted from [28]).

(ReLU), a a piecewise linear function that will output the input directly if it is positive, otherwise, it will output zero as shown in Figure 2.13.

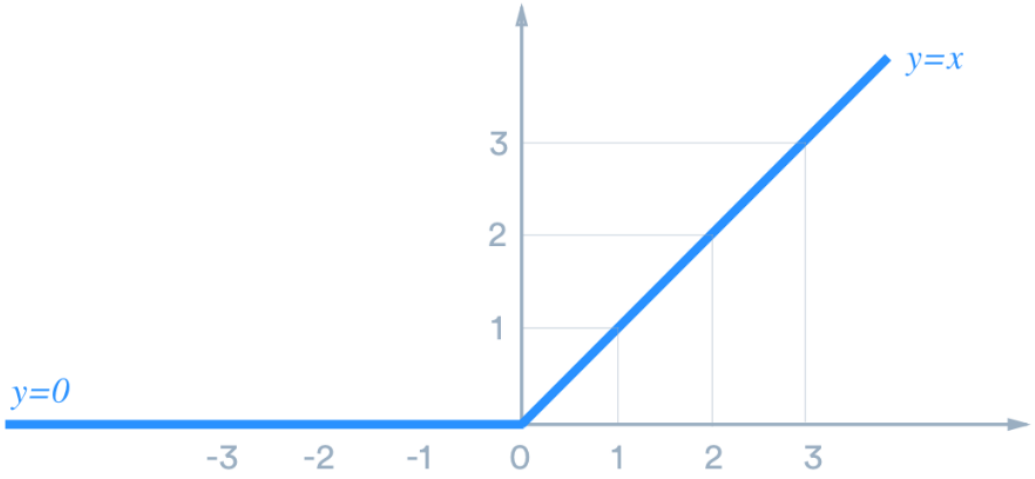


FIGURE 2.13: Graphical representation of the ReLU activation function (adapted from [29]).

DL algorithms use layers of neural-networks to convert raw input data to higher-level information. There are different kinds of DL algorithms such as Convolutional Neural Network (CNN), Multilayer Perceptron (MLP), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM) and so on.

A MLP is a network of artificial neurons with multiple hidden layers between input and output layers. These neurons usually create a complex network of different layers. Neurons from one layer pass signals to other neurons in the next layer. Figure 2.14 shows a schematic representation of a MLP.

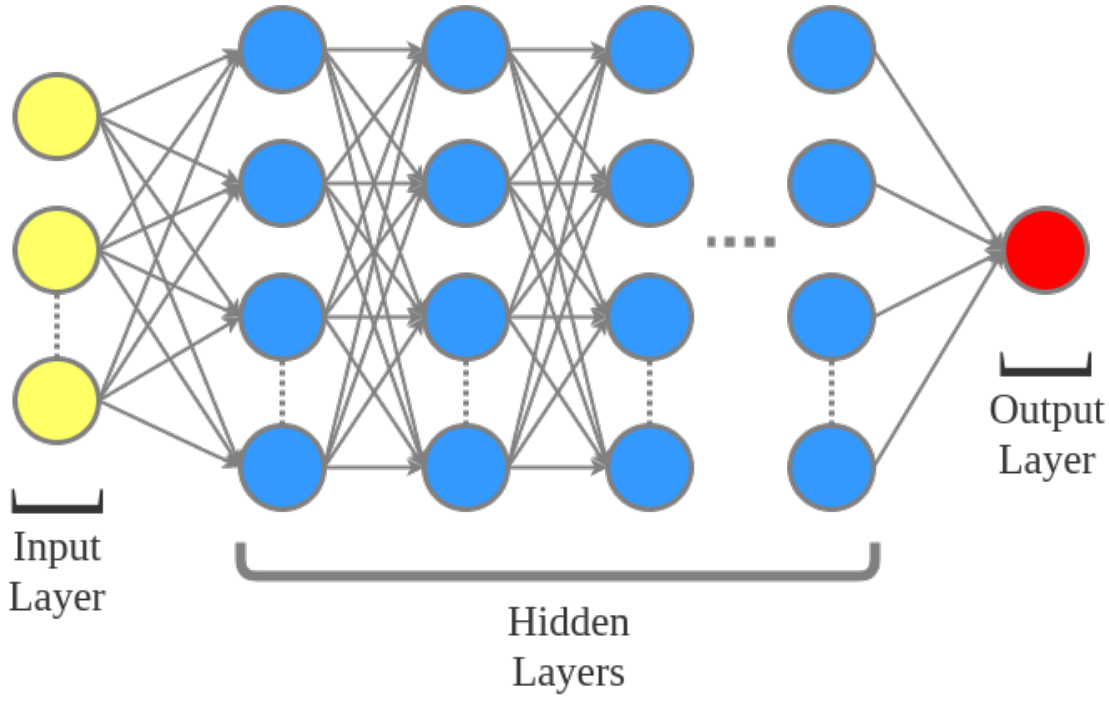


FIGURE 2.14: A MLP with multiple hidden layers [30].

The output of the input layer works as input to the first hidden layer. This process will continue until the final layer is reached. The output of the final layer will give the final prediction. Each layer can have one or more neurons and each neuron uses a threshold value in the form of an activation function to pass the signal to the next neuron. Two neurons of consecutive layers are connected with a parameter called weight. The function of the weight is to transform the input data within the hidden layers. The network parameters are modified to reduce the value of a function that expresses the difference between the actual output and the ground truth, called *loss* function. The loss function minimization is done via gradient-based optimization, where the gradient is computed using the backpropagation algorithm. The gradient descent is a widely used optimization method to update the weights by calculating the derivative of the error with respect to the weights of the network. While training the model, MLP uses a backpropagation algorithm to provide feedback to the network based on the output. The goal of the backpropagation algorithm is to update each of the weights several times step-by-step, thereby minimizing the error and gradually increase the overall performance [27].

The learning rate is a hyper-parameter, which determines the adjustment of the weights with respect to the loss gradient. The range of the learning rate is usually between 0 to 1.

This process of updating the weights will continue until the loss function is minimum or until a specific stopping criteria is verified.

While MLPs are usually applied over a set of features selected and extracted from the data, a CNN is a DL algorithm which uses a series of convolutions with different filters to automatically learn important features directly from the raw data. The convolution layer contains filters that pass over the data to capture the optimal features. For a 1-D signal $x_n = [x_1, x_2, \ldots, x_N]$, if it has K number of classes to classify and N is the signal length then initially a 1D convolution method is used to extract the optimal features from the raw input data by applying a series of 1-D convolutions with different 1-D kernels. This process is achieved by sliding a kernel $h(n)$ with length of W samples along the input data [11]. In this way, the ith output $y_i(n)$ from the Conv1D layer can be expressed by:

$$y_i(n) = \sum_{k=0}^{W-1} h_i(k)x(n-k)$$

If the the dimension of the feature is high, it can be reduced by using pooling function. Pooling can be min, max, or average pooling. Pooling layers merge semantically similar features found in the previous activation map and help control overfitting [31]. The most common layer is the max-pooling which extract the maximum, as shown in Figure 2.15 a max-pooling layer is applied with a 2×2 kernel size (filter dimensions) and a stride (how much the kernel is shifted during the convolution) of 2 .

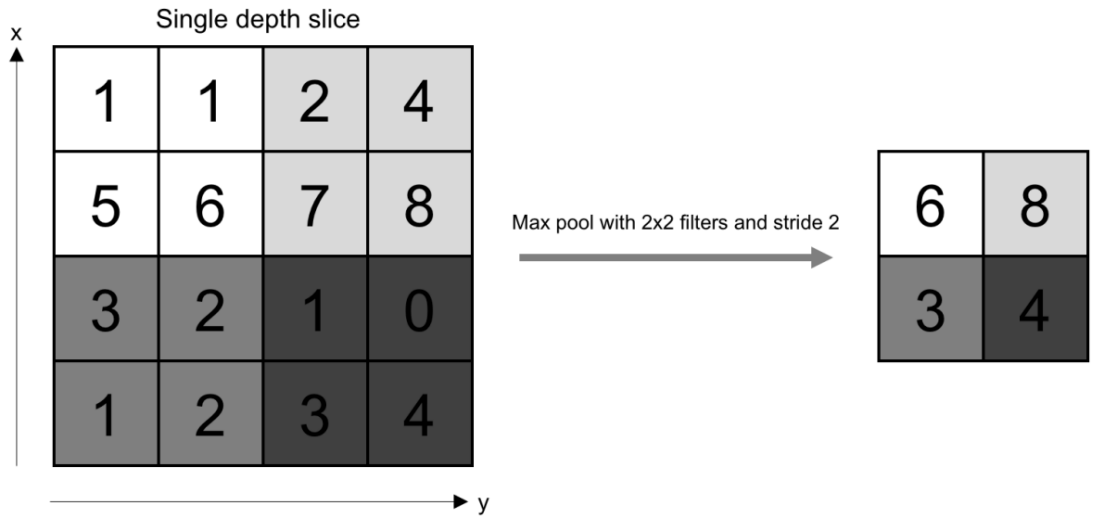


FIGURE 2.15: Example of a max-pooling operation [32].

Finally, the pooled features are passed into a fully connected layer for the final classification. Figure 2.16 represents schematically an example of a CNN architecture showing several convolutional layers to extract the features and a MLP to combine them and obtain the classification.

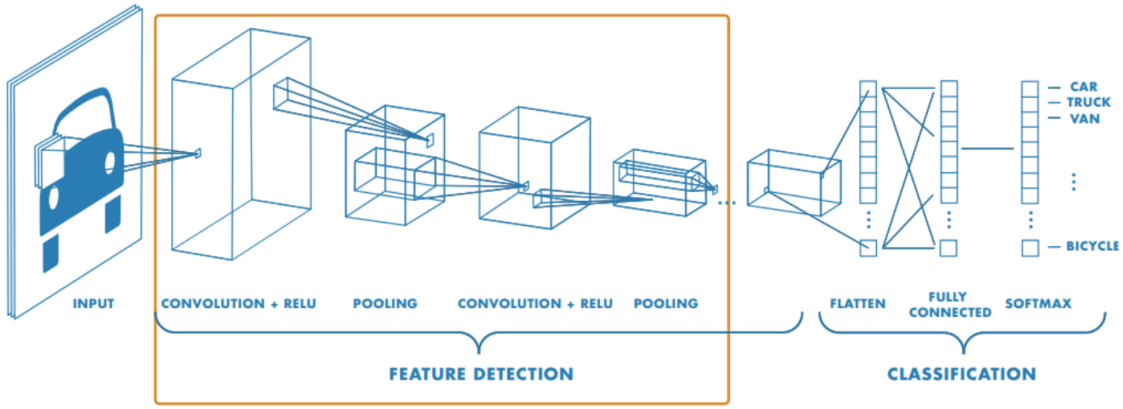
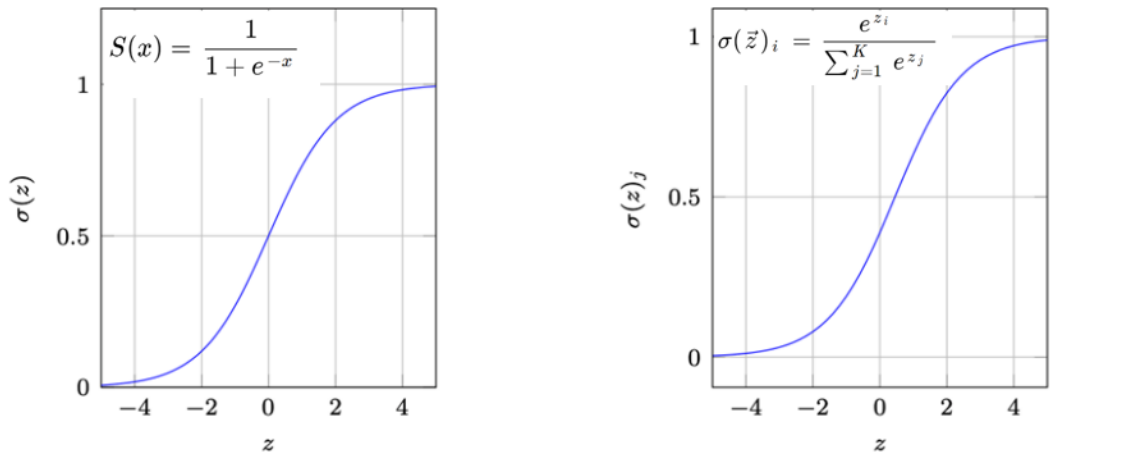


FIGURE 2.16: Convolutional neural network architecture that classifies input images as belonging to a number of categories including cars, trucks, vans and bicycles [18].

To obtain the estimated probability for each class, sigmoid or softmax activation function are typically used at the final output of the fully connected layer. If the classification is binary then Sigmoid activation function is generally used. If it is a multi-class classification then Softmax activation is typically used. Softmax and sigmoid activation functions are presented in Figure 2.17.



(a) Sigmoid activation function.

(b) Softmax activation function.

FIGURE 2.17: Graphical representation and equations of Sigmoid (left) and softmax (right) activation functions [33].

2.2.4 Transfer learning

Transfer learning is a machine learning technique in which a model created for a certain task is re-purposed on a second related task. It is particulary useful in DL, since most sophisticated models require access to vast amounts of data (not always available), time and computational power [34].

DL requires vast amounts of large labeled datasets, for example, ImageNet, one of the largest datasets for image classification (with over 14 million images) and very powerful computing resources to solve many challenging computer vision problems. Transfer learning presents an attractive proposition in solving real-world problems such as medical image recognition tasks, where there is a shortage of labeled datasets, as they usually require experts to label the data [35]. In this context, two transfer learning techniques have been widely applied for image recognition tasks: (1) Pretrained networks as a feature extractor and (2) fine-tuning a pretrained network. In principle, transfer learning translates knowledge that has already been learned in one domain (source) and applied to solve a new task in a different but related problem (target) [36].

The most common use of transfer learning in the context of deep learning consist in the following workflow:

1. Take layers from a previously trained model.
2. Freeze them, so as to avoid destroying any of the information they contain during future training rounds.
3. Add some new, trainable layers on top of the frozen layers.
4. Train the new layers on the new dataset.

A last, optional step, is fine-tuning, which consists of unfreezing the entire model (or part of it), and re-training it on the new data with a very low learning rate. This can potentially achieve improvements, by incrementally adapting the pretrained features to the new data [37].

2.2.5 Model performance evaluation

A confusion matrix is an $N \times N$ matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual

target values with those predicted by the machine learning model. Figure 2.18 shows a basic representation of confusion matrix for binary classification.

		Predicted	
		Negative **(N)** -	Positive **(P)** +
Actual	Negative -	True Negative **(TN)**	**False Positive (FP) Type I Error**
	Positive +	**False Negative (FN) Type II Error**	True Positive **(TP)**

FIGURE 2.18: Basic representation of a 2x2 confusion matrix for binary classification [38].

Data from the confusion matrix can be used to calculate several classification metrics such as sensitivity/recall, specificity and precision. This metrics can be used to assess and compare model's performances. The sensitivity/recall indicates the true positive rate and measures the proportion of the correctly identified actual positives. The specificity indicates the true negative rate and measures the proportion of the correctly identified actual negatives. Out of predictive positive, how many of them are actual positive is defined by the precision metric. All of these parameters can be calculated by using the confusion matrix. Another important metric used to evaluate the classification model is known as accuracy, which is the number of correctly predicted data points out of all the data points. Sensitivity, specificity, and accuracy can be calculated by using the following formulas:

$$Sensitivity/recall = \frac{TP}{TP + FN} \tag{2.3}$$

$$Specificity = \frac{TN}{FP + TN} \tag{2.4}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.6}$$

In equations (2.3)-(2.6), TP corresponds to the number of true positives, for example when diagnosing a disease, represents the number of sick people correctly identified as sick. TN correspond to the total of true negatives (the number of healthy people correctly identified as healthy, FP (false positives) is the number of people healthy incorrectly diagnosed as sick and FN (false negatives) is the number of unhealthy people incorrecly classified as healthy.

Moreover, precision and recall can both be taken into account through their harmonic mean, generally referred as the F1-score, calculated as follows:

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.7}$$

The receiver operating characteristics (ROC) curve is a graphical plot that shows the classification performance of a model as its decision threshold varies [39]. The ROC curve is created by plotting the true positive rate (TPR), on the y-axis, against the false positive rate (FPR), on the x-axis, at various threshold settings. The TPR is equal to sensitivity calculated by equation (2). The FPR can be calculate as $(1 - specificity)$.

ROC curves summarize the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds [40]. The area under the ROC curve (AUC) is a global measure of the classification performance, measure the model's ability to distiguish between the target classes. An AUC of 0.5 represents a test with no discriminating ability (no better than chance), while an AUC of 1.0 represents a test with perfect discrimination. Figure 2.19 shows a plot of ROC curves from different classifiers. The dashed line represents a random classifier.

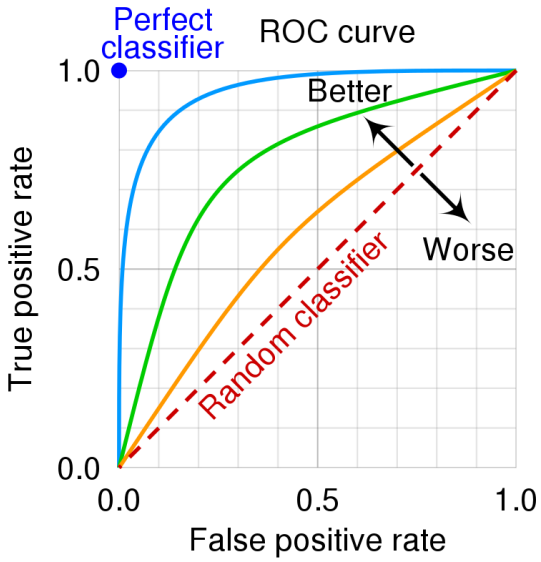


FIGURE 2.19: Representation of ROC curves from different classifiers [41].

Chapter 3

State of the art

In this chapter, a detailed characterization of the current studies that perform multimodal analysis of ECG and PCG signals is provided. A thorough literature systematic search and review was conducted with the objective of creating an extensive reference list of related research and identify current and future trends in the particular scope of this study.

The publicly available databases, that provide simultaneously acquired ECG and PCG and represent a valuable tool for researchers in this field, are also presented in this chapter.

3.1 Multimodal analysis of ECG and PCG

3.1.1 Search Methodology

Two online research databases were used in the search process: PubMed [42] and IEEEXplore [42]. Medical Subject Headings (MeSH) terms, a medical metadata nomenclature system, were taken into account in the search query design. The designed search query is a combination of 3 components, that aims to combine the different aspects of the intended search. MeSH terms are written in bold:

Component 1 - Related to ECG:

(**"electrocardiography"** OR "electrocardiogram" OR "ECG" OR "EKG")

Component 2 - Related to PCG:

(**"phonocardiography"** OR "phonocardiogram" OR "PCG" OR **"auscultation"** OR **"heart sounds"**)

Component 3 - Related to analysis/outcome:

(**"diagnosis"** OR **"classification''** OR **"machine learning"** OR **"deep learning"** OR "analysis" OR "multimodal")

The resulting complete search string is the following:

(**"electrocardiography"** OR "electrocardiogram" OR "ECG" OR "EKG") AND (**"phonocardiography"** OR "phonocardiogram" OR "PCG" OR **"auscultation"** OR **"heart sounds"**) AND (**"diagnosis"** OR **"classification''** OR **"machine learning"** OR **"deep learning"** OR "analysis" OR "multimodal")

A publication date range was applied to the search, selecting articles that were published from 2010 to 2022. The query generated a total of 439 results (269 from PubMed and 170 from IEEEXplore). From the 439 articles a total of 14 were duplicates, remaining therefore 425 unique articles.

3.2 Systematic Review Methodology

In this section, the systematic review process is described. The main goal of this set of steps, is to filter the publications that fit the eligibility criteria, i.e., studies in which simultaneously acquired PCG and ECG signals are analysed and features from both are used to potentially obtain clinically relevant information, such as a possible diagnosis or estimation of physiological parameters that can be used to further characterize the subject.

To provide an efficient systematic review process, the following sequential procedure was designed and applied to the 425 unique publications:

1. **Superficial selection** - Assessment of the articles elligibility by analysing titles and abstracts;

2. **Deep selection** - Analysis of the full text publications and applying the eligibility criteria to a more restrictive level;

3. **Information extraction and presentation** - Extraction and presentation of several important aspects related to the selected articles, like the objectives of the study and the results obtained.

From superficial selection, a total of 378 (89%) of the articles were excluded as they did not fulfill the inclusion criteria.

Deep selection was performed on the previously selected 47 articles, from which 35 did not meet the selection criteria and thus were excluded, therefore remaining 12. By reference list analysis, one other eligible article not present in the original extraction was

found, being then added to the set of selected articles, making a total of 13 final resulting articles from this phase.

Therefore a total of 426 articles (including 1 from reference lists) were analysed, 378 were excluded in the first step (remaining 48) and the second selection step excluded 35 publications, thus remaining 13. Figure 3.1 presents a flowchart of the selection process.

It is important to note that a big fraction of the total 412 excluded publications in both selection steps corresponded to articles that only analysed either ECG or PCG but not both (45%), clinical case reports and observational studies (10%), about the development of multimodal acquisition systems (5%), related to other physiological signals (4%), among other excluding reasons (36%).

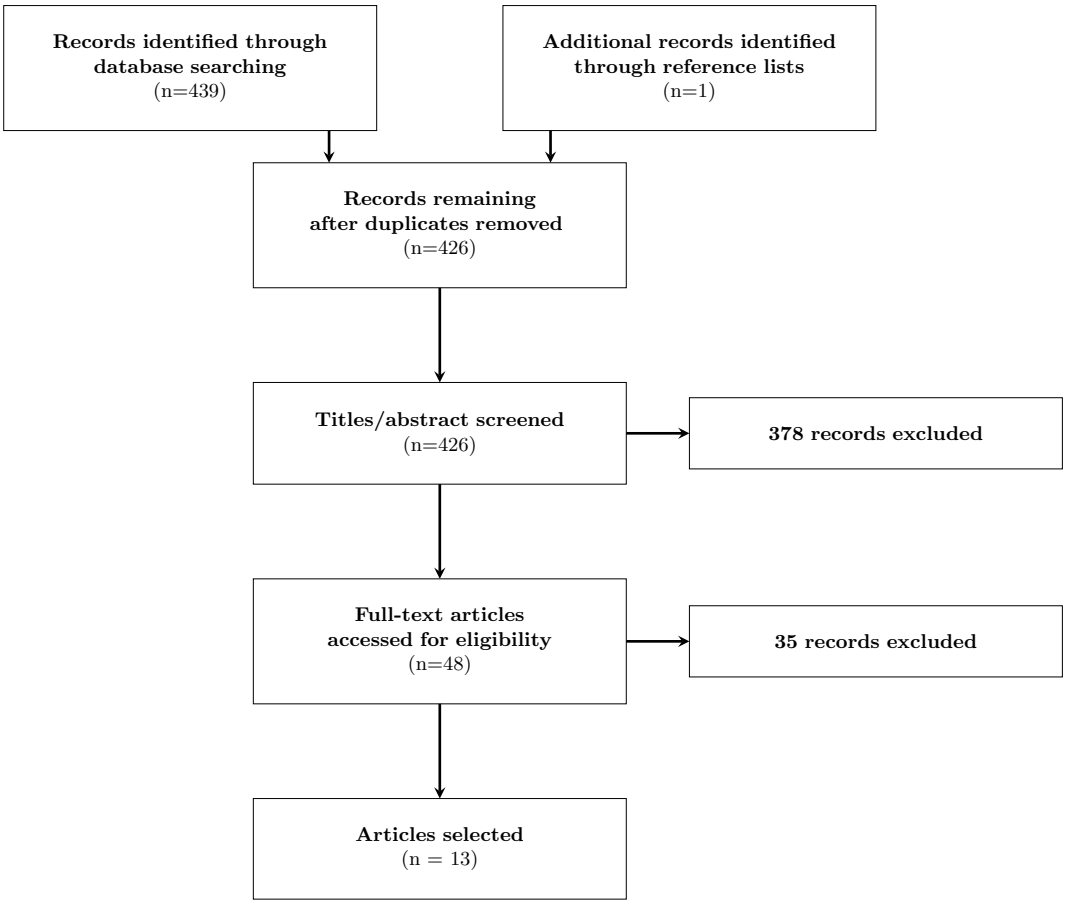


FIGURE 3.1: Flowchart of the systematic review process

3.2.1 Results

Information extracted from the selected articles is presented in Table 3.2.1.

Hettiarachchi et al. [43] introduced a novel dual-convolutional neural network (CNN) based approach using transfer learning to tackle the problem of having limited amounts of simultaneous PCG and ECG data that is publicly available, while having the potential to adapt to larger datasets. The hypothesis was tested using data from a publicly available database, PhysioNet 2016 Challenge (training set A) further described in section 3.3. Comparisons with methods which used single or dual modality data, namely making the comparison with the dual modality SVM model used by Chakir et al. [44] (later described in this chapter), show that the developed method can lead to better performance. Furthermore, results show that individually collected ECG or PCG waveforms are able to provide transferable features which could effectively help to make use of a limited number of synchronized PCG and ECG waveforms and still achieve significant classification performance.

EL-Bouridy and EL-Batouty [45] used integrated cardiograph (ICG) scanned images composed of 12-lead ECG and 5-probe PCG for each recorded subject. The proposed algorithm consists of three fundamental steps. The initial step is image reading and denoising using a 2-D adaptive noise-removal filtering. The second step is the pre-extracion, that is divided into two channels, the statistical channel, and the decompositional channel. Another set of features designates "post extracted" (such as the standard deviation) are obtained from the pre-extracted features. The third step is an ANN classifier that consists of three layers using feed forward back propagation computation with momentum term the activation function between the input layer and the hidden layer is log-sigmoid, and the activation function between the hidden layer and the output layer is linear. The best obtained results regarding classification accuracy were 96.82 %. Information regarding the size and source of the dataset was not provided in the publication.

In Balbin et al. [46] simultaneous PCG and ECG signals were acquired from 20 subjects by an hardware apparatus assembled by the authors. The ECG recording is classified by a CNN model specifically tuned for ECG. The PCG is transformed into Mel Frequency Cepstral Coefficients before being classified by a separate CNN model specifically tuned for PCG. Given that both models were trained separately, each model outputs its own classification. The researchers have created a summary table to unify the classification from both models, producing 3 classes: normal, abnormal and noisy. A sensitivity of 100% was obtained, while specificity was 77.78%, showing an accuracy of 80%.

Zhang et al. [47] proposed an effective method for mining cardiac mechano-electric

coupling information and to evaluate its ability to distinguish patients with varying degrees of coronary artery stenosis (VDCAS). Five minutes of electrocardiogram and phonocardiogram signals collected synchronously from 191 VDCAS patients were used to construct several interval time series (e.g. heartbeat interval –systolic time interval (STI)). Several features were computed, namely the cross sample entropy (XSampEn), cross fuzzy entropy (XFuzzyEn), joint distribution entropy (JDistEn). Subsequently, SVM recursive feature elimination and XGBoost were utilized for feature selection and classification, respectively. Results showed that the joint analysis of XSampEn, XFuzzyEn, and JDistEn had the best ability to distinguish patients with VDCAS and can effectively capture the cardiac mechano-electric coupling information of patients with VDCAS, which can provide valuable information for clinicians to diagnose coronary heart disease (CHD). The classification accuracy of distinguishing between severe CHD and mild-to-moderate CHD groups, severe CHD and chest pain and chest pain and normal coronary angiography (CPNCA) groups, and mild-to-moderate CHD and CPNCA groups were 80.43%, 76.59%, and 75.00%, respectively.

H. Li et al. [48] wrote a paper that aims to differentiate between Coronary Artery Disease (CAD) and non-CAD groups. A novel dual-input neural network that integrates the feature extraction and deep learning methods is developed. First, the ECG and PCG features are extracted from multiple domains, and the information gain ratio is used to select important features. On the other hand, the ECG signal and the decomposed PCG signal (at four scales) are concatenated as a five-channel signal. Then, the selected features and the five-channel signal are fed into the proposed network composed of a fully connected model and a deep learning model. The results show that the classification performance of either feature extraction or deep learning is insufficient when using only ECG or PCG signal, and combining the two signals improves the performance. Further, when using the proposed network, the best result is obtained with accuracy, sensitivity, specificity, and G-mean of 95.62%, 98.48%, 89.17%, and 93.69%, respectively.

X. Li et al. [49] did a study whose purpose was to determine the potential of synchronized analysis of PCG and ECG in identifying patients with depressed left ventricular ejection fraction (dLVEF). A total of 189 patients (76 with dLVEF; 113 with normal ejection fraction) were enrolled. All were admitted to the hospital because of dyspnea or chest discomfort. PCG and ECG signals were automatically analyzed using wavelet analysis and uti- lized to determine electromechanical activation time (EMAT), EMAT/RR, S1-S2 time,

and S1-S2/RR. EMAT in the dLVEF group was significantly higher than that in the control group. ROC curve analysis allowed to determine the optimal EMAT cutoff point of 104 ms (records were classified dLVEF when EMAT was equal or bigger than the cutoff point), having a sensitivity and specificity for the diagnosis of dLVEF of 92.1% and 92%, respectively. The authors concluded that the PCG and ECG signal index EMAT contributes to the diagnosis of dLVEF.

Singh et al. [50] developed a method for heart abnormality classification using PCG and ECG. Both signals were initially preprocessed with the purpose of noise removal. Wavelet decomposition, Hilbert transforms, Homomorphic filtering and power spectral distributions were used to extract time-frequency features from the PCG signal. The Pan-Tompkins algorithm was used to extract QRS based features. The extracted features from PCG and ECG signals were independently trained and tested using different classifiers (SVM, KNN, and Ensemble) and compared with the merged features of both the PCG and ECG signals. The proposed model was validated using publicly available dataset 'A' of PhysioNet 2016/ CinC challenge. The results show that ECG and PCG signals can efficiently be employed for predicting cardiovascular disorders.

Yupapin et al. [51] developed a system for the detection of preliminary heart defects composed by two subsystems. A set of 80 synchronous ECG and PCG signals were used. The relationship between both signals is determined as an impulse response of a system, where the decision is made based on the linear predictive coding coefficients of a heart's impulse response. The decision is made by the back propagation neural network from the impulse response signal, the accuracy obtained was 90% and 85% for NPVs and PPVs, respectively. The other subsystem is based on phase space of the signal (ECG or PCG). The MSE value obtained by comparing the distance vector of the testing signal with the reference distance vector is judged by the likelihood ratio test result. This technique provided 100% accuracy for decision making. The results from both techniques show that the impulse response-based method can be used primarily to detect a heart abnormality, whereas the phase space-based approach can be used to indicate whether the heart defect is caused from the abnormal ECG signal and/or abnormal PCG signal.

H. Li et al. [52] also wrote a paper that describes the development of a multi-input CNN framework that integrates time, frequency, and time-frequency domain deep features of ECG and PCG for CAD detection. Simultaneously recorded ECG and PCG signals from 195 subjects are used. The proposed framework consists of 1-D and 2-D CNN

models and uses signals, spectrum images, and time-frequency images of ECG and PCG as inputs. The framework combining multi-domain deep features of two-modal signals is very effective in classifying non-CAD and CAD subjects, achieving an accuracy, sensitivity, and specificity of 96.51%, 99.37%, and 90.08%, respectively. The comparison with existing studies demonstrates that the proposed method is very competitive in CAD detection. The proposed approach is very promising in assisting the real-world CAD diagnosis, especially under general medical conditions.

In the paper by Chakir et al. [44], a subsample of 100 multimodal records from the Physionet 2016 challenge was used to make a comparison between the performance of the PCG-based features and that of the features extracted from synchronous PCG and ECG recordings is presented. For that, the ROC curve and values of accuracy, AUC, sensitivity, and specificity are compared to select the best classifier for each feature combination and then to select the more pertinent biomarkers from these two resulting classification models. This paper demonstrates that a merging of ECG and PCG leads to higher performance of heart condition assessment than the diagnosis based on PCG recordings alone (Accuracy: 92.5% vs. 82.5%, AUC: 95.05% vs. 90.66%, sensitivity: 92.31% vs. 76.92% and specificity: 92.86% for both).

Y. Li et al. [53], perfomed a study about the application of cardiac electromechanical delay variability (EMDV) analysis to the detection of coronary heart disease. The authors extracted the beat-to-beat EMD from 5-min simultaneously recorded electrocardiogram and phonocardiogram signals in 30 patients with coronary artery disease and 30 healthy control subjects, and studied its variability using the same methods as applied for HRV. An SVM with 10-fold cross-validation was used for classification. The results suggest that the EMDV analysis that could potentially be helpful for detecting CAD noninvasively.

Klum et al. [54], performed multimodal analysis (using PCG and ECG) to estimate respiratory rate, pre-ejection period (PEP) and left ventricular ejection time (LVET) Respiratory rates were estimated using a time-delay neural network having as input several PCG and ECG features with mean absolute errors below 1.2 bpm, and the respiratory signal yielded a correlation of 0.66. Regarding the PEP and LVET estimations, the Pan-Tompkins algorithm was used to detect the R-peak in the ECG signal. P, Q, S and T waves were identified using successive rule based windowing and local maximum and minimum detection. The start, end, and peak of S1 and S2 within the stethoscope PCG signals

were detected using a modified approach of the empirical wavelet transform and instantaneous phase based method, combined with an ECG based PCG peak classification step. The PEP was estimated with a mean error (ME) of 0.4 ms and an mean absolute error (MAE) of 25.1 ms, which translated to 21.3% relative error. The LVET was estimated with an ME of -3.6 ms and an MAE of 30.5 ms, which was a relative error of 10.0%.

Zeng Y. et al.[6] proposed a parallel multimodal method for left ventricular dysfunction (LVD) identification based on synchronous analysis of PCG and ECG signals. The database used consisted of 1046 synchronous ECG and PCG recordings from patients with 173 LVD and 873 normal patients. Signals were preprocessed and spectrograms were obtained using short-time Fourier transform. Two-layer bidirectional gate recurrent unit was used to extract features in the time domain, and the data were classified using residual network 18. This research confirmed that fused ECG and PCG signals yielded better performance than ECG or PCG signals alone, with an accuracy of 93.27%, precision of 93.34%, sensitivity of 93.27%, and F1-score of 93.27%.

Table 3.1: Summary of the analysed publications.

Authors (year of publication)	Focus	Database	Main Model/Method	Results
Hettiarachchi et al. (2021) [43]	Binary classification (normal or abnormal)	Physionet 2016 challenge dataset	Transfer Learning and CNN	With transfer learning: Sen = 87.72%, Spe = 87.5%, Acc = 87.67%, AUC = 93.75%, G-mean = 87.6%

Continued on next page

Table 3.1: Summary of the analysed publications. (Continued)

Authors (year of publication)	Focus	Database	Main Model/Method	Results
EL-Bouridy and EL-Batouty (2021) [45]	Binary classification (normal or abnormal)	ICG scanned images composed of 12-lead ECG and 5-probe PCG (size and origin of dataset not specified)	ANN	Acc up to 96.82%
Balbin et al. (2021) [46]	Classification (normal, abnormal, noisy)	Data gathered from 20 subjects	CNN	Sen = 100%, Spe = 77.78% and Acc = 80%
Zhang et al. (2021) [47]	Classification: distinguish patients with VDCAS (severe vs. mild-to moderate CHD, severe CHD vs CPNCA, mild-to-moderate CHD vs. CPNCA)	5 min of signals collected synchronously from 191 VDCAS patients	XGBoost	Acc in distinguishing each group bigger than 75%

Continued on next page

Table 3.1: Summary of the analysed publications. (Continued)

Authors (year of publication)	Focus	Database	Main Model/Method	Results
H. Li et al. (2019) [48]	Classification: differentiate between Coronary Artery Disease (CAD) and non-CAD groups	Data acquired from 195 subjects	ANN	Acc = 95.62%, Sem = 98.48%, Spe = 89.17% and G-mean = 93.69%
X. Li et al. (2020) [49]	Identifying patients with depressed left ventricular ejection fraction (dLVEF)	A total of 189 patients (76 with dLVEF; 113 with normal ejection fraction)	ROC curve analysis	EMAT in the dLVEF group was significantly higher than that in the control group
Singh et al. (2021) [50]	Binary classification (normal or abnormal)	Physionet 2016 challenge dataset	SVM (proposed method), KNN and Ensemble	SVM: Sen = 94.00%, Spe = 90.00% and Acc = 93.13%
Yupapin et al. (2011) [51]	Binary classification (normal or abnormal)	80 records	Back propagation ANN	PPV = 90%, NPV = 85%
H. Li et al. (2021) [52]	Classification (CAD and non-CAD)	Data acquired from 195 subjects	CNN	Acc = 96.51%, Sen = 99.37%, Spe = 90.08%

Continued on next page

Table 3.1: Summary of the analysed publications. (Continued)

Authors (year of publication)	Focus	Database	Main Model/Method	Results
Chakir et al. (2020) [44]	Binary classification (normal or abnormal)	Subsample of 100 records from Physionet 2016 challenge	SVM was the proposed model	Acc = 92.5%, AUC = 95.05%, Sen = 92.31%, Spe = 92.86%
Y. Li et al. (2019) [53]	Study the impact of adding EMDV to the detection of CAD	60 records (30 from CAD subjects and 30 from healthy control group)	SVM	Adding EMDV increased the classification accuracy from 72.9% to 95.8%
Klum et al. (2020) [54]	Estimate respiratory rate, LVET and PEP	Data from 10 healthy subjects	Time-delay NN for the Respiratory rate, Pan-Tompkins and wavelet transform for PEP and LVET	Respiratory rate MAE below 1.2 bpm, signal correlation of 0.66. LVET and PEP estimation errors were 10% and 21%, respectively.
Zeng et al. (2022) [6]	LVD diagnosis	Records from 1046 subjects (173 with LVD and 873 healthy)	Residual Network	Acc = 93.27%, Sen = 93.27%, and F1-score of 93.27%.

From the selected set it can be concluded that most of the articles are recent (54% were published since 2021), having only one article published prior to 2018. This may indicate a growing interest in the particular topic of this review. Further characterizing the selected

set, publications can be divided in two groups: (a) articles that focus on classification (mostly binary) and (b) articles that aim to estimate a certain potentially relevant cardiac parameter and evaluate its importance.

A significant fraction of the articles presented (8 out of 12) use deep learning approaches. Most of these studies acknowledge that one of the limitations to their studies is the lack of data, namely publicly available datasets containing multimodal data. Transfer learning, as performed by Hettiarachchi et al. [43], can be further explored and can lead to interesting and even more promising results since databases that contain only ECG or only PCG are much more abundant.

Moreover, articles that feature NN either present a model that combines the signals/features in the first stages (early fusion) or a model that has basically separate pipelines for each signal, combining only the resulting features in the last stages of the network (late fusion). A comparison between both these approaches is yet to be made for this particular problem in order to possibly identify the most effective method.

3.2.2 Other considerations

It is worth mentioning that during the review process, several publications were analysed in which simultaneously captured PCG and ECG signals are used, but the resulting publication is not within the particular scope of this study. Namely, several articles in which PCG segmentation is aided by ECG's reference points identification and no further multimodal analysis is performed [55–58].

3.3 Databases

One of the many challenges in performing studies regarding ECG and PCG multimodal analysis is the relatively low volume of publicly available data when compared to individual signals datasets (only PCG or only ECG) [43].

Multimodal analysis of simultaneously recorded ECG and PCG can provide interesting insights into the inter-relationship between the mechanical and electrical mechanisms of the heart, combining advantages from both signals, potentially improving disease diagnosis and estimation of certain features/parameters. Moreover, synchronized data can be used also for other applications, namely:

- Testing the performance of synthetic signal generators that attempt to simulate a signal having the other as input, for example generating artificial ECG from real PCG records as done by McSharry et al. [59] ;
- Use one of the signals as a gold standard reference to train or evaluate the performance of a single input model, such as QT interval estimation from PCG with ECG determined QT interval as gold standard, studied by Sbrollini et al. [60].

Table 3.2 contains a summary of 3 publicly available databases that will be further described in the following subsections.

TABLE 3.2: Summary of publicly available databases that contain simultaneously recorded ECG and PCG.

Database Name	Number of Subjects	Number of Records	Record Length	Number of Healthy Records	Number of Pathological Records
PhysioNet 2016 Challenge (training set A) [61]	Unknown	405	Variable duration (9 - 37 s)	117	288
EPHNOGRAM [62]	24	69	0.5 or 30 min	69	0
GUARDIAN Vital Sign Data [63]	11	259	approx. 60 s	259	0

3.3.1 PhysioNet 2016 Computing in Cardiology Challenge Dataset

PhysioNet is a very important web-based resource supplying well-characterized physiologic signals and related open-source tools to the biomedical research community. It provides a public service of the Research Resource for Complex Physiologic Signals, a cooperative project started by researchers at Boston's Beth Israel Deaconess Medical Center/Harvard Medical School, Boston University, MIT and McGill University [64].

PhysioNet promotes challenges, that invite participants to deal with clinically challenging questions that are currently either unsolved or not very well-solved. PhysioNet has been co-hosting a challenge annually, collaborating with the Computing in Cardiology (CinC) conference [65].

In 2016, the PhysioNet/CinC challenge focused on the classification of heart sound recordings [61]. This challenge aimed the development of algorithms to classify PCG

recordings, to determine from a single short recording, whether the subject should be referred for an expert diagnosis.

The challenge provides one of the largest public collection of PCG recordings, collected from several contributors around the world from both healthy and pathological subjects, including children and adults, allowing participants and researchers to potentially develop accurate and robust algorithms . It is unknown exactly how many records each subject contributed, the number of contributions may range between one and six PCG recordings per subject. The recordings have been sampled at 2000 Hz and each recording contains one PCG signal provided as .wav format [61].

The training set is divided in five databases (A through E) corresponding to a total of 3126 PCG recording of varying length (from 5 seconds up to 120 seconds) [61]. Although the challenge is focused exclusively in classification having PCG as the only input, most of the records in training dataset A contain also an ECG lead. Among the 409 records present in training set A, a total of 405 (99.02%) records present a simultaneously captured ECG lead with .dat format and sampled at 2000 Hz.

On this particular multimodal subset, there are 288 pathologic records, including variety of illnesses, namely, heart valve defects and coronary artery disease, and the remaining 117 are classified as normal [61, 66]. Signal duration ranges from 9.625 seconds to 36.502 seconds, with an average duration of 32.56 seconds. The total length is approximately 13187 seconds (220 minutes). Figure 3.2 presents a histogram of the time length distribution of this subset.

3.3.2 EPHNOGRAM

The electro-phono-cardiogram (EPHNOGRAM) project focused on the development of low-cost and low-power devices for recording simultaneous ECG and PCG data, with auxiliary channels for capturing environmental audio noise. The current database, recorded by version 2.1 of the developed hardware, has been acquired from 24 healthy adults aged between 23 and 29 (average: 25.4 ± 1.9 years) in 30min stress-test sessions during several states (resting, walking, running and biking conditions, using indoor fitness center equipment). The dataset also contains several 30s sample records acquired during rest conditions [62].

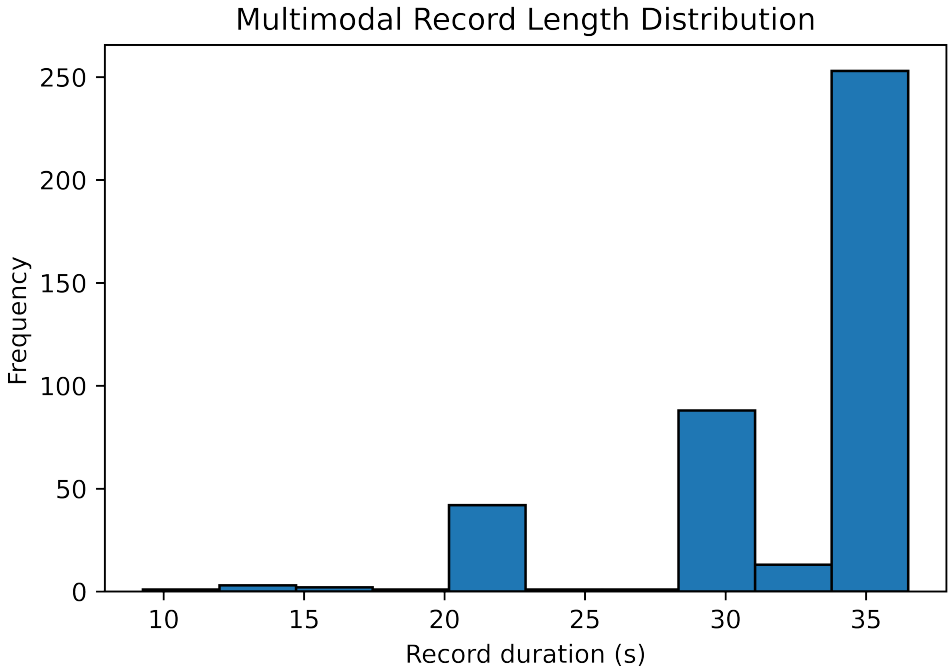


FIGURE 3.2: Multimodal records duration histogram for the Physionet dataset.

The dataset consists of 69 simultaneous ECG and PCG recordings, each with a duration of 30 seconds (8 records) and 30 minutes (61 records), acquired synchronously from a three-lead ECG and a single PCG stethoscope.

The 30 minutes records of the dataset were acquired in an indoor sports center. A structured interview determined that the participants were in good physical condition and none reported symptoms of cardiovascular disorder.

The resulting dataset is also available through PhysioNet [62, 64].

3.3.3 GUARDIAN Vital Sign Data

The GUARDIAN Vital Sign Data is a dataset designed for the validation of the feasibility of radar-based heart sound detection and the algorithm's functionality provided by simultaneously evaluating the ECG and PCG reference data [63, 67].

This dataset consists of synchronised data which are acquired using a Six-Port-based radar system operating at 24 GHz, a digital stethoscope, an ECG, and a respiration sensor. A total of 11 test subjects were measured in different defined scenarios and at several measurement positions such as at the carotid, the back, and several frontal positions on the thorax. It was intended for each record to have 60 seconds. A total of 259 records were produced, whose duration is slighty less than 60 seconds (mainly due to synchronization)

[63]. Around 223 minutes of data were acquired at scenarios such as breath-holding, post-exercise measurements, and while speaking.

Although the main objective in the acquisition of this dataset was to assess the feasibility of radar-based heart sound detection, since reference ECG and PCG recordings are provided, this dataset is also suitable for multimodal ECG and PCG analysis.

The dataset also provides an overview file, in which the recordings of each person are described. Noted there are the exact times of the recordings, which also serve as unique file identifers in the file names, subjective ratings of the signal qualities of the diferent sensor signals of a recording, and the exact positions and scenarios of the measurements.

The whole dataset is freely available at figshare [68]. All the records are stored in .mat (MATLAB) format.

Chapter 4

Methodology

In this chapter, a detailed description of the methods and techniques explored and implemented is provided. Implementation aspects such as data selection, preprocessing, feature extraction and model design are presented.

Due to the low amount of multimodal PCG and ECG data, a transfer learning approach is explored to leverage knowledge obtained from related domains to improve performance in the target task. This consists in the detection of abnormalities in the multimodal records leading to a binary classification in which the positive class (1) corresponds to the presence of abnormalities and the negative class (0) is attributed to normal (healthy) subjects. Since individual signal databases, of either PCG or ECG, have a much higher data volume than the available multimodal datasets, it is possible to extract knowledge from their interpretation and transfer it to the specific multimodal problem, as describer further in this chapter.

A data pipeline that allows signal pre-processing and the generation of scalograms is developed and explained in this chapter. Scalograms are used, since they have proven to be effective in the detection and classification of cardiac signals both individually and in a multimodal fashion [43, 50, 69–72].

As signals in the available datasets usually have variable lengths, a fixed length window of 5 seconds is used with no overlapping portions for length regularization and data augmentation [72]. Record portions with less than 5 seconds are discarded. The 5 seconds signal samples are extracted consecutively, not segmented taking into account specific signal components, such as the R-wave for the ECG, pre-processed, transformed by CWT and the resulting scalogram is exported as an image file. The scalograms are then used

as input in the model development, including model training and evaluation. Figure 4.1 illustrates the described process.

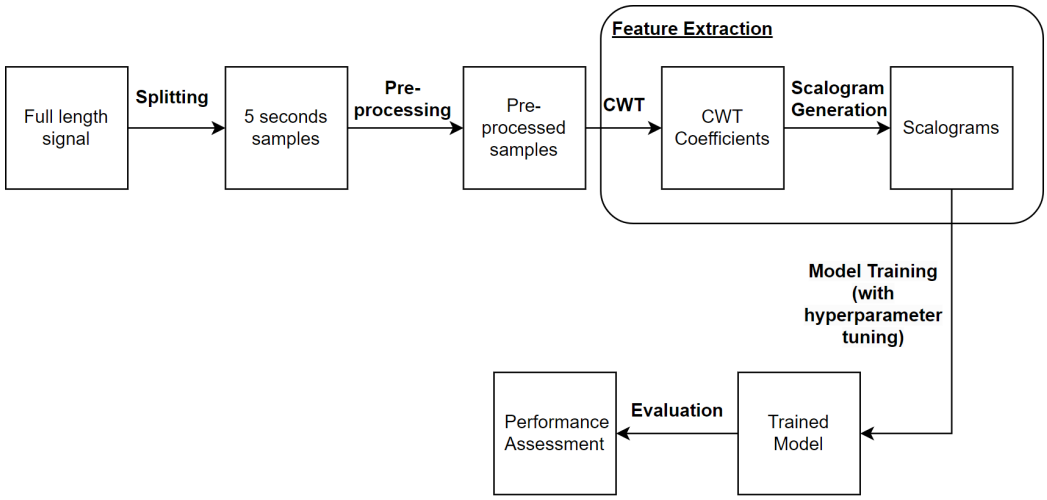


FIGURE 4.1: Graphical representation of the machine learning pipeline.

4.1 Datasets

In this section, the datasets that form the foundation of the learning process are described, including the multimodal dataset and the individual datasets used for transfer learning.

4.1.1 Multimodal Dataset

Among the databases presented in Section 3.3, the PhysioNet 2016 Computing in Cardiology challenge dataset (training-set A) is the only one that contains records classified as abnormal. Since the main focus of this work is to study methods to classify multimodal PCG and ECG records, only the database from Physionet will be used. The other two available multimodal datasets contain only records acquired from healthy subjects, thus hindering the ability to conduct clinically relevant tests regarding the capabilities of the proposed models in detecting abnormalities. Combining the three datasets could influence the model's behavior by adding heterogeneity due to differences such as the sensors used, signal resolution, noise, clinical signal capture procedure, frequency of sampling and different pre-processing techniques [73].

Moreover, the selected dataset with respect to the target class contains a portion of 71.1 % abnormal records and the remaining 28.9 % correspond to normal records.

Signal information regarding the binary classification is extracted from the header file (.hea format) for each record. PCG and ECG signals are extracted from .wav and .dat files. Standardization is applied on the input signals in order to rescale them amplitude wise.

More information regarding this particular dataset is presented in Section 3.3.1.

4.1.2 ECG

The 2017 PhysioNet/CinC Challenge aims to encourage the classification of short ECG single lead recordings in 4 classes: normal sinus rhythm, atrial fibrillation (AF), abnormal alternative rhythm (designated as "other rythms"), or too noisy to be classified [74]. Including a total of 8528 single lead ECG recordings with durations lasting from 9 s to just over 60 s is one of the largest ECG databases publicly available. Table 4.1 shows the data profile for the original dataset categories.

The provided ECG recordings are sampled at 300 Hz and already band pass filtered by the AliveCor device, having a frequency range of 0.5-40 Hz [74]. All data are provided in MATLAB V4 WFDB-compliant format (each including a .mat file containing the ECG and a .hea file containing the waveform information).

Figure 4.2 shows a plot of example waveforms for each of the original dataset labels [74, 75].

TABLE 4.1: Original data profile for the PhysioNet 2017 training set [74].

Type	Number of Recordings	Mean Time Length (s)
Normal	5154	31.9
AF	771	31.6
Other Rhythm	2557	34.1
Noisy	46	27.1
Total	8528	32.5

To match the binary classes present in the multimodal dataset the AF and other rythms are set to be abnormal. Records classified as too noisy are removed since they do not fit either target binary classes. Resulting in a total of 8482 recordings, of which, 5154 are normal (60.8%) and 3328 are abnormal (39.2%).

4.1.3 PCG

Besided the training-set A, the Physionet 2016 Computing in Cardiology challenge dataset provides 5 other training sets that contain PCG only. This subset of the Physionet dataset

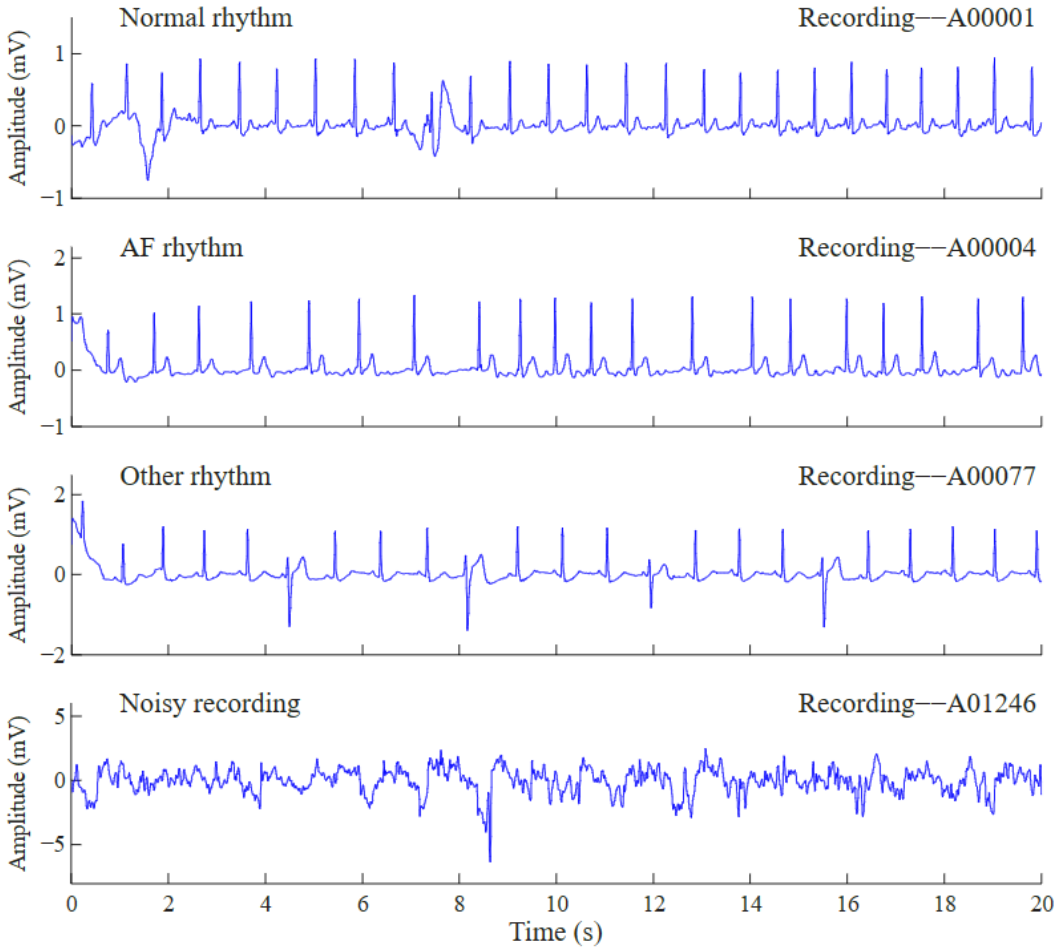


FIGURE 4.2: Examples of ECG waveforms for the categories provided in the Physionet 2017 challenge [76].

includes a total of 2831 records, of which 373 are abnormal (13.2%) and 2458 are normal (86.8%) [61, 75], as presented in Table 4.2.

Like the training-set A, the sampling frequency is 2000 Hz, the target class presented is binary and the file structure is identical.

As the training-set A is already used as the source of multimodal signals, it is not considered in the PCG dataset for individual signals, mainly to reduce the risk of overfit and avoid data leaking.

TABLE 4.2: Number of normal and abnormal recordings for each database in the training set excluding training-set A. [61]

Database name	Abnormal Recordings	Normal Recordings
Training-b	104	386
Training-c	24	7
Training-d	28	27
Training-e	183	1958
Training-f	34	80
Total	373	2458

4.2 Pre-processing

Short signal frames with a duration of 5 seconds are extracted consecutively (with no overlapping) from the original records for both the PCG and ECG signals. The window size of 5 seconds is a common choice in studies regarding ECG and PCG [50, 77, 78]. Each frame is pre-processed before being submitted to CWT for scalogram generation.

For both signals, band-pass filters are applied. In the field of signal processing, a filter is a device or process that suppresses unwanted components or features from a signal. The most commonly used filters are low-pass, high-pass, band-pass and band-stop. The main characteristics that describe a filter are its type, order and cutoff frequency [79].

For the ECG signal the pre-processing consists on applying a 4^{th} order digital band-pass Butterworth filter with cufoff frequencies of 0.5 Hz and 100 Hz, as it encapsulates most of the useful frequency range for the ECG signal while attenuating noise and signals artefacts such as baseline wander[80, 81]. The filtered signal is then standardized using the following formula:

$$z = \frac{x - \mu}{\sigma} \tag{4.1}$$

in which, x representation the signal data points, μ represents the mean value of the signal and σ represents the standard deviation.

The PCG signal is also treated in a similar fashion in which a 4^{th} order digital bandpass Butterworth filter with cufoff frequencies of 25 Hz and 400 Hz, to remove low and high frequency noise while keeping the fundamental heart sounds and murmurs components [82, 83].

An additional spike removal algorithm is applied inspired on the work develop by Singh et al. [78, 84]. The steps for the spike removal algorithm are as follows:

Spike removal

1. Based on a 0.5-second window, divide the PCG recording.
2. Compute the maximum absolute amplitude (MAA) for each sliding window.
3. If the value of MAA equal three times the median values of MAA then advance to step (4) else continue.
   I. Determine the window with the highest MAA value.
   II. With the previous window as reference MAA, the noise spikes are computed from the respected location.
   III. Determine the last zero-crossing point using the starting location of the noise spike, which is just before the MAA point.
   IV. Determine the first zero-crossing point using the end location of the noise spike which is just after the MAA point.
   V. The determined noise spike is displaced by zeros.
   VI. Start again from step (2).
4. Tasks completed.

Like the ECG, the PCG is also standardized.

Figure 4.3 shows the result of the pre-processing for a raw PCG sample with noticeable noise and spikes.

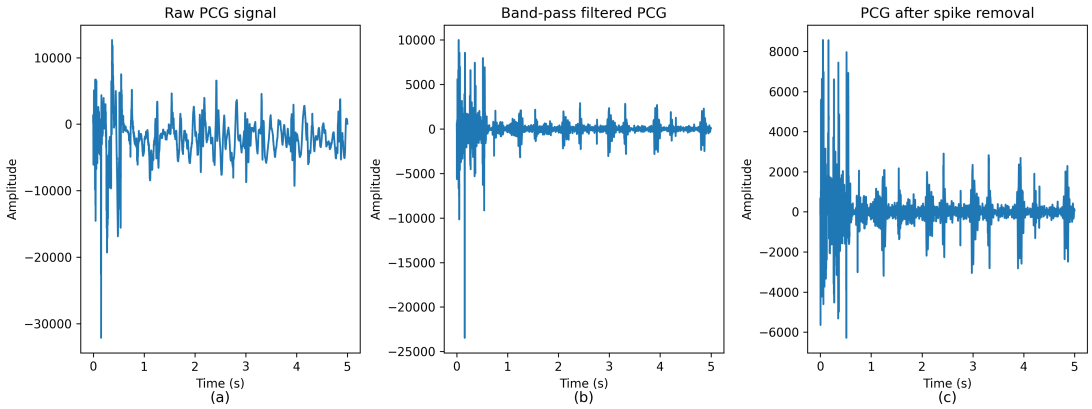


FIGURE 4.3: Plot of a PCG signal (a) through different stages of pre-processing: band-pass filter (b) and spike removal (c).

4.3 Scalogram Generation

After pre-processing, CWT is applied to PCG and ECG. Due to the distinct nature and properties of the signals, CWT parameters (such as the wavelet used) are different for PCG and ECG.

The complex Morlet is selected as the mother wavelet for the ECG CWT, as it has proven to be effective in this task [43]. The bandwith parameter and central frequency are set to 1.5 Hz and 1 Hz, respectively. For the ECG signals sampled at 2000 Hz, the scale parameter goes from 20 to 500, as the corresponding pseudo-frequencies encapsulate most of the ECG's fundamental frequency content. For the ECG signals sampled at 300 Hz (Physionet 2017) the scale ranges from 3 to 100, taking into account the influence of the sampling period in the calculation of the corresponding pseudo-frequencies, as shown in Equation 2.2.

The Morlet wavelet is chosen to perform the PCG CWT, as it is frequenctly used for this analysis [43, 85]. Since all the PCG signals have a sampling frequency of 2000 Hz, the scale parameter range is set from 4 to 100, taking into account the Morlet's default center frequency of 0.8125 Hz and the PCG's fundamental frequency content.

Figure 4.4 shows examples of scalograms generated from multimodal samples.

Table 4.3 presents the number of scalograms (corresponding to 5 seconds) generated (divided per class) for each of the datasets used, as well as, the number of recordings per dataset. Regarding the multimodal dataset, since each record contains 2 signals, each pair of scalograms is only counted once in the table calcultations.

TABLE 4.3: Numbers of recordings, abnormal and normal scalograms generated for each database.

Database	Records	Abnormal Scalograms	Normal Scalograms
Physionet 2017 (ECG only)	8482	21187	32115
Physionet 2016 (Multimodal)	405	1824	738
Physionet 2016 sets b,c,d,e,f (PCG only)	2831	1306	9119
Total	11718	25842	41971

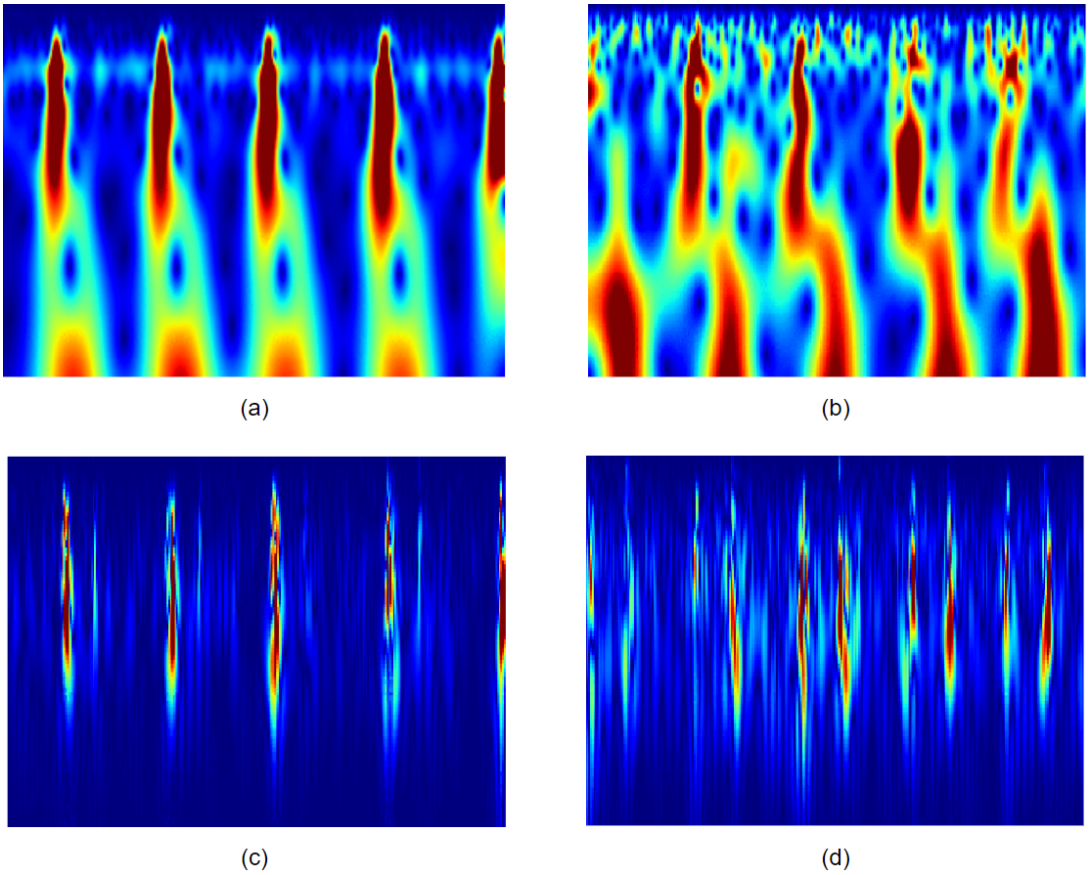


FIGURE 4.4: Scalograms generated from the multimodal dataset: (a) and (c) represent, respectively, the ECG and PCG from the same normal multimodal sample , (b) and (d) represent the ECG and PCG, respectively, for an abnormal multimodal sample.

4.4 Models

The base model implemented is derived from the VGG-16 architecture. The VGG-16 network was created by Karen Simonyan and Andrew Zisserman for the task of image classification from the University of Oxford in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition"[86]. The VGG16 model achieves almost 92.7% top-5 test accuracy in ImageNet (a dataset consisting of more than 14 million images belonging to nearly 1000 classes). It was one of the most popular models submitted to ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014 . It replaces the large kernel-sized filters with several 3×3 kernel-sized filters one after the other, making significant improvements over AlexNet [87]. Figure 4.5 shows the VGG-16 network.

The VGG-16 architecture consists of six blocks (five convolutional blocks and 1 classification block). The input consists of 224x224 RGB images. Each of the first two blocks

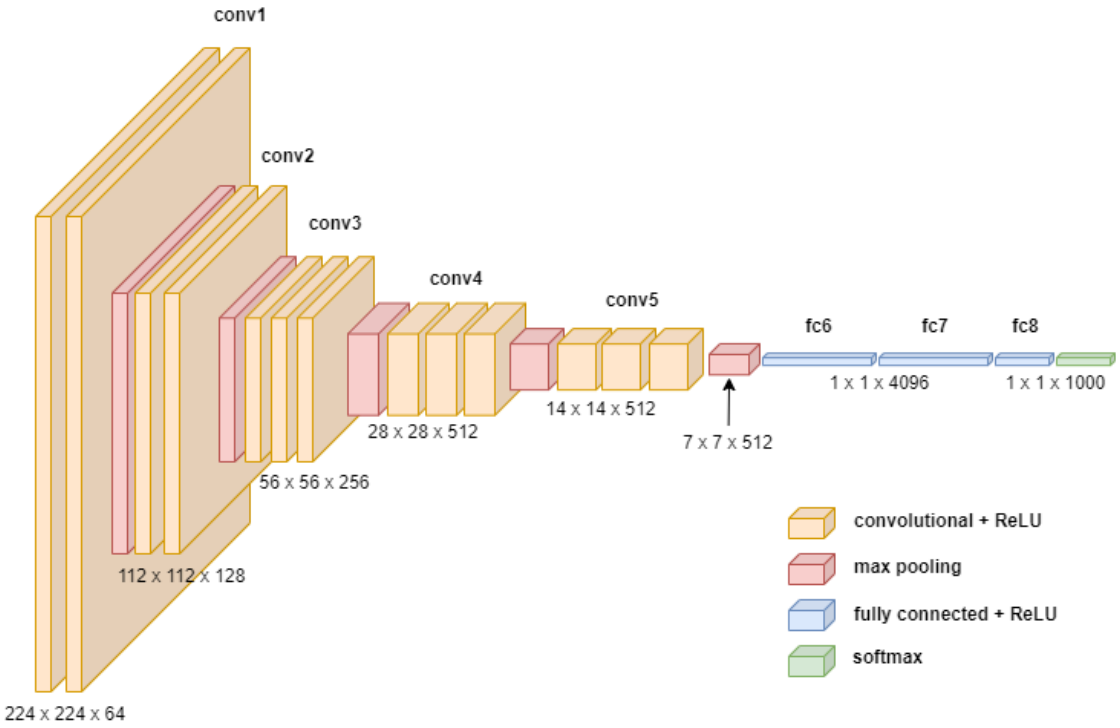


FIGURE 4.5: The original VGG-16 network.

contains two convolutional layers and one pooling layer, each of the next three blocks contains three convolutional layers and one pooling layer, and finally, the last block contains three fully connected layers. Therefore, the network, in total, has 13 convolutional layers, 5 pooling layers, and 3 fully connected layers. Each of the convolutional operations in this network is performed using filters of kernel size 3x3 and followed by an activation operation using the ReLU function [88].

A modification is made to the model, that affects only the classifier portion. The convolution portion is kept unchanged to allow the use of transfer learning of the pre-trained CNN on the ImageNet dataset. After the convolution section a flatten layer is added followed by a hidden dense layer with N neurons and ReLU activation, N is set to be tunable, a 0.5 dropout layer (to prevent overfitting) and a 1 neuron dense layer with a sigmoid activation.

Figure 4.6 shows the architecture of the custom VGG-16 network.

The presented modified network takes a single 224x224 RGB image as input and outputs the predicted binary classification probability. Since the main classification task of this study is multimodal, a dual input network is developed consisting essentially of two branches (one for PCG scalograms and the other ECG scalograms) of the individual custom network whose output after the N neuron fully connected network is concatenated

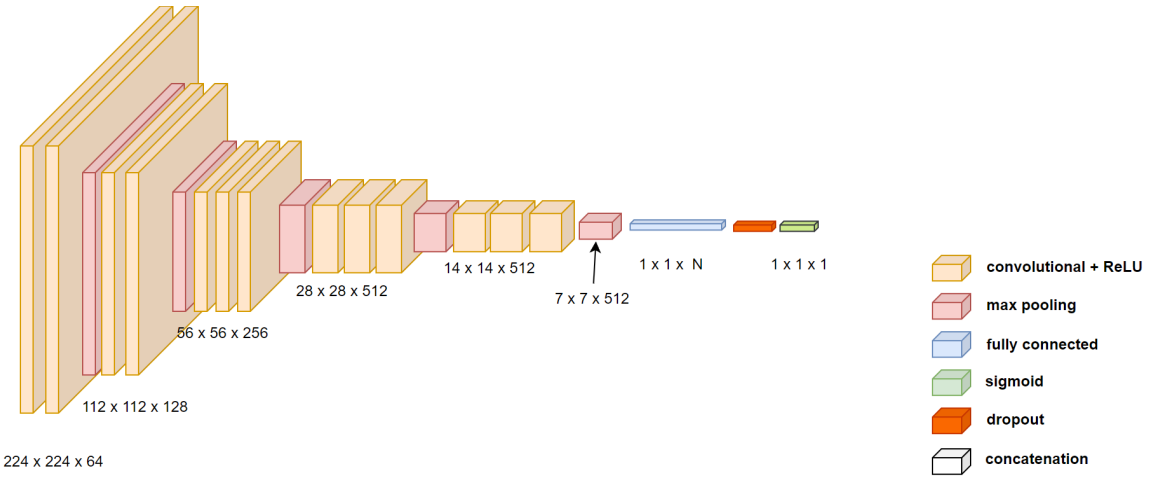


FIGURE 4.6: The custom VGG-16 network.

and passed through a multimodal classifier that consists of another N fully connected dense layer with ReLU activation, a dropout layer and a dense layer with 1 neuron and sigmoid output. The resulting multimodal network is shown in Figure 4.7.

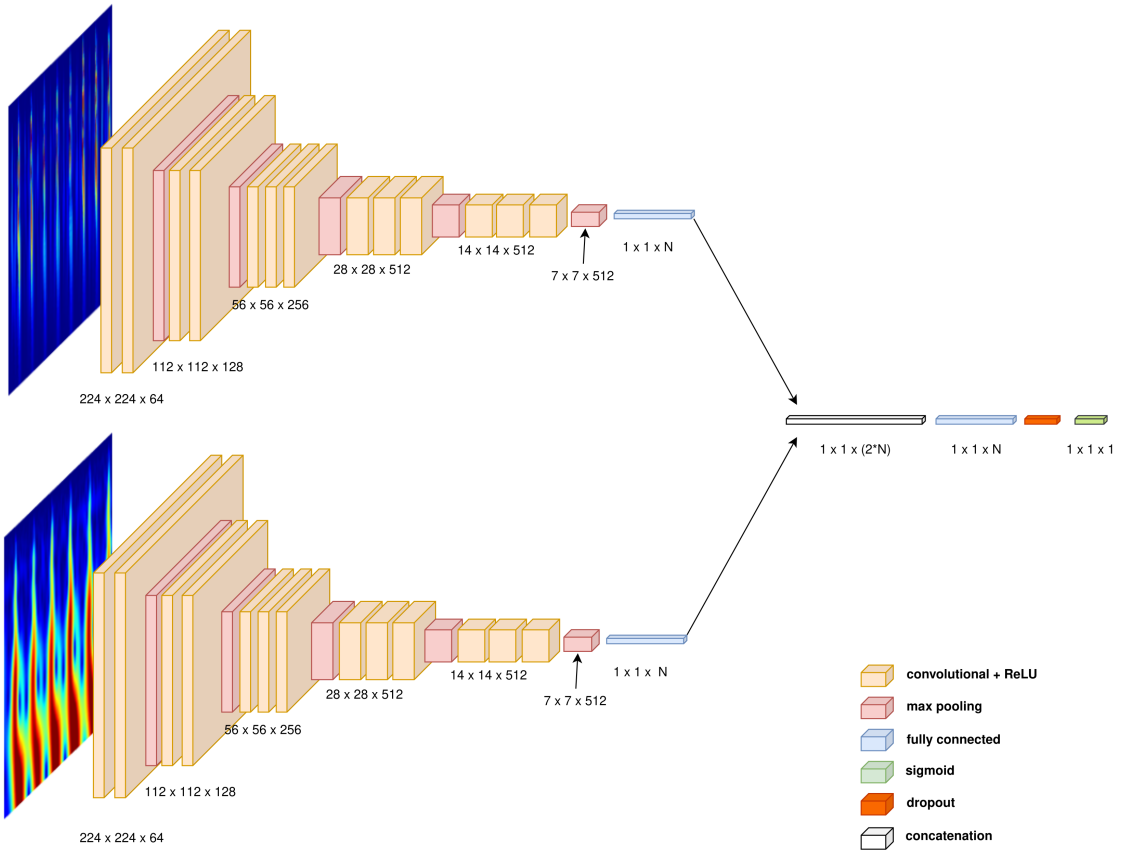


FIGURE 4.7: The custom multimodal VGG-16 network.

4.5 Data Splitting

The datasets are split into training and test. The split is stratified (meaning that the ratio of abnormal/normal records is kept the same in test and training) and record-wise to ensure that samples from the same record are not in both sets and the proportion used is 70% for training and 30% for test.

4.6 Model Selection

Throughout the experiments, grid search with 5-fold cross validation (GridSearchCV) is applied for hyperparameter tuning. Cross validation is applied only on the training set to keep the test set unseen as shown in Figure 4.8.

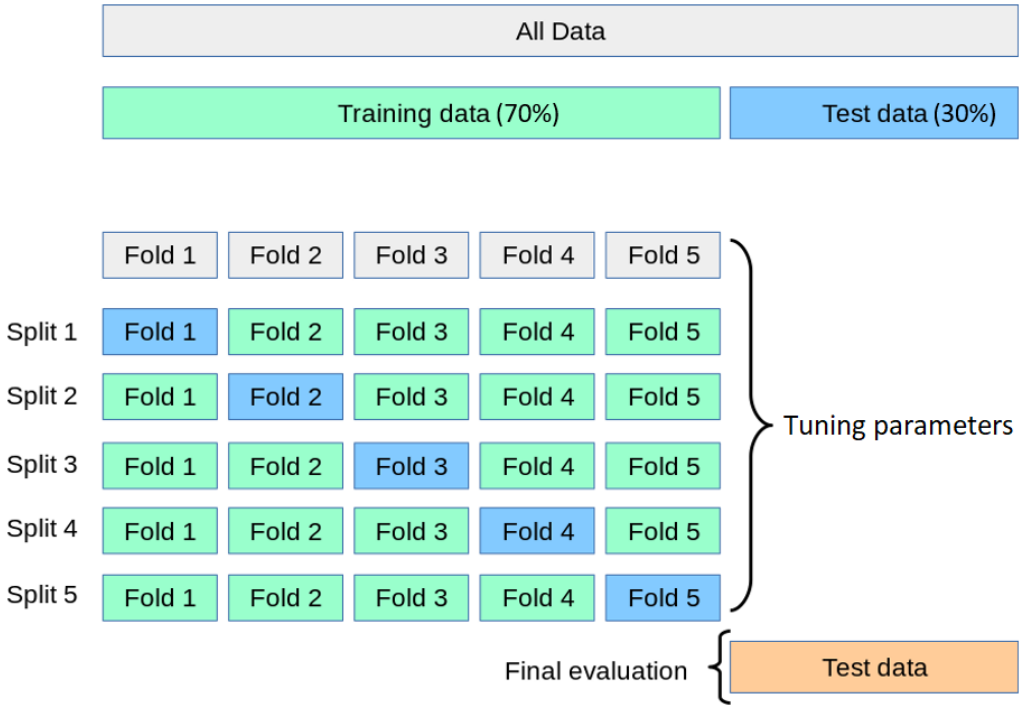


FIGURE 4.8: Schematic representation of the 5-fold cross validation applied on the training set (adapted from [89]).

In the cross validation data splitting, a group stratified split [90] is applied in order to both ensure that the class distribution across each fold's sets remains the same and making the cross validation sets exclusive, meaning that samples from the same group (record) cannot end up on different cross validation sets.

The grid search results are analyzed for each parameter combination. The average scoring metrics for sample-wise classification are calculated and the model with the best performance is selected. ROC AUC is the metric used to rank the models, since it summarizes the performance of a model across all possible decision thresholds. The selected model is retrained on the full train set and tested on the unseen test set for evaluation. Figure 4.9 illustrates the described process.

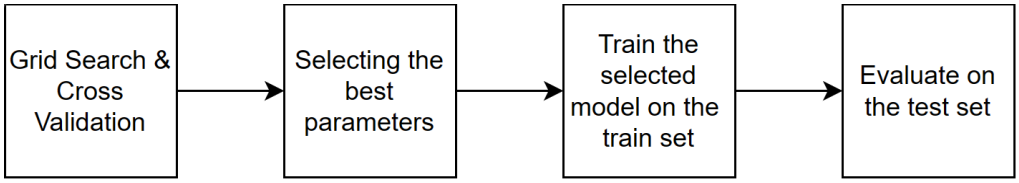


FIGURE 4.9: The model selection and evaluation process.

4.7 Classification

Predictions obtained from samples can be combined to obtain recordwise classification. Therefore two classification modalities are used:

- Frame-wise evaluation: Short signal samples are classified individually. If the probability determined by the model exceeds 0.5 the sample is classified as abnormal, otherwise the sample is classified as normal.
- Record-wise evaluation: The results from the frame-wise classification are taken into account aggregated by the respective record.

A soft voting probabilistic approach is taken to obtain the record-wise predicitions. For each record, a set of samples is evaluated by the model resulting in a set of predicted probabilities. The mean of predicted probabilities is computed and if it exceeds 0.5 the record is classified as abnormal.

Figure 4.10 shows an example of a 20 seconds record (record-A) that is split into 4 non-overlapping 5-seconds frames.

For example, if the predicted probabilities for frames 1-4 are 0.25, 0.50, 0.75, 0.80, respectively, the average value is 0.575, meaning that the record would be classified as abnormal since the average exceeds the applied threshold (0.5).

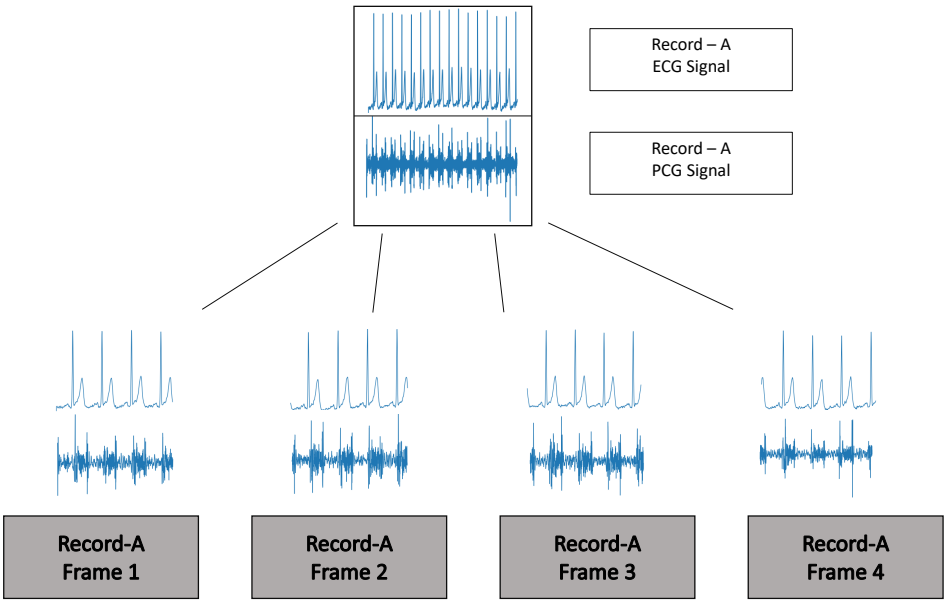


FIGURE 4.10: A multimodal record being splitted into multiple frames.

4.8 Experiments

This section provides a comprehensive overview of the experimental design and the strategies employed in this study. A detailed description of the hardware and software configurations used in the experiments, outlining the specific equipment, systems, and software versions utilized. Additionally, this section delves into experimental setttings used in the study, elaborating on the learning strategies applied.

4.8.1 Experimental Setup

The experiments presented in this work are programmed in Python 3.7.6 and run in a 64-bits Windows 10 operating system. The system hardware includes an AMD Ryzen 3600 CPU (with 6 cores and 12 logical processors), 16 GB of random access memory (RAM) and a NVIDIA GeForce GTX 1660 Super GPU with 1408 CUDA cores. Model implementations were performed using Keras, a deep learning API written in Python, running on top of the machine learning platform TensorFlow [91]. PyWavelets, an open source wavelet transform software for python, was used to perform CWT [92].

4.8.2 Experimental settings

Using the architectures described in Section 4.4, three experimental multimodal settings are developed:

- **Setting 1**: No transfer learning. The multimodal network is trained from scratch on the multimodal dataset using the custom multimodal architecture.
- **Setting 2**: Transfer learning from ImageNet. The multimodal network receives knowledge from the pre-trained network on ImageNet.
- **Setting 3**: Transfer learning and fine-tuning. The custom single input network is trained on the individual datasets, meaning the Physionet 2017 for ECG and the Physionet 2016 (sets b, c, d, e, f) for PCG. Knowledge is transferred to the multimodal network.

On setting 1, the multimodal network is trained completely from scratch, showing the performance without the use of transfer learning.

On setting 2, the feature extraction block for both branches of the multimodal network uses weights from the VGG-16 network pre-trained on ImageNet. The convolutional blocks are frozen and the classifier portion is trained from scratch.

On setting 3, learning is done firstly on the individual datasets. The feature extraction block is initializated with weights from a pre-trained VGG-16 network trained on ImageNet. The initial 3 convolutional blocks are frozen while the remaining 2 are finetuned. The classifier portion is trained from scratch. The individual learning pipeline consists in the following sequence:

1. GridSearchCV is performed on the training set. The best set of hyperparameters is selected.
2. Evaluation of the previously selected model by training on the full individual training set and testing on the test set.
3. Training on the full individual (single modality) PCG or ECG dataset (training and test sets).
4. The finetuned feature extractor weights are transferred to the respective branch on the multimodal network.

Figure 4.11 shows a schematic representation of the described individual learning process.

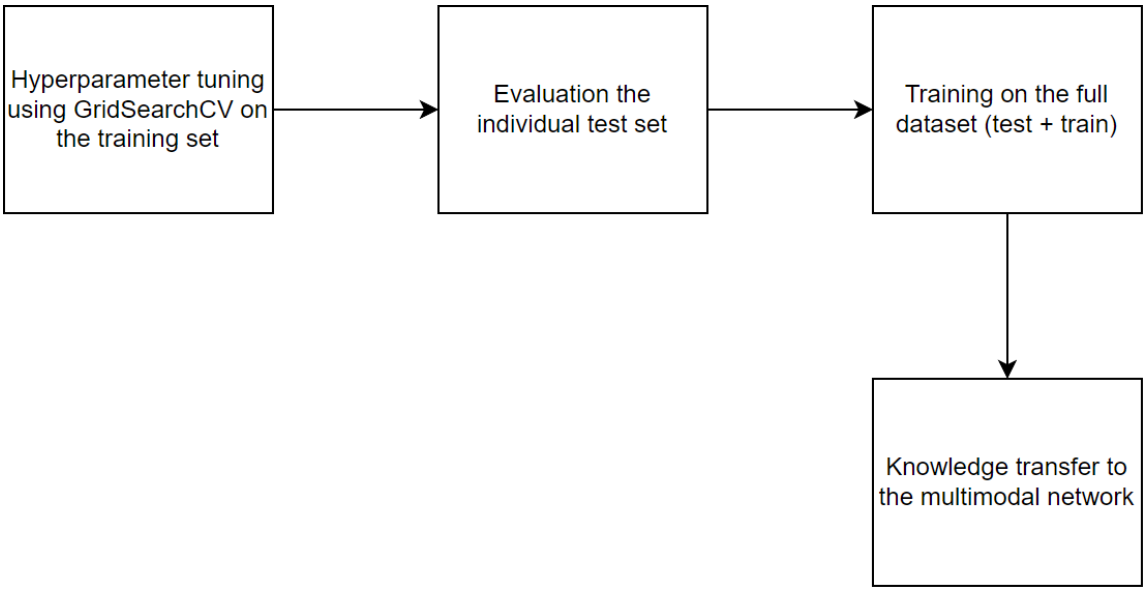


FIGURE 4.11: The individual learning pipeline.

The remaining portion of the multimodal network is trained from scratch on the multimodal dataset.

Figure 4.12 shows the architecture of the custom individual network with annotations regarding transfer learning.

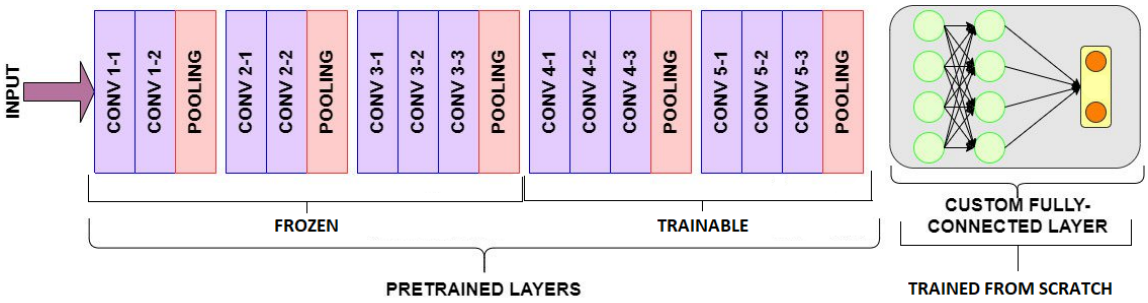


FIGURE 4.12: The custom individual network used on setting 3 (adapted from [93]).

Binary cross entropy is used as the loss function across all models/settings, since it is the standard and most commonly used for binary classification problems [94].

Regarding the optimizer, Adam (Adaptive Moment Estimation) is used, since it is computationally efficient and broadly adopted for deep learning applications [95].

Since the class distributions in the datasets used throughout the experimental settings are not balanced, the use of class weighs is experimented upon, assigning higher weights to the minority class resulting in a reduced bias towards the majority class. Class weights

are calculated based on the dataset's inverse of the class frequencies. The weight for each class is computed by dividing the total number of samples by the product of the number of classes and the number of samples in that class [96]. To incorporate these class weights during training, a weighted binary cross-entropy loss function is employed, which adjusts the contribution of each class to the overall loss according to its weight.

In addition to the multimodal settings, baseline single input settings are also employed. These networks serve as a baseline, providing a comparative measure for the performance of individual and multimodal approaches. The single input networks are trained from scratch, similarly to setting 1, considering the individual signals from the target multimodal dataset. Meaning that the individual samples are considered separately, disregarding the multimodal component.

Table 4.4 shows the set of parameters used on the GridSearchCV process for each setting. The number of neurons parameter corresponds to the amount of neurons per fully connected hidden layer in the classifier portion of the network.

TABLE 4.4: The grid search parameters used in the experiments.

Setting	Model	Batch Size(s)	Learning Rate(s)	Epochs	Number of Neurons	Class Weights
Baseline	Individual - PCG	32	1e-6, 1e-5, 5e-5	10	32	True, False
Baseline	Individual - ECG	32	1e-6, 1e-5, 5e-5	10	32	True, False
1	Multimodal	8	1e-6, 1e-5, 0.0005	30	128	True, False
2	Multimodal	8	1e-6, 1e-5, 5e-5	10, 20	32	True, False
3	Individual - PCG	32	5e-7, 1e-6	30	128	True, False
3	Individual - ECG	32	1e-6	20, 30	128	True, False
3	Multimodal	8	0.0001, 0.001, 0.005	20	32	True, False

Chapter 5

Results

In this chapter, the outcomes from the experiments described on Section 4.8.2 are presented.

Firstly, the GridSearchCV process results are shown. This technique allows the determination of the optimal set of parameters for each model.

Furthermore, the final evaluation results of the selected models is presented and analyzed, providing an understanding of the predictive capabilities of the model on unseen data, for both sample-wise and record-wise classification.

5.1 Hyperparameter Tuning

In this section, results regarding the hyperparameter tuning through GridSearchCV are presented for the previously mentioned experimental settings. Average cross validation train and test scores are displayed, as well as, loss and accuracy score curves to show the models predictive power and generalization capabilities.

5.1.1 Baseline

5.1.1.1 PCG

Table 5.1 shows the GridSearchCV results for the baseline PCG only model (trained on the PCG portion of the multimodal dataset).

The best set of parameters (highest test ROC AUC score) for this model corresponds to a learning rate of 1e-5, batch size of 32, 10 epochs, 32 neurons for the dense hidden layer, and without the inclusion of class weights.

TABLE 5.1: The GridSearchCV results for the individual PCG network trained from scratch (baseline).

Parameters	Mean Train ROC AUC	Mean Test ROC AUC	Mean Train F1-score	Mean Test F1-score
Learning Rate: 1e-05, Batch Size: 32, Epochs: 10, Neurons: 32, Class Weights: False	0.809	**0.626**	0.852	0.829
Learning Rate: 1e-05, Batch Size: 32, Epochs: 10, Neurons: 32, Class Weights: True	0.776	0.578	0.709	0.604
Learning Rate: 5e-05, Batch Size: 32, Epochs: 10, Neurons: 32, Class Weights: False	0.663	0.607	0.826	0.825
Learning Rate: 1e-06, Batch Size: 32, Epochs: 10, Neurons: 32, Class Weights: True	0.664	0.540	0.697	0.643
Learning Rate: 1e-06, Batch Size: 32, Epochs: 10, Neurons: 32, Class Weights: False	0.691	0.574	0.832	0.801
Learning Rate: 5e-05, Batch Size: 32, Epochs: 10, Neurons: 32, Class Weights: True	0.604	0.563	0.478	0.495

Figure 5.1, shows the training and validation loss and accuracy score curves, throughout the training epochs.

5.1.1.2 ECG

Table 5.1 shows the GridSearchCV results for the baseline ECG only model (trained on the ECG portion of the multimodal dataset).

The best set of parameters for this model corresponds to a learning rate of 1e-5, batch size of 32, 10 epochs, 32 neurons for the dense hidden layer, and with the inclusion of class weights.

Figure 5.2, shows the training and validation loss and accuracy score curves, throughout the training epochs.

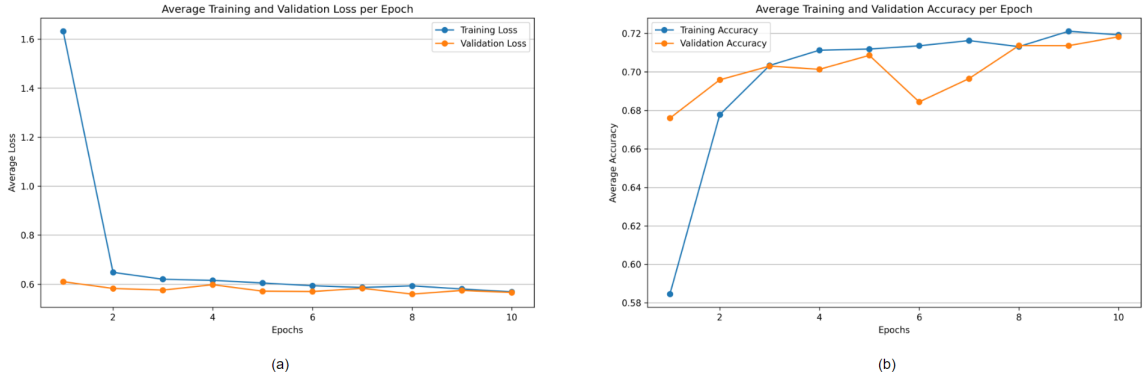


FIGURE 5.1: The baseline PCG model average loss (a) and average accuracy (b) for the training and validation during cross validation.

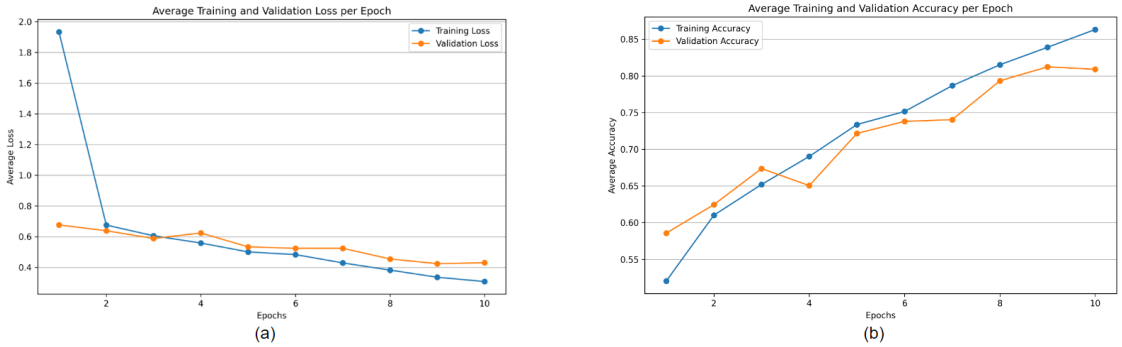


FIGURE 5.2: The baseline ECG model average loss (a) and average accuracy (b) for the training and validation during cross validation.

5.1.2 Setting 1

Table 5.3 shows the GridSearchCV of the setting 1 model trained from scratch on the multimodal dataset.

The set of parameters responsible for the best performance are: learning rate of 1e-5, batch size of 8, 30 epochs, no class weights and 32 neurons for the dense hidden layer.

Figure 5.3, shows the training and validation loss and accuracy score curves, throughout the training epochs.

5.1.3 Setting 2

Table 5.4 shows the GridSearchCV results of the setting 2 model.

The set of parameters responsible for the best performance are: learning rate of 5e-5, batch size of 8, 20 epochs, no class weights and 32 neurons for the dense hidden layer.

Figure 5.4, shows the training and validation loss and accuracy score curves, throughout the training epochs.

TABLE 5.2: The GridSearchCV results for the individual ECG network trained from scratch (baseline).

Parameters	Mean Train ROC AUC	Mean Test ROC AUC	Mean Train F1-score	Mean Test F1-score
Learning Rate: 1e-05, Batch Size: 32, Epochs: 10, Neurons: 32, Class Weights: True	0.993	**0.875**	0.962	0.863
Learning Rate: 1e-05, Batch Size: 32, Epochs: 10, Neurons: 32, Class Weights: False	0.982	0.800	0.945	0.815
Learning Rate: 5e-05, Batch Size: 32, Epochs: 10, Neurons: 32, Class Weights: True	0.827	0.745	0.660	0.590
Learning Rate: 5e-05, Batch Size: 32, Epochs: 10, Neurons: 32, Class Weights: False	0.839	0.734	0.851	0.811
Learning Rate: 1e-06, Batch Size: 32, Epochs: 10, Neurons: 32, Class Weights: True	0.878	0.707	0.839	0.753
Learning Rate: 1e-06, Batch Size: 32, Epochs: 10, Neurons: 32, Class Weights: False	0.847	0.674	0.869	0.806

5.1.4 Setting 3

In setting 3, firstly the individual models GridSearchCV results are displayed. The selected individual (single input) models are retrained on the full individual dataset and knowledge is transferred to the respective branch of the multimodal network. Multimodal GridSearchCV results are also shown in this section.

5.1.4.1 Individual - ECG

Table 5.5 shows the GridSearchCV results for the individual finetuned ECG model (trained on the Physionet 2017 dataset).

TABLE 5.3: The GridSearchCV results for the multimodal network trained from scratch on the multimodal dataset (setting 1).

Parameters	Mean Train ROC AUC	Mean Test ROC AUC	Mean Train F1-score	Mean Test F1-score
Learning Rate: 1e-05, Batch Size: 8, Epochs: 30, Neurons: 128, Class Weights: False	0.989	**0.869**	0.969	0.867
Learning Rate: 1e-05, Batch Size: 8, Epochs: 30, Neurons: 128, Class Weights: True	0.977	0.862	0.941	0.845
Learning Rate: 1e-06, Batch Size: 8, Epochs: 30, Neurons: 128, Class Weights: True	0.980	0.814	0.857	0.734
Learning Rate: 1e-06, Batch Size: 8, Epochs: 30, Neurons: 128, Class Weights: False	0.982	0.812	0.953	0.831
Learning Rate: 0.0005, Batch Size: 8, Epochs: 30, Neurons: 128, Class Weights: False	0.571	0.563	0.830	0.830
Learning Rate: 0.0005, Batch Size: 8, Epochs: 30, Neurons: 128, Class Weights: False	0.500	0.500	-	-

The best set of parameters for this model corresponds to a learning rate of 1e-6, batch size of 32, 30 epochs, 128 neurons for the dense hidden layer, and without the use of class weights.

5.1.4.2 Individual - PCG

Results from the GridSearchCV for the individual PCG network finetuned on the Physionet 2016 PCG-only dataset (sets b, c, d, e, f) are displayed on Table 5.6.

The best set of parameters for this model corresponds to a learning rate of 1e-6, batch size of 32, 30 epochs, 128 neurons for the dense hidden layer, and the use of class weights.

5.1.4.3 Multimodal

The setting 3 multimodal GridSearchCV results are displayed on Table 5.7.

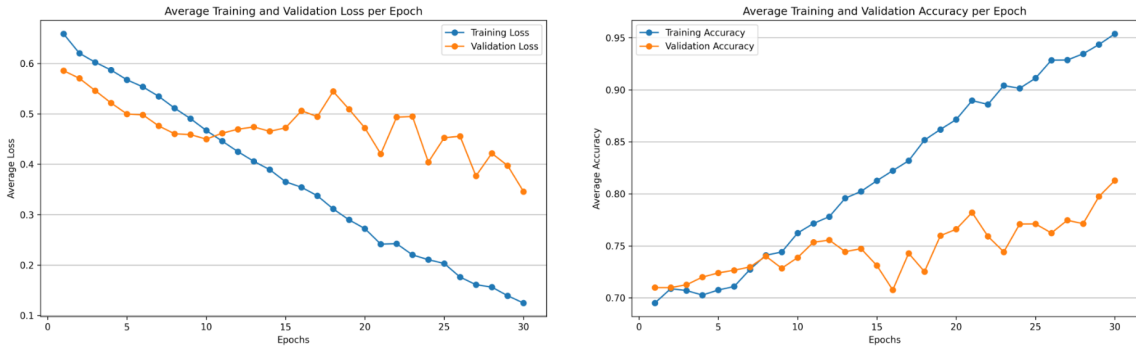


FIGURE 5.3: The setting 1 model average accuracy (a) and average loss (b) for the training and validation during cross validation.

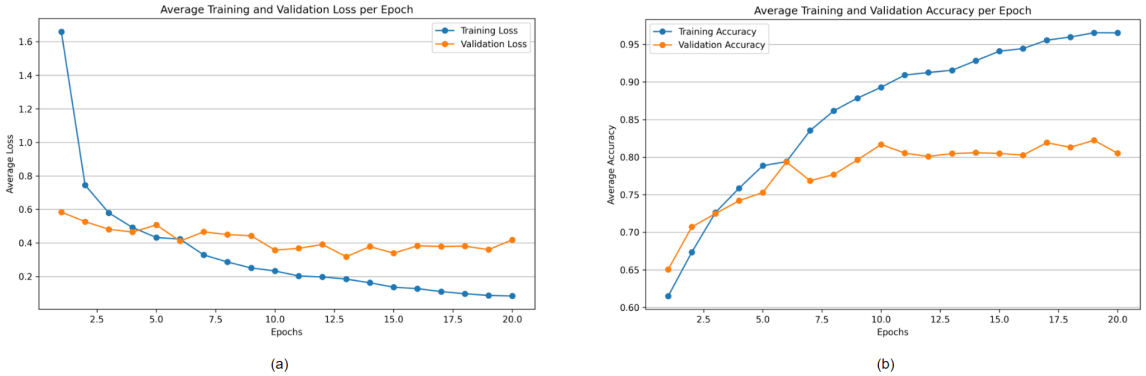


FIGURE 5.4: The setting 2 model average loss (a) and average accuracy (b) for the training and validation during cross validation.

The best set of parameters for this model corresponds to a learning rate of 5e-5, batch size of 8, 20 epochs, 32 neurons for the dense hidden layer and without the use of class weights.

Figure 5.5 shows the training and validation loss and accuracy score curves, throughout the training epochs.

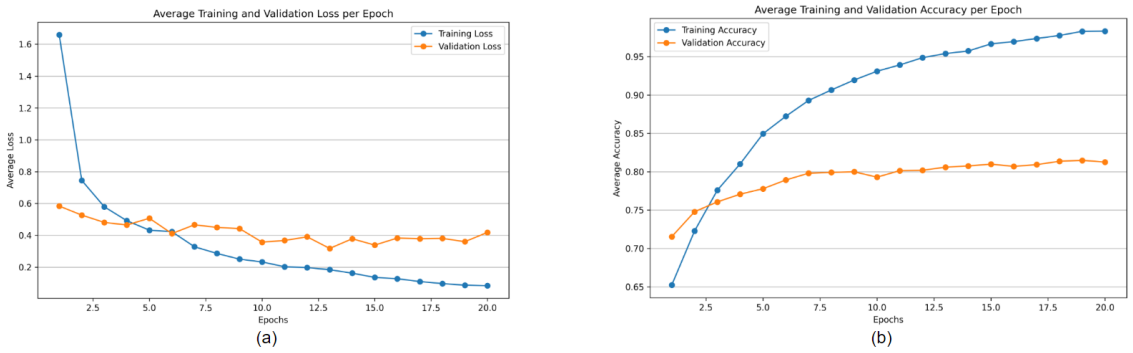


FIGURE 5.5: The setting 3 model average loss (a) and average accuracy (b) for the training and validation during cross validation.

TABLE 5.4: The GridSearchCV results for the multimodal network with ImageNet weights (setting 2).

Parameters	Mean Train ROC AUC	Mean Test ROC AUC	Mean Train F1-score	Mean Test F1-score
Learning Rate: 5e-05, Batch Size: 8, Epochs: 20, Neurons: 32, , Class Weights: False	1.000	**0.882**	0.989	0.855
Learning Rate: 5e-05, Batch Size: 8, Epochs: 20, Neurons: 32, Class Weights: True	1.000	0.869	0.985	0.858
Learning Rate: 5e-05, Batch Size: 8, Epochs: 10, Neurons: 32, Class Weights: False	0.984	0.860	0.962	0.863
Learning Rate: 5e-05, Batch Size: 8, Epochs: 10, Neurons: 32, Class Weights: True	0.995	0.852	0.949	0.811
Learning Rate: 1e-05, Batch Size: 8, Epochs: 20, Neurons: 32, Class Weights: False	0.998	0.835	0.985	0.849
Learning Rate: 1e-05, Batch Size: 8, Epochs: 20, Neurons: 32, Class Weights: True	0.994	0.824	0.950	0.809
Learning Rate: 1e-05, Batch Size: 8, Epochs: 10, Neurons: 32, Class Weights: True	0.956	0.759	0.889	0.739
Learning Rate: 1e-05, Batch Size: 8, Epochs: 10, Neurons: 32, Class Weights: False	0.962	0.753	0.930	0.822
Learning Rate: 1e-06, Batch Size: 8, Epochs: 20, Neurons: 32, Class Weights: True	0.794	0.658	0.751	0.696
Learning Rate: 1e-06, Batch Size: 8, Epochs: 20, Neurons: 32, Class Weights: False	0.748	0.615	0.839	0.796
Learning Rate: 1e-06, Batch Size: 8, Epochs: 10, Neurons: 32, Class Weights: False	0.684	0.614	0.796	0.775
Learning Rate: 1e-06, Batch Size: 8, Epochs:10, Neurons: 32, Class Weights: False	0.647	0.584	0.678	0.638

TABLE 5.5: The GridSearchCV results for the individual network finetuned on the Physionet 2017 ECG-only dataset.

Parameters	Mean Train ROC AUC	Mean Test ROC AUC	Mean Train F1-score	Mean Test F1-score
Learning Rate: 1e-06, Batch Size: 32, Epochs: 30, Neurons: 128, Class Weights: False	0.933	**0.855**	0.819	0.738
Learning Rate: 1e-06, Batch Size: 32, Epochs: 20, Neurons: 128, Class Weights: False	0.916	0.845	0.780	0.700
Learning Rate: 1e-06, Batch Size: 32, Epochs: 30, Neurons: 128, Class Weights: True	0.948	0.845	0.846	0.731
Learning Rate: 1e-06, Batch Size: 32, Epochs: 20, Neurons: 128, Class Weights: True	0.917	0.832	0.822	0.701

TABLE 5.6: The GridSearchCV results for the individual network finetuned on the Physionet 2016 PCG-only dataset (sets b, c, d, e, f).

Parameters	Mean Train ROC AUC	Mean Test ROC AUC	Mean Train F1-score	Mean Test F1-score
Learning Rate: 1e-6, Batch Size: 32, Epochs: 30, Neurons: 128, Class Weights: True	0.996	**0.969**	0.853	0.730
Learning Rate: 1e-6, Batch Size: 32, Epochs: 30, Neurons: 128, Class Weights: False	0.999	0.967	0.947	0.673
Learning Rate: 5e-7, Batch Size: 32, Epochs: 30, Neurons: 128, Class Weights: True	0.986	0.956	0.800	0.696
Learning Rate: 5e-7, Batch Size: 32, Epochs: 30, Neurons: 128, Class Weights: False	0.990	0.955	0.860	0.639

TABLE 5.7: The GridSearchCV results for the multimodal finetuned network (setting 3).

Parameters	Mean Train ROC AUC	Mean Test ROC AUC	Mean Train F1-score	Mean Test F1-score
Learning Rate: 5e-5, Batch Size: 8, Epochs: 20, Neurons: 32, Class Weights: True	1.000	**0.845**	0.994	0.867
Learning Rate: 5e-5, Batch Size: 8, Epochs: 20, Neurons: 32, Class Weights: False	1.000	0.839	0.998	0.868
Learning Rate: 1e-5, Batch Size: 8, Epochs: 20, Neurons: 32 , Class Weights: False	0.977	0.806	0.945	0.863
Learning Rate: 1e-5, Batch Size: 8, Epochs: 20, Neurons: 32 , Class Weights: True	0.977	0.804	0.932	0.813
Learning Rate: 1e-6, Batch Size: 8, Epochs: 20, Neurons: 32, Class Weights: True	0.746	0.595	0.728	0.658
Learning Rate: 1e-6, Batch Size: 8, Epochs: 20, Neurons: 32, Class Weights: False	0.702	0.588	0.831	0.809

5.2 Evaluation

This section is devoted to the presentation of the final evaluation results. The selected models are retrained on the full training set, tested and sample-wise classification is obtained. Combining each record's samples predictions allows the calculation of the record-wise prediction.

5.2.1 Sample-wise classification

Scoring metrics obtained for sample-wise classification of the selected models are shown on Table 5.8. Additionaly, the confusion matrices for all the experimental settings are displayed on Figure 5.6.

TABLE 5.8: Sample-wise testing scores.

Setting	ROC AUC	Recall	Precision	Accuracy	F1-score
Baseline - PCG	0.739	**0.949**	0.749	0.734	0.837
Baseline - ECG	0.865	0.875	**0.861**	0.808	0.868
Setting 1	0.859	0.909	0.831	0.802	0.868
Setting 2	**0.888**	0.920	0.851	**0.827**	**0.884**
Setting 3	0.809	0.817	0.848	0.763	0.832

Setting 2 model presents the best overall performance, outperforming the other settings across all metrics presented, besides recall and precision.

Although, the baseline PCG model achieves the highest recall, it has the lowest scores in the other metrics as it seems to exhibit a high bias towards classifying records as abnormal (positive) as shown by the number of false positives in the confusion matrix.

Setting 2 outperforms the baseline models regarding ROC AUC, accuracy and F1-score, demonstrating the effectiveness of the multimodal approach.

Setting 3 fails to outperform the baseline ECG model, showing that improvements have to be made in this modality in order to surpass the baseline.

Figure 5.7 shows the sample-wise classification ROC curves for all the experimental settings.

5.2.2 Record-wise classification

Record-wise testing scores are presented on Table 5.9.

Figure 5.8 shows the confusion matrices for each of the experiments.

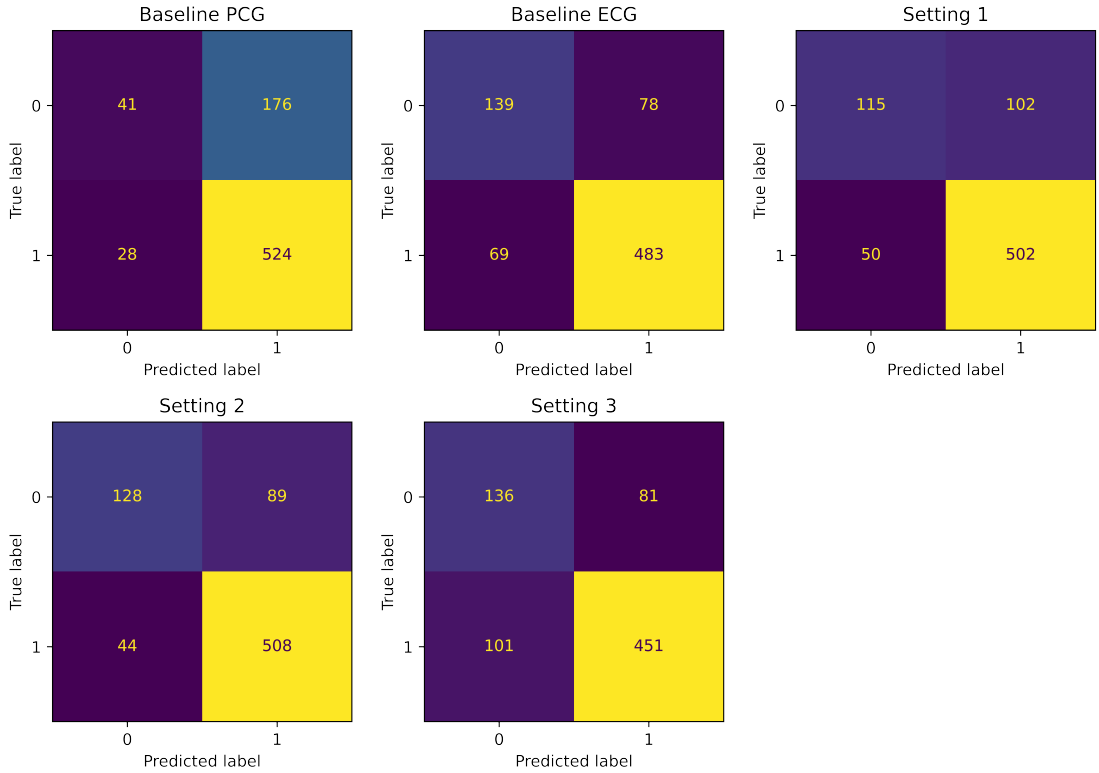


FIGURE 5.6: The test sample-wise classification confusion matrices for baseline PCG, baseline ECG, setting 1, setting 2 and setting 3 models.

TABLE 5.9: Record-wise testing scores.

Setting	ROC AUC	Recall	Precision	Accuracy	F1-score
Baseline - PCG	0.797	**0.966**	0.743	0.738	0.840
Baseline - ECG	0.885	0.885	0.856	0.811	0.870
Setting 1	0.861	0.931	0.826	0.811	0.876
Setting 2	**0.913**	0.931	0.844	**0.828**	**0.885**
Setting 3	0.839	0.851	**0.860**	0.795	0.855

Record-wise scores show a general increase in ROC AUC when compared with the sample-wise respective across all models, demonstrating a performance gain associated with the combination of several samples per record.

Setting 1 has a very close performance to the ECG baseline model, having better results in terms of recall and f1-score, identical accuracy but worse ROC AUC and precision.

Setting 1 and setting 2 models have identical performance regarding recall. However, setting 2 classifies the negative records better having higher score across the other metrics.

Setting 3 has the highest precision across all the record-wise classification score. However it is outperformed by the baseline ECG on the other metrics.

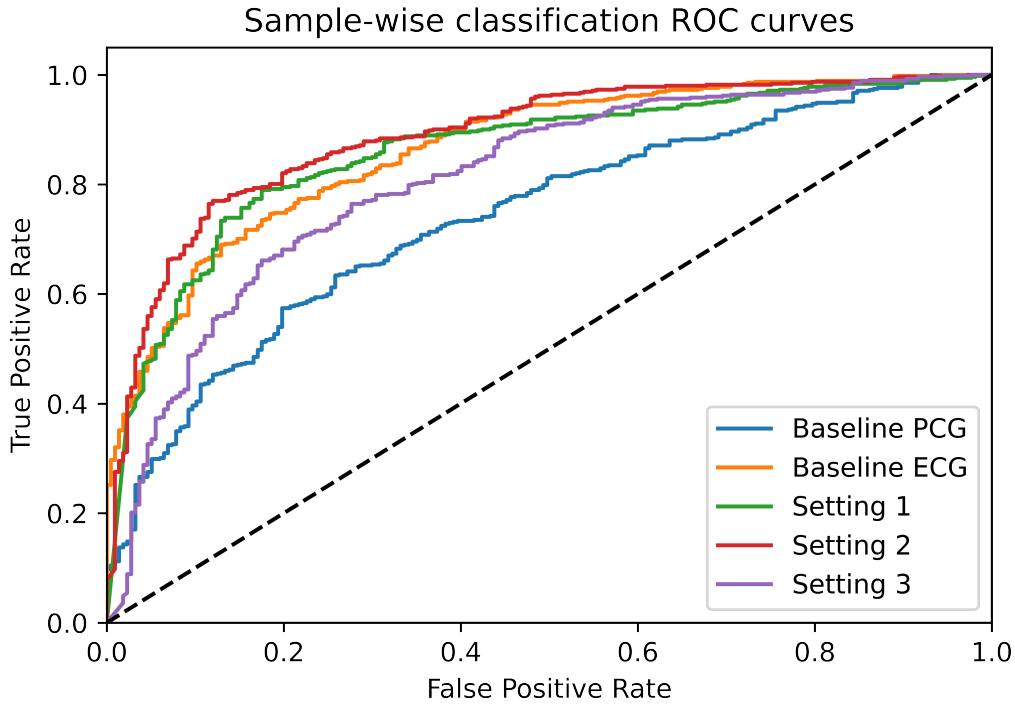


FIGURE 5.7: The sample-wise classification ROC curves.

The baseline PCG model exhibits the same behavior as in sample-wise classification, showing a significant bias towards classifying records as positive.

Figure 5.9 shows the record-wise classification ROC curves for the experiments performed in this study.

Setting 2 has the best overall results, like in the sample-wise classification, achieving the highest score in all the metrics besides recall.

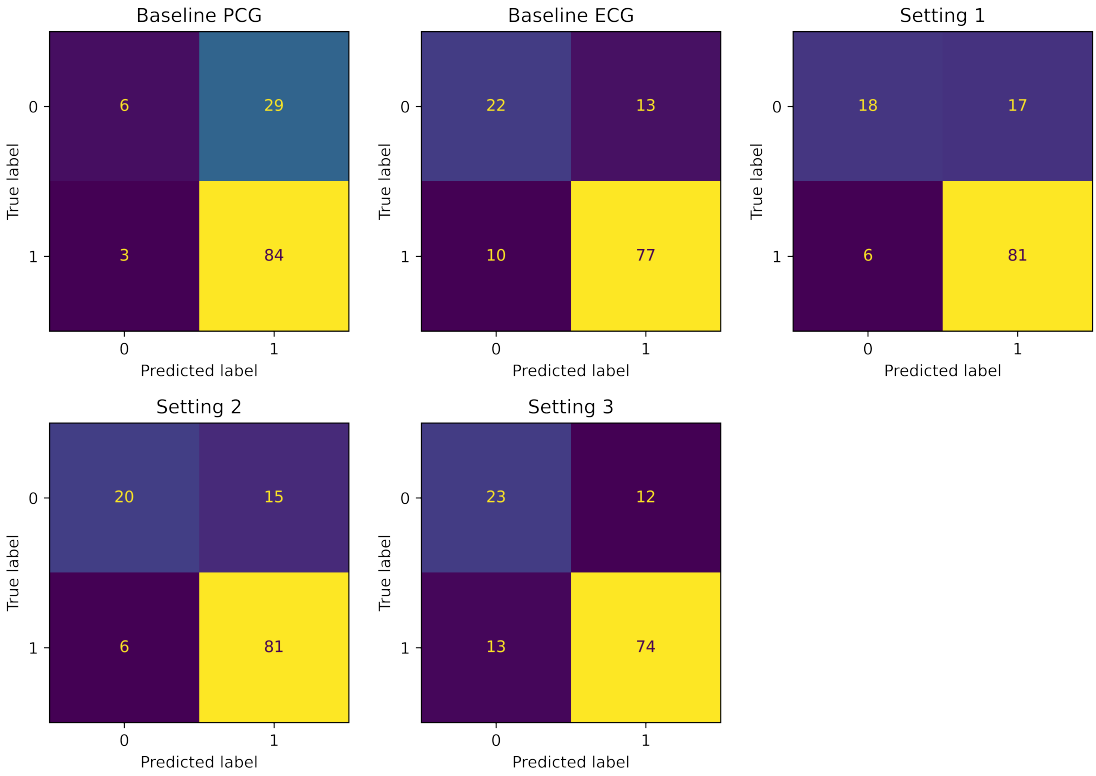


FIGURE 5.8: The test record-wise classification confusion matrices for baseline PCG, baseline ECG, setting 1, setting 2 and setting 3 models.

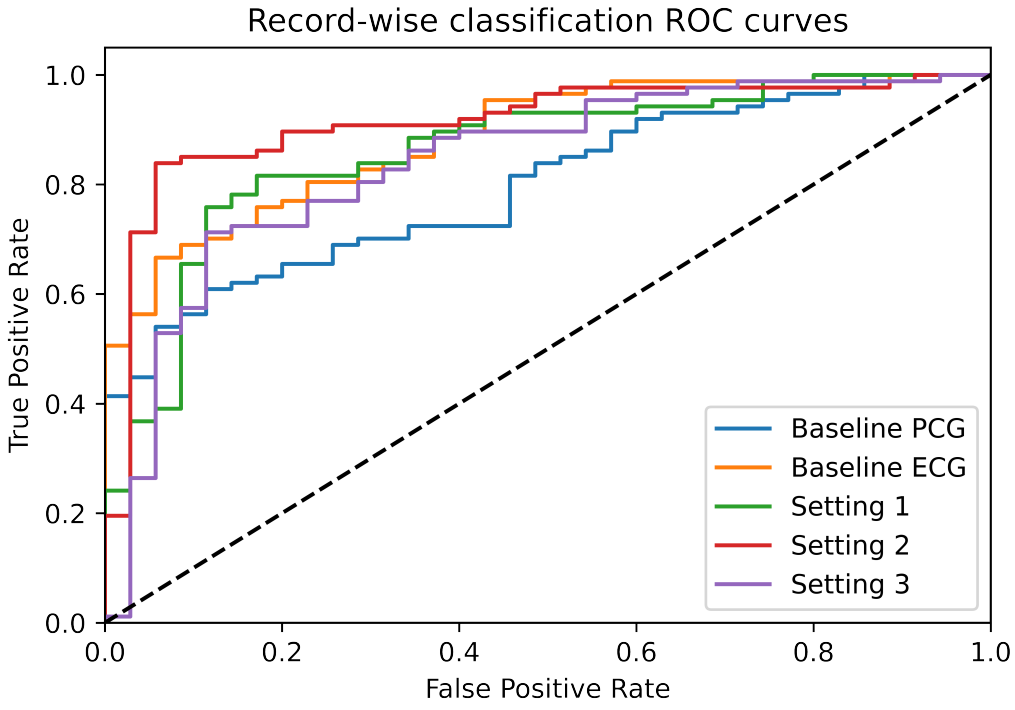


FIGURE 5.9: The record-wise classification ROC curves.

5.3 Discussion

A substancial portion of the models show signs of overfitting, having significantly better training scores when compared with testing scores on the GridSearchCV. This could be improved upon by reducing model complexity, adding regularization techniques and using early stopping to stop training when the model's performance starts degrading [97].

Setting 3 failed to outperform the ECG baseline, this might be explained by several factors namely, differences in the label distribution of the datasets used for finetuning when compared with the target dataset, especially regarding the PCG dataset, Physionet 2016 PCG only (datasets b, c, d, e, f) has only 14% of abnormal samples, meanwhile the multimodal dataset has roughly 70%. The use of other larger and more balanced datasets for finetuning, such as, the CirCor Digiscope Phonocardiogram Dataset [98], could also contribute to improvements in this setting.

Nevertheless, the setting 2 model achieved an overall better performance when compared with the other experimental settings, demonstrating the potential of transfer learning and multimodal classification.

As expected, settings whose best parameter combination included the use of class weights (baseline ECG and setting 3) have results less biased towards the majority class, performing better in classifying the minority (negative) class, as shown in Figure 5.8.

Table 5.10 shows a comparison of the setting 2 performance with other author implementations of multimodal PCG and ECG approaches that used the multimodal Physionet 2016 dataset.

TABLE 5.10: Comparison of the best performing experiments developed in this study and other multimodal state of the art works .

Author	**Method**	**Modality**	**Accuracy (%)**	**AUC (%)**	**Recall (%)**
Chakir et al. [44] (2020)	SVM	Record-wise	92.50	95.05	92.31
Hettiarachchi et al. [43] (2021)	CNN	Record-wise	87.67	93.75	87.72
Hettiarachchi et al. [43] (2021)	CNN	Sample-wise	81.45	85.83	81.71
This study - setting 2	CNN	Sample-wise	82.70	88.77	92.02
This study - setting 2	CNN	Record-wise	82.79	91.26	93.10

The developed setting 2 model shares a similar performance (specially regarding sample-wise classification) to the Hettiarachch et al. CNN approach that used the full 405 records from the Physionet 2016 multimodal dataset (split into 70% training, 10% validation and 20% test).

The SVM based method implemented by Chakir et al. [44], outperforms the approaches developed on this work. The results obtained by this author were based on a small subsample of 100 simultaneous records (split into 60% training and 40% test) from the Physionet 2016 multimodal dataset and classification was performed using a set of handcrafted features which might not be able to capture diverse and complex abnormalities that may appear on a larger dataset.

Chapter 6

Conclusions

PCG and ECG have been used separately for decades to detect cardiac abnormalities. The main goal of this thesis is to show the effectiveness and potential of the multimodal analysis of simultaneously recorded ECG and PCG signals. DL approaches, based on CNN, were developed and implemented.

The main current limitation in the developing of machine learning based multimodal PCG and ECG classification algorithms is the lack of publicly available data. Transfer learning approaches were developed in order to leverage knowledge acquired in other domains.

The most successful approach, combining multimodal analysis and transfer learning, achieved scores of 82.79%, 91.26 %, 88.52% for accuracy, ROC AUC and F1-score, respectively, highlighting the potential of this techniques.

Although the objectives of this work were fulfilled, in general, significant improvements can still be made, which can ultimately enhance the performance. Section 6.1 provides an in depth analysis of the current limitations and future research directions.

6.1 Future work

In the course of this study, several current limitations and potential avenues for future research have been identified. These opportunities for further improvement, exploration and development are outlined below:

- The expansion of the individual datasets used for finetuning: by incorporating datasets such as the CirCor Digiscope Phonocardiogram Dataset and the PhysioNet/-Computing in Cardiology Challenge 2020 dataset [98, 99] .

- The use of image data augmentation techniques to increase the training size and attenuate overfitting [100].
- Dealing with data imbalance by using techniques besides class weights, such as resampling, which can lead to less biased models.
- The use of other popular features, such as the spectogram and MFCC, to find the most effective features and possibly combine them.
- The use of other CNN architectures, such as residual networks and GoogLeNet [101, 102]. Early fusion based archictectures can also be explored.
- Besides the implemented soft voting record-wise classification, other approaches could also be studied, namely hard voting, which can lead to improvements in record-wise performance.
- The use of segmentation techniques, such as the Pan-Tompkins algorithm for ECG and hidden Markov models for PCG, to study the impact of the use of segmentation in the performance [103].
- Exploration of different parameters related with the scalogram generation, such as segment sizes, mother wavelet used, scale range and colormap to find the parameters that can lead to the most effective features.

Bibliography

[1] W. H. Organization, "Cardiovascular diseases (CVDs)," Jun. 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [Cited on page 1.]

[2] H. Vieira, N. Costa, L. P. Coelho, and J. Alves, "Real-time Modeling of Abnormal Physiological Signals in a Phantom for Bioengineering Education," in *2020 IEEE Global Engineering Education Conference (EDUCON)*, Apr. 2020, pp. 1206–1211, iSSN: 2165-9567. [Cited on page 1.]

[3] H. Vieira, N. Costa, J. F. A. Alves, and L. P. Coelho, "Simulation of Abnormal Physiological Signals in a Phantom for Bioengineering Education," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 16, no. 14, p. 107, Nov. 2020. [Online]. Available: https://online-journals.org/index.php/i-joe/article/view/16941 [Cited on pages 1 and 10.]

[4] I.-I. d. Telecomunicações, "DigiScope2 making strides in cardiac pathology screening with new AI technology," Nov. 2020. [Online]. Available: https://www.it.pt/News/NewsPost/4627 [Cited on page 1.]

[5] T. Lisboa, "BITalino à conquista do mundo," Jun. 2017. [Online]. Available: https://tecnico.ulisboa.pt/pt/noticias/bitalino-a-conquista-do-mundo/ [Cited on page 1.]

[6] Y. Zeng, S. Yang, X. Yu, W. Lin, W. Wang, J. Tong, and S. Xia, "A multimodal parallel method for left ventricular dysfunction identification based on phonocardiogram and electrocardiogram signals synchronous analysis," *Mathematical Biosciences and Engineering*, vol. 19, no. 9, pp. 9612–9635, 2022. [Online]. Available: http://www.aimspress.com/article/doi/10.3934/mbe.2022447 [Cited on pages 2, 34, and 37.]

[7] J. E. Hall, *Guyton and Hall Textbook of Medical Physiology - Elsevier eBook on VitalSource*, 12th ed. Saunders, Jun. 2010. [Cited on page 6.]

[8] "File:Diagram of the human heart (cropped).svg - Wikimedia Commons." [Online]. Available: https://commons.wikimedia.org/wiki/File:Diagram_of_the_human_heart_(cropped).svg [Cited on pages ix and 6.]

[9] J. R. Mitchell and J.-J. Wang, "Expanding application of the Wiggers diagram to teach cardiovascular physiology," *Advances in Physiology Education*, vol. 38, no. 2, p. 170, Jun. 2014, publisher: American Physiological Society. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4056172/ [Cited on pages ix and 7.]

[10] J. Oliveira, "Subject-driven supervised and unsupervised Hidden Markov Models for heart sound segmentation in real noisy environments," Ph.D. dissertation, 2018. [Cited on pages 8 and 9.]

[11] T. H. Chowdhury, K. N. Poudel, and Y. Hu, "Time-Frequency Analysis, Denoising, Compression, Segmentation, and Classification of PCG Signals," *IEEE Access*, vol. 8, pp. 160 882–160 890, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9183915/ [Cited on pages ix, 9, and 21.]

[12] "What is Apical Pulse: Definition and Process of Measurement." [Online]. Available: https://www.topregisterednurse.com/apical-pulse-definition-process-measurement/ [Cited on pages ix and 9.]

[13] P. Ganesan, M. Sterling, S. Ladavich, and B. Ghoraani, "Computer-Aided Clinical Decision Support Systems for Atrial Fibrillation," in *Computer-aided Technologies - Applications in Engineering and Medicine*, R. Udroiu, Ed. InTech, Dec. 2016. [Online]. Available: http://www.intechopen.com/books/computer-aided-technologies-applications-in-engineering-and-medicine/computer-aided-clinical-decision-support-systems-for-atrial-fibrillation [Cited on pages ix and 10.]

[14] C. b. Agateller, "Schematic diagram of normal sinus rhythm for a human heart as seen on ECG (with English labels)." Jan. 2007. [Online]. Available: https://commons.wikimedia.org/wiki/File:SinusRhythmLabels.svg [Cited on pages ix and 12.]

[15] T. H. Chowdhury, "Application of Signal Processing and Deep Hybrid Learning in Phonocardiogram and Electrocardiogram Signals to Detect Early Stage Heart Diseases," Ph.D. dissertation, Middle Tennessee State University, 2022. [Cited on page 13.]

[16] "Classification of Heart Sound Recordings: The PhysioNet/Computing in Cardiology Challenge 2016 v1.0.0." [Online]. Available: https://physionet.org/content/challenge-2016/1.0.0/ [Cited on pages ix and 13.]

[17] A. Aghabayli, "Machine Learning Applied to Building Information Models," Ph.D. dissertation, Universidade do Minho, 2021. [Cited on pages 14 and 15.]

[18] N. E. Sahla, "A Deep Learning Prediction Model for Object Classification," Ph.D. dissertation, Delft University of Technology, 2018. [Cited on pages ix, 14, and 22.]

[19] IBM, "What is Supervised Learning?" Jun. 2021. [Online]. Available: https://www.ibm.com/cloud/learn/supervised-learning [Cited on page 15.]

[20] N. Kehtarnavaz, "CHAPTER 7 - Frequency Domain Processing," in *Digital Signal Processing System Design (Second Edition)*, N. Kehtarnavaz, Ed. Burlington: Academic Press, Jan. 2008, pp. 175–196. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123744906000076 [Cited on page 16.]

[21] T. M. , "Scalogram Computation in Signal Analyzer - MATLAB & Simulink." [Online]. Available: https://www.mathworks.com/help/signal/ug/scalogram-computation-in-signal-analyzer.html [Cited on pages 16 and 17.]

[22] H. S. Kvalnes and M. S. Lysø, "Comparison of Wavelet Transforms and STFTs in Classification of Outdoor Noise," Master's thesis, NTNU, 2020, accepted: 2021-09-15T16:58:10Z ISSN: 3452-7877. [Online]. Available: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2778110 [Cited on pages ix and 16.]

[23] T. M. d. C. Pereira, "Biometric authentication and identification through electrocardiogram signals," Master's thesis, 2021, accepted: 2022-02-17T18:31:07Z. [Online]. Available: https://repositorio.ul.pt/handle/10451/51392 [Cited on page 17.]

[24] T. M. , Inc., "Scale to frequency - MATLAB scal2frq." [Online]. Available: https://www.mathworks.com/help/wavelet/ref/scal2frq.html [Cited on page 17.]

[25] "Available Wavelet Families · ContinuousWavelets.jl." [Online]. Available: https://docs.juliahub.com/ContinuousWavelets/JbYB7/1.0.0/coreType/ [Cited on pages ix and 17.]

[26] J. L. C. Loong, K. S. Subari, M. K. Abdullah, N. N. Ahmad, and RosliBesar, "Comparison Of Mfcc And Cepstral Coefficients As A Feature Set For Pcg Biometric Systems," Aug. 2010, publisher: Zenodo. [Online]. Available: https://zenodo.org/record/1071702 [Cited on page 18.]

[27] D. Caschili, "Optimization of CNN-Based Object Detection Algorithms for Embedded Systems," p. 57, 2019. [Cited on pages 18 and 20.]

[28] S. Saxena, "Artificial Neuron Networks(Basics) | Introduction to Neural Networks," Oct. 2017. [Online]. Available: https://becominghuman.ai/artificial-neuron-networks-basics-introduction-to-neural-networks-3082f1dcca8c [Cited on pages ix and 19.]

[29] L. Danqing, "A Practical Guide to ReLU. Start using and understanding ReLU… | by Danqing Liu | Medium," Nov. 2017. [Online]. Available: https://medium.com/@danqing/a-practical-guide-to-relu-b83ca804f1f7 [Cited on pages ix and 19.]

[30] "Deep Learning Tutorial." [Online]. Available: https://www.tutorialkart.com/deep-learning/ [Cited on pages ix and 20.]

[31] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. da Silva, "A comparative analysis of object detection metrics with a companion open-source toolkit," *Electronics*, vol. 10, no. 3, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/3/279 [Cited on page 21.]

[32] R. Shanmugamani, *Deep Learning for Computer Vision: Expert techniques to train advanced neural networks using TensorFlow and Keras*. Birmingham Mumbai: Packt Publishing, Jan. 2018. [Cited on pages ix and 21.]

[33] Nomidl, "Difference between Sigmoid and Softmax activation function?" Apr. 2022. [Online]. Available: https://www.nomidl.com/deep-learning/what-is-the-difference-between-sigmoid-and-softmax-activation-function/ [Cited on pages ix and 22.]

[34] J. Brownlee, "A Gentle Introduction to Transfer Learning for Deep Learning," Dec. 2017. [Online]. Available: https://machinelearningmastery.com/transfer-learning-for-deep-learning/ [Cited on page 23.]

[35] R. S. de Gélis, "Transfer learning techniques in time series analysis," Ph.D. dissertation, KTH ROYAL INSTITUTE OF TECHNOLOGY, STOCKHOLM, SWEDEN, 2022. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1647721/FULLTEXT01.pdf [Cited on page 23.]

[36] J. Boit, "The Effectiveness of Transfer Learning Systems on Medical Images," Ph.D. dissertation, Dakota State University, Dakota, 2020. [Cited on page 23.]

[37] K. Team, "Keras documentation: Transfer learning & fine-tuning." [Online]. Available: https://keras.io/guides/transfer_learning/ [Cited on page 23.]

[38] A. Suresh, "What is a confusion matrix?. Everything you Should Know about... | by Anuganti Suresh | Analytics Vidhya | Medium," Nov. 2020. [Online]. Available: https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5 [Cited on pages ix and 24.]

[39] S. Narkhede, "Understanding AUC - ROC Curve | by Sarang Narkhede | Towards Data Science," Jun. 2018. [Online]. Available: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 [Cited on page 25.]

[40] J. Brownlee, "How to Use ROC Curves and Precision-Recall Curves for Classification in Python," Aug. 2018. [Online]. Available: https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/ [Cited on page 25.]

[41] Z. H. Hoo, J. Candlish, and D. Teare, "What is an ROC curve?" *Emergency Medicine Journal*, vol. 34, no. 6, pp. 357–359, Jun. 2017. [Online]. Available: https://emj.bmj.com/lookup/doi/10.1136/emermed-2017-206735 [Cited on pages ix and 26.]

[42] PubMed, "PubMed." [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/ [Cited on page 27.]

[43] R. Hettiarachchi, U. Haputhanthri, K. Herath, H. Kariyawasam, S. Munasinghe, K. Wickramasinghe, D. Samarasinghe, A. De Silva, and C. U. S. Edussooriya, "A

Novel Transfer Learning-Based Approach for Screening Pre-Existing Heart Diseases Using Synchronized ECG Signals and Heart Sounds," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. Daegu, Korea: IEEE, May 2021, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9401093/ [Cited on pages 30, 34, 38, 43, 49, and 72.]

[44] F. Chakir, A. Jilbab, C. Nacir, and A. Hammouch, "Recognition of cardiac abnormalities from synchronized ECG and PCG signals," *Physical and Engineering Sciences in Medicine*, vol. 43, no. 2, pp. 673–677, Jun. 2020. [Online]. Available: https://link.springer.com/10.1007/s13246-020-00875-2 [Cited on pages 30, 33, 37, 72, and 73.]

[45] M. E. EL-Bouridy and A. S. EL-Batouty, "An Intelligent High Accuracy & hybrid Identification for Heart diseases Diagnosis," in *2021 International Telecommunications Conference (ITC-Egypt)*. Alexandria, Egypt: IEEE, Jul. 2021, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9513892/ [Cited on pages 30 and 35.]

[46] J. R. Balbin, A. I. T. Yap, B. D. Calicdan, and L. A. M. Bernabe, "Arrhythmia Detection using Electrocardiogram and Phonocardiogram Pattern using Integrated Signal Processing Algorithms with the Aid of Convolutional Neural Networks," in *2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACIS)*. Shah Alam, Malaysia: IEEE, Jun. 2021, pp. 146–151. [Online]. Available: https://ieeexplore.ieee.org/document/9495913/ [Cited on pages 30 and 35.]

[47] H. Zhang, X. Wang, C. Liu, Y. Li, Y. Liu, Y. Jiao, T. Liu, H. Dong, and J. Wang, "Discrimination of Patients with Varying Degrees of Coronary Artery Stenosis by ECG and PCG Signals Based on Entropy," *Entropy*, vol. 23, no. 7, p. 823, Jun. 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/7/823 [Cited on pages 30 and 35.]

[48] H. Li, X. Wang, C. Liu, Y. Wang, P. Li, H. Tang, L. Yao, and H. Zhang, "Dual-Input Neural Network Integrating Feature Extraction and Deep Learning for Coronary Artery Disease Detection Using Electrocardiogram and Phonocardiogram," *IEEE Access*, vol. 7, pp. 146 457–146 469, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8846698/ [Cited on pages 31 and 36.]

[49] X.-C. Li, X.-H. Liu, L.-B. Liu, S.-M. Li, Y.-Q. Wang, and R. H. Mead, "Evaluation of left ventricular systolic function using synchronized analysis of heart sounds and the electrocardiogram," *Heart Rhythm*, vol. 17, no. 5, pp. 876–880, May 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1547527120300771 [Cited on pages 31 and 36.]

[50] S. A. Singh, S. A. Singh, N. D. Devi, and S. Majumder, "Heart Abnormality Classification Using PCG and ECG Recordings," *Computación y Sistemas*, vol. 25, no. 2, May 2021. [Online]. Available: https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3447 [Cited on pages 32, 36, 43, and 47.]

[51] P. Yupapin, Wardkein, P. Yupapin, Phanphaisarn, Koseeyaporn, Roeksabutr, Roeksabutr, Wardkein, and Koseeyapon, "Heart detection and diagnosis based on ECG and EPCG relationships," *Medical Devices: Evidence and Research*, p. 133, Aug. 2011. [Online]. Available: http://www.dovepress.com/heart-detection-and-diagnosis-based-on-ecg-and-epcg-relationships-peer-reviewed-article-MDER [Cited on pages 32 and 36.]

[52] H. Li, X. Wang, C. Liu, P. Li, and Y. Jiao, "Integrating multi-domain deep features of electrocardiogram and phonocardiogram for coronary artery disease detection," *Computers in Biology and Medicine*, vol. 138, p. 104914, Nov. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0010482521007083 [Cited on pages 32 and 36.]

[53] Y. Li, X. Wang, C. Liu, L. Li, C. Yan, L. Yao, and P. Li, "Variability of Cardiac Electromechanical Delay With Application to the Noninvasive Detection of Coronary Artery Disease," *IEEE Access*, vol. 7, pp. 53 115–53 124, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8692727/ [Cited on pages 33 and 37.]

[54] M. Klum, M. Urban, T. Tigges, A.-G. Pielmus, A. Feldheiser, T. Schmitt, and R. Orglmeister, "Wearable Cardiorespiratory Monitoring Employing a Multimodal Digital Patch Stethoscope: Estimation of ECG, PEP, LVET and Respiration Using a 55 mm Single-Lead ECG and Phonocardiogram," *Sensors*, vol. 20, no. 7, p. 2033, Apr. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/7/2033 [Cited on pages 33 and 37.]

[55] J. S. F. Botha, C. Scheffer, W. W. Lubbe, and A. F. Doubell, "Autonomous auscultation of the human heart employing a precordial electro-phonocardiogram and ensemble empirical mode decomposition," *Australasian Physical & Engineering Sciences in Medicine*, vol. 33, no. 2, pp. 171–183, Jun. 2010. [Online]. Available: http://link.springer.com/10.1007/s13246-010-0021-9 [Cited on page 38.]

[56] S.-Y. Lee, P.-W. Huang, J.-R. Chiou, C. Tsou, Y.-Y. Liao, and J.-Y. Chen, "Electrocardiogram and Phonocardiogram Monitoring System for Cardiac Auscultation," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, no. 6, pp. 1471–1482, Dec. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8876684/

[57] J. Oliveira, C. Sousa, and M. T. Coimbra, "Coupled hidden Markov model for automatic ECG and PCG segmentation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. New Orleans, LA: IEEE, Mar. 2017, pp. 1023–1027. [Online]. Available: http://ieeexplore.ieee.org/document/7952311/

[58] N. F. Hikmah, A. Arifin, T. A. Sardjono, and E. A. Suprayitno, "A signal processing framework for multimodal cardiac analysis," in *2015 International Seminar on Intelligent Technology and Its Applications (ISITIA)*. Surabaya, Indonesia: IEEE, May 2015, pp. 125–130. [Online]. Available: http://ieeexplore.ieee.org/document/7219966/ [Cited on page 38.]

[59] P. McSharry, G. Clifford, L. Tarassenko, and L. Smith, "A dynamical model for generating synthetic electrocardiogram signals," *IEEE Transactions on Biomedical Engineering*, vol. 50, no. 3, pp. 289–294, Mar. 2003. [Online]. Available: http://ieeexplore.ieee.org/document/1186732/ [Cited on page 39.]

[60] A. Sbrollini, M. Morettini, I. Marcantoni, and L. Burattini, "Model-Based Estimation of Electrocardiographic QT Interval From Phonocardiographic Heart Sounds in Healthy Subjects," Dec. 2020. [Online]. Available: http://www.cinc.org/archives/2020/pdf/CinC2020-158.pdf [Cited on page 39.]

[61] C. Liu, D. Springer, Q. Li, B. Moody, R. A. Juan, F. J. Chorro, F. Castells, J. M. Roig, I. Silva, A. E. W. Johnson, Z. Syed, S. E. Schmidt, C. D. Papadaniil, L. Hadjileontiadis, H. Naseri, A. Moukadem, A. Dieterlen, C. Brandt, H. Tang, M. Samieinasab, M. R. Samieinasab, R. Sameni, R. G. Mark, and G. D. Clifford, "An

open access database for the evaluation of heart sound algorithms," *Physiological Measurement*, vol. 37, no. 12, pp. 2181–2213, Dec. 2016. [Online]. Available: https://iopscience.iop.org/article/10.1088/0967-3334/37/12/2181 [Cited on pages xi, 39, 40, 46, and 47.]

[62] A. Kazemnejad, P. Gordany, and R. Sameni, "EPHNOGRAM: A Simultaneous Electrocardiogram and Phonocardiogram Database," 2021, version Number: 1.0.0 Type: dataset. [Online]. Available: https://physionet.org/content/ephnogram/1.0.0/ [Cited on pages 39, 40, and 41.]

[63] K. Shi, S. Schellenberger, C. Will, T. Steigleder, F. Michler, J. Fuchs, R. Weigel, C. Ostgathe, and A. Koelpin, "A dataset of radar-recorded heart sounds and vital signs including synchronised reference sensor signals," *Scientific Data*, vol. 7, no. 1, p. 50, Dec. 2020. [Online]. Available: http://www.nature.com/articles/s41597-020-0390-1 [Cited on pages 39, 41, and 42.]

[64] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000 (June 13), circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215. [Cited on pages 39 and 41.]

[65] "PhysioNet Challenges." [Online]. Available: https://physionet.org/about/challenge/ [Cited on page 39.]

[66] "PhysioNet/CinC Challenge 2016: Training Sets." [Online]. Available: https://archive.physionet.org/pn3/challenge/2016/ [Cited on page 40.]

[67] C. Will, "Radar-Based Heart Sound Detection," *SCiEntifiC REPOrTS*, p. 14, 2018. [Cited on page 41.]

[68] S. Schellenberger, "GUARDIAN Vital Sign Data," Aug. 2019, publisher: figshare. [Online]. Available: https://figshare.com/collections/GUARDIAN_Vital_Sign_Data/4633958 [Cited on page 42.]

[69] E. Benmalek, J. Elmhamdi, and A. Jilbab, "ECG scalogram classification with CNN micro-architectures," *Research on Biomedical Engineering*, vol. 38, no. 2, pp.

325–335, Jun. 2022. [Online]. Available: https://link.springer.com/10.1007/s42600-021-00188-7 [Cited on page 43.]

[70] O. Ozaltin and O. Yeniay, "A novel proposed CNN–SVM architecture for ECG scalograms classification," *Soft Computing*, vol. 27, no. 8, pp. 4639–4658, Apr. 2023. [Online]. Available: https://link.springer.com/10.1007/s00500-022-07729-x

[71] Y.-H. Byeon, S.-B. Pan, and K.-C. Kwak, "Intelligent Deep Models Based on Scalograms of Electrocardiogram Signals for Biometrics," *Sensors*, vol. 19, no. 4, p. 935, Feb. 2019. [Online]. Available: http://www.mdpi.com/1424-8220/19/4/935

[72] J. Gelpud, S. Castillo, M. Jojoa, B. Garcia-Zapirain, W. Achicanoy, and D. Rodrigo, "Deep Learning for Heart Sounds Classification Using Scalograms and Automatic Segmentation of PCG Signals," in *Advances in Computational Intelligence*, I. Rojas, G. Joya, and A. Català, Eds. Cham: Springer International Publishing, 2021, vol. 12861, pp. 583–596, series Title: Lecture Notes in Computer Science. [Online]. Available: https://link.springer.com/10.1007/978-3-030-85030-2_48 [Cited on page 43.]

[73] M. R. Dammeyer, D. Engineer, D. Hansen, and D. Engineer, "The Impact of ECG Sensor Design on Signal Noise," p. 6. [Cited on page 44.]

[74] G. Clifford, C. Liu, B. Moody, L.-w. Lehman, I. Silva, Q. Li, A. Johnson, and R. Mark, "AF Classification from a Short Single Lead ECG Recording: the Physionet Computing in Cardiology Challenge 2017," Sep. 2017. [Online]. Available: http://www.cinc.org/archives/2017/pdf/065-469.pdf [Cited on pages xi and 45.]

[75] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, no. 23, Jun. 2000. [Online]. Available: https://www.ahajournals.org/doi/10.1161/01.CIR.101.23.e215 [Cited on pages 45 and 46.]

[76] Physionet, "AF Classification from a Short Single Lead ECG Recording: The PhysioNet/Computing in Cardiology Challenge 2017 v1.0.0," 2017. [Online]. Available: https://physionet.org/content/challenge-2017/1.0.0/ [Cited on pages ix and 46.]

[77] K. Ali and C. E. Hughes, "A Unified Transformer-based Network for multimodal Emotion Recognition," Aug. 2023, arXiv:2308.14160 [cs]. [Online]. Available: http://arxiv.org/abs/2308.14160 [Cited on page 47.]

[78] S. A. Singh, T. G. Meitei, and S. Majumder, "Short PCG classification based on deep learning," in *Deep Learning Techniques for Biomedical and Health Informatics*. Elsevier, 2020, pp. 141–164. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780128190616000069 [Cited on page 47.]

[79] D. , "Signal filtering, Signal suppression, Signal processing | Dewesoft." [Online]. Available: https://training.dewesoft.com/online/course/filters [Cited on page 47.]

[80] A. Zyout, H. Alquran, W. A. Mustafa, and A. M. Alqudah, "Advanced Time-Frequency Methods for ECG Waves Recognition," *Diagnostics*, vol. 13, no. 2, p. 308, Jan. 2023. [Online]. Available: https://www.mdpi.com/2075-4418/13/2/308 [Cited on page 47.]

[81] L. G. Tereshchenko and M. E. Josephson, "Frequency content and characteristics of ventricular conduction," *Journal of Electrocardiology*, vol. 48, no. 6, pp. 933–937, Nov. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0022073615002915 [Cited on page 47.]

[82] S. A. Singh, S. Majumder, and M. Mishra, "Classification of short unsegmented heart sound based on deep learning," in *2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*. Auckland, New Zealand: IEEE, May 2019, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8826991/ [Cited on page 47.]

[83] K. N. Khan, F. A. Khan, A. Abid, T. Olmez, Z. Dokur, A. Khandakar, M. E. H. Chowdhury, and M. S. Khan, "Deep learning based classification of unsegmented phonocardiogram spectrograms leveraging transfer learning," *Physiological Measurement*, vol. 42, no. 9, p. 095003, Sep. 2021. [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-6579/ac1d59 [Cited on page 47.]

[84] H. Naseri, M. Homaeinezhad, and H. Pourkhajeh, "Noise/spike detection in phonocardiogram signal as a cyclic random process with non-stationary period interval," *Computers in Biology and Medicine*, vol. 43, no. 9, pp. 1205–1213,

Sep. 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0010482513001364 [Cited on page 47.]

[85] L. Zhong, X. Guo, A. Ji, and X. Ding, "A Robust Envelope Extraction Algorithm for Cardiac Sound Signal Segmentation," in *2011 5th International Conference on Bioinformatics and Biomedical Engineering*. Wuhan, China: IEEE, May 2011, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/document/5781655/ [Cited on page 49.]

[86] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd International Conference on Learning Representations (ICLR 2015)*, 2015, publisher: Computational and Biological Learning Society. [Online]. Available: https://ora.ox.ac.uk/objects/uuid:60713f18-a6d1-4d97-8f45-b60ad8aebbce [Cited on page 50.]

[87] M. u. Hassan, "VGG16 - Convolutional Network for Classification and Detection," Nov. 2018. [Online]. Available: https://neurohive.io/en/popular-networks/vgg16/ [Cited on page 50.]

[88] A. M. Al-Amaren, "Development of Deep Convolutional Neural Network Techniques for Edge Detection in Images," Ph.D. dissertation, Concordia University, Montréal, Québec, Canada, Jun. 2022. [Cited on page 51.]

[89] A. Kumar, "K-Fold Cross Validation - Python Example," Mar. 2023. [Online]. Available: https://vitalflux.com/k-fold-cross-validation-python-example/ [Cited on pages x and 53.]

[90] "GroupShuffleSplit." [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.model_selection.GroupShuffleSplit.html [Cited on page 53.]

[91] K. Team, "Keras documentation: About Keras." [Online]. Available: https://keras.io/about/ [Cited on page 55.]

[92] G. Lee, R. Gommers, F. Waselewski, K. Wohlfahrt, and A. O'Leary, "PyWavelets: A Python package for wavelet analysis," *Journal of Open Source Software*, vol. 4, no. 36, p. 1237, Apr. 2019. [Online]. Available: http://joss.theoj.org/papers/10.21105/joss.01237 [Cited on page 55.]

[93] J. McDermott, "Hands-on Transfer Learning with Keras and the VGG16 Model." [Online]. Available: https://www.learndatasci.com/tutorials/hands-on-transfer-learning-keras/ [Cited on pages x and 57.]

[94] J. Brownlee, "How to Choose Loss Functions When Training Deep Learning Neural Networks," Jan. 2019. [Online]. Available: https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/ [Cited on page 57.]

[95] ——, "Gentle Introduction to the Adam Optimization Algorithm for Deep Learning," Jul. 2017. [Online]. Available: https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/ [Cited on page 57.]

[96] K. Singh, "How to Improve Class Imbalance using Class Weights in Machine Learning?" Oct. 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/ [Cited on page 58.]

[97] J. Brownlee, "Overfitting and Underfitting With Machine Learning Algorithms," Mar. 2016. [Online]. Available: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/ [Cited on page 72.]

[98] J. Oliveira, F. Renna, P. D. Costa, M. Nogueira, C. Oliveira, C. Ferreira, A. Jorge, S. Mattos, T. Hatem, T. Tavares, A. Elola, A. B. Rad, R. Sameni, G. D. Clifford, and M. T. Coimbra, "The CirCor DigiScope Dataset: From Murmur Detection to Murmur Classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 6, pp. 2524–2535, Jun. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9658215/ [Cited on pages 72 and 75.]

[99] E. A. Perez Alday, A. Gu, A. J Shah, C. Robichaux, A.-K. Ian Wong, C. Liu, F. Liu, A. Bahrami Rad, A. Elola, S. Seyedi, Q. Li, A. Sharma, G. D. Clifford, and M. A. Reyna, "Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020," *Physiological Measurement*, vol. 41, no. 12, p. 124003, Dec. 2020. [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-6579/abc960 [Cited on page 75.]

[100] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, and F. Shen, "Image Data Augmentation for Deep Learning: A Survey," 2022, publisher: arXiv Version

Number: 1. [Online]. Available: https://arxiv.org/abs/2204.08610 [Cited on page 76.]

[101] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2015, publisher: arXiv Version Number: 1. [Online]. Available: https://arxiv.org/abs/1512.03385 [Cited on page 76.]

[102] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, Jun. 2015, pp. 1–9. [Online]. Available: http://ieeexplore.ieee.org/document/7298594/ [Cited on page 76.]

[103] C. S. Lima and M. J. Cardoso, "PHONOCARDIOGRAM SEGMENTATION BY USING HIDDEN MARKOV MODELS," 2007. [Cited on page 76.]