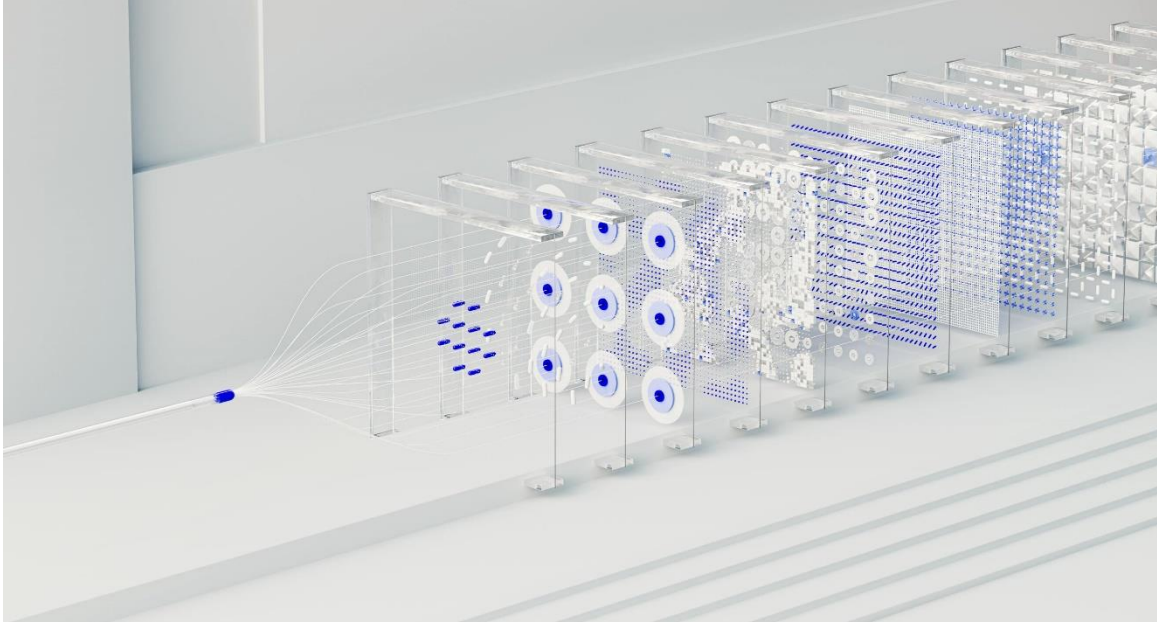
Deep learning for heart sounds and electrocardiogram signal analysis: On the impact of multimodal data for explainable AI

Bruno Filipe Ferreira da Assunção Oliveira
Master's Degree on Data Science
Department of Computer Science
Faculty of Sciences of the University of Porto
2024



Deep learning for heart sounds and electrocardiogram signal analysis: On the impact of multimodal data for explainable AI

Bruno Filipe Ferreira da Assunção Oliveira

Dissertation carried out as part of the Master's Degree on Data Science

Department of Computer Science

2024

Supervisor
Francesco Renna, Assistant Professor, Faculty of Sciences of the University of Porto

Co-supervisor
Miguel Coimbra, Full Professor, Faculty of Sciences of the University of Porto

Acknowledgements

I am deeply grateful to my family, whose unwavering support and encouragement have been the foundation of this journey. Special thanks to my sister and Paulo for their constant help and understanding, as well as to my aunt and uncle, who guided me through new challenges. I am especially thankful to my grandmother and mother for their endless love, care, and life lessons. Without them, I wouldn’t be here today.

I also want to extend my heartfelt thanks to my supervisors, Professor Miguel Coimbra and Francesco Renna, for their patience, guidance, and trust. Their insights and expertise were invaluable in helping me navigate even the most difficult moments. I am grateful for their availability and willingness to simplify complex ideas, making this achievement possible. My thanks also go to the entire FCUP working group, who supported me in both work and in light-hearted conversations, creating a friendly and welcoming environment. Additionally, I appreciate the INESC-TEC team for their knowledge-sharing through presentations and for organizing engaging team-building events and conferences.

Finally, I would like to thank my long-standing friends, whose friendships span over a quarter of a century, as well as my university friends. Whether near or far, they have always believed in me and have become my second family. I am forever grateful for their presence in my life.

This journey was not without its challenges, but it has been deeply rewarding. In the end, it all came together, as Confucius once said: "Our greatest glory is not in never falling, but in rising every time we fall."

Resumo

Doenças cardiovasculares continuam a ser uma das principais causas de mortalidade em todo o mundo, sublinhando a necessidade de métodos de diagnóstico mais eficientes, acessíveis e precisos. A integração de abordagens multimodais no rastreio coronário pode melhorar a precisão do diagnostico, reduzindo ao mesmo tempo os custos e aumentando a acessibilidade. Modelos de deep learning, aliados a técnicas de inteligência artificial explicável (XAI), podem ajudar significativamente os profissionais de saúde, fornecendo informações sobre o desempenho dos modelos e a sua tomada de decisão.

Neste trabalho, desenvolvemos cinco modelos distintos para a classificação binária (normal/anormal) de sons cardíacos (PCG) e eletrocardiogramas (ECG) sincronizados. Estes modelos incluem uma rede neural convolucional unidimensional (1D CNN) para ECG, uma CNN bidimensional (2D CNN) aplicada tanto a PCG como a ECG, e dois modelos híbridos multimodais: um modelo de fusão precoce (combinando as estruturas baseadas em 2D CNN para ECG e PCG) e um modelo de fusão tardia (integrando a 1D CNN para ECG e a 2D CNN para PCG). Os nossos resultados mostram que os modelos multimodais, em particular o modelo de fusão precoce, demonstraram maior estabilidade em todas as métricas, com menor variabilidade. O modelo de fusão precoce superou os modelos que se baseavam num único sinal, alcançando uma pontuação ROC-AUC de 0,840, em comparação com 0,793 para o 2D CNN de PCG e 0,829 para o 2D CNN de ECG.

Para interpretar melhor as previsões dos modelos, utilizámos duas metodologias de XAI: Gradient-weighted Class Activation Mapping (Grad-CAM) e Shapley Additive Explanations (SHAP). Estas técnicas destacaram características-chave nos sinais de ECG e PCG, como o complexo QRS, ondas P e T no ECG, e os sons cardíacos fundamentais S1 e S2 no PCG. O Grad-CAM identificou correlações mais amplas entre as características, enquanto o SHAP forneceu uma visão mais detalhada sobre a importância de características individuais.

Os resultados sublinham as vantagens da combinação de sinais de ECG e PCG para melhorar o rastreio de doenças cardiovasculares. Além disso, os insights complementares oferecidos pelo Grad-CAM e SHAP aumentam a interpretabilidade do

processo de tomada de decisão do modelo, abrindo caminho para a futura integração destes modelos em ambientes clínicos e telemedicina.

Palavras-chave: Eletrocardiograma (ECG), Fonocardiograma (PCG), Redes Neurais Convolucionais (CNN), Modelos Multimodais, Gradient-weighted Class Activation Mapping (Grad-CAM), Shapley Additive Explanations (SHAP).

Abstract

Cardiovascular diseases remain one of the leading causes of mortality worldwide, underscoring the need for more efficient, accessible, and accurate diagnostic methods. The integration of multimodal approaches in coronary screening can enhance diagnostic precision while reducing costs and expanding accessibility. Deep learning models, coupled with explainable artificial intelligence (XAI) techniques, can significantly assist clinicians in this process by providing insights into model performance and decision-making.

In this work, we developed five distinct models for binary classification (normal/abnormal) of synchronized heart sounds (PCG) and electrocardiograms (ECG). These models include a one-dimensional convolutional neural network (1D CNN) for ECG, a two-dimensional CNN (2D CNN) applied to both PCG and ECG, and two multimodal hybrid models: an early fusion model (combining 2D CNN backbones for ECG and PCG) and a late fusion model (integrating the 1D CNN for ECG and 2D CNN for PCG). Our findings show that multimodal models, particularly the early fusion model, demonstrated greater stability across all metrics, with reduced variability. The early fusion model outperformed models that relied on either signal independently, achieving a higher ROC-AUC score of 0.840 compared to 0.793 for PCG 2D CNN and 0.829 for ECG 2D CNN.

To further interpret model predictions, we employed two XAI methodologies: Gradient-weighted Class Activation Mapping (Grad-CAM) and Shapley Additive Explanations (SHAP). These techniques highlighted key features in both ECG and PCG signals, such as the QRS complex, P and T waves in ECG, and the fundamental heart sounds S1 and S2 in PCG. Grad-CAM identified broader correlations between features, while SHAP provided more detailed insights into individual feature importance.

The results underscore the advantages of combining ECG and PCG signals for improved cardiovascular disease screening. Additionally, the complementary insights offered by Grad-CAM and SHAP enhance the interpretability of the model's decision-making process, paving the way for future integration of such models in clinical settings and telemedicine.

Keywords: Electrocardiogram (ECG), Phonocardiogram (PCG), Convolutional Neural Networks (CNN), Multimodal models, Gradient-weighted Class Activation Mapping (Grad-CAM), Shapley Additive Explanations (SHAP).

Table of Contents

List of Tables

List of Figures

List of Abbreviations

FCUP	FACULTY OF SCIENCES OF THE UNIVERSITY OF PORTO
UP	UNIVERSITY OF PORTO
CVD	CARDIOVASCULAR DISEASES
CT	COMPUTED TOMOGRAPHY
MRI	MAGNETIC RESONANCE IMAGING
ECG	ELECTROCARDIOGRAM
PCG	PHONOCARDIOGRAM
XAI	EXPLAINABLE ARTIFICIAL INTELLIGENCE
GRAD-CAM	GRADIENT-WEIGHTED CLASS ACTIVATION MAPPING
SHAP	SHAPLEY ADDITIVE EXPLANATIONS
SA	SINOATRIAL
AV	ATRIOVENTRICULAR
S1	FIRST HEART SOUND
S2	SECOND HEART SOUND
S3	THIRD HEART SOUND
S4	FOURTH HEART SOUND
CNN	CONVOLUTIONAL NEURAL NETWORKS
RELU	RECTIFIED LINEAR UNIT
ADAM	ADAPTIVE MOMENT ESTIMATION
CAD	CORONARY ARTERY DISEASE
BI-GRU	BIDIRECTIONAL GATED RECURRENT UNIT
PHY16	PHYSIONET 2016
CWT	CONTINUOUS WAVELET TRANSFORM
AUC	AREA UNDER THE CURVE
LSTM	LONG SHORT-TERM MEMORY
SVM	SUPPORT-VECTOR MACHINE
GA	GENETIC ALGORITHM
MFCC	MEL-FREQUENCY CEPSTRAL COEFFICIENT
HVDS	HEART VALVE DEFECTS
GAP	GLOBAL AVERAGE POOLING
VIT	VISION TRANSFORMERS
MHSA	MULTI HEAD SELF-ATTENTION
MLP	MULTILAYER PERCEPTRON
GELU	GAUSSIAN ERROR LINEAR UNIT

CMR	CROSS-MODALITY REGION-AWARE
MFO	MULTI-SCALE FEATURE OPTIMIZATION
DFA	DENSE FEATURE AGGREGATION
MR-NET	MULTI-BRANCH RESIDUAL NETWORK ARCHITECTURE
SE	SQUEEZE-AND-EXCITATION
PCA	PRINCIPAL COMPONENT ANALYSIS
SNR	NORMAL SINUS RHYTHM
AF	ATRIAL FIBRILLATION
IAVB	FIRST-DEGREE ATRIOVENTRICULAR BLOCK
LBBB	LEFT BUNDLE BRANCH BLOCK
RBBB	RIGHT BUNDLE BRANCH BLOCK
PAC	PREMATURE ATRIAL CONTRACTION
PVC	PREMATURE VENTRICULAR CONTRACTION
STD	ST-SEGMENT DEPRESSION
STE	ST-SEGMENT ELEVATION
AS	AORTIC STENOSIS
MR	MITRAL REGURGITATION
MS	MITRAL STENOSIS
CHD	CONGENITAL HEART DISEASE
LIME	LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS
LOGMT	LOG-MEL TRANSFORMATION
STFT	SHORT-TIME FOURIER TRANSFORMATION
HHT	HILBERT–HUANG TRANSFORMATION
ST	STOCKWELL TRANSFORM
MDN	MULTI-SCALE DENSE NETWORK
MARNN	MULTI-HEAD RECURRENT NEURAL NETWORK
MA	MULTI-HEAD SELF-ATTENTION MECHANISM
ACC	ACCURACY
SEN	SENSITIVITY
SPE	SPECIFICITY
PREC	PRECISION
MVP	MITRAL VALVE PROLAPSE
AD	AORTIC DISEASE
MPC	MISCELLANEOUS PATHOLOGICAL CONDITIONS

MAA	MAXIMUM ABSOLUTE AMPLITUDE
DCT	DISCRETE COSINE TRANSFORM
DWT	DISCRETE WAVELET TRANDFORM
ROC	RECEIVER OPERATING CHARACTERISTIC
AUC	AREA UNDER THE CURVE
EF	EARLY FUSION
LF	LATE FUSION
TP	TRUE POSITIVES
TN	TRUE NEGATIVES
FP	FALSE POSITIVES
FN	FALSE NEGATIVES
TPR	TRUE POSITIVE RATE
FPR	FALSE POSITIVE RATE
UQ	UNCERTAINTY QUANTIFICATION

1. Introduction

1.1. Overview

Cardiovascular diseases (CVD) were the leading cause of death globally in 2019, accounting for 32% of all fatalities. Notably, 75% of these CVD-related deaths occurred in low- and middle-income countries, primarily due to limited access to primary healthcare, essential services, and specialized medical equipment [1]. Advanced imaging techniques like Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and echocardiography provide detailed insights into cardiac function and structure. However, their high costs, advanced equipment, and the need for specialized personnel restrict their use in lower-income regions, making them less feasible for initial screenings [2]. In contrast, traditional methods such as cardiac auscultation and electrocardiogram (ECG) analysis are more accessible and practical for preliminary evaluations. Cardiac auscultation, which involves listening to heart sounds with a stethoscope or observing their graphical representation through a phonocardiogram (PCG), is a quick, non-invasive method that provides valuable information about heart function and structure. It allows clinicians to assess heart rate, rhythm, and valve function, offering insights into the mechanical aspects of the heart [3]. Similarly, the ECG records the electrical activity of the heart, making it a powerful, non-invasive, and cost-effective tool for evaluating the heart's electrical condition [4]. Combining the analysis of both PCG and ECG signals during routine check-ups enhances the accuracy of heart health assessments by integrating complementary information from these two methods, ultimately improving early detection and screening outcomes [5].

1.2. Motivation

While advanced signal processing and machine learning techniques have shown promise in automatically diagnosing diseases from both PCG and ECG signals, research has predominantly focused on using these signals separately [6], [7]. However, there is a growing trend towards combining PCG and ECG signals, reflecting a shift towards more integrated diagnostic approaches [8], [9].

Despite the impressive performance of deep learning models in various predictive and classification tasks, they are often criticized for their lack of transparency. In medical applications, understanding how a model arrives at its decisions is crucial, as healthcare

professionals often require clear explanations to trust and utilize these predictions effectively [10]. As a result, Explainable Artificial Intelligence (XAI) has emerged as an increasingly important area of research, particularly in the context of biomedical signals [11].

The primary goal of this work is to explore the benefits of combining synchronous ECG and PCG data, utilizing deep learning methodologies for the simultaneous classification of these signals, specifically in distinguishing between normal and abnormal cases. Given the scarcity of publicly available extensive datasets containing synchronous ECG and PCG data, this research addresses a significant gap in the field. Additionally, various XAI techniques are applied to gain insights into the models' behaviour, enabling healthcare professionals to build the trust needed to inform and support their clinical decisions. The use of multimodal data further enhances the explainability of the models, as it provides a more comprehensive understanding of the relationship between the two signals, leading to richer and more reliable interpretations of the model’s decisions.

1.3. Objectives

For this thesis scope, these are the outlined objectives:

- Analyse the current state on multimodal deep learning models that apply synchronous ECG and PCG for heart abnormality detection.
- Analyse the current XAI techniques employed in ECG or PCG deep learning classifiers.
- Develop new deep learning algorithms that can classify synchronous ECG and PCG signals, with ablation studies to compare the gain in performance and demonstrate/discuss if multimodality is preferred for clinical evaluation and application instead of single modularity.
- Apply different XAI techniques and discuss the differences and advantages of each one of them.
- Discuss and propose improvements of the current work, to strengthen the transparency and trustiness of the developed models.

1.4. Contributions

This work yielded some important contributions to the fields of cardiac signal classification and explainability:

- Analysis of the up to date state-of-the-art in deep learning multimodal ECG and PCG classification.
- Implementation of two XAI techniques, namely Grad-CAM and SHAP, for multimodal classification, along with a comparative analysis of the results across single modality, early, and late multimodal models, to evaluate gradient activations in each approach.
- An accepted article in a national conference about the development of a multimodal model for synchronous ECG and PCG classification: Oliveira, Bruno, et al. “Multimodal Deep Learning for Synchronous Heart Sounds and Electrocardiogram Classification” RECPAD Portuguese Conference on Pattern Recognition. (RECPAD 2023).
- The application of Grad-CAM on the Multimodal model for synchronous ECG and PCG classification, as well as demonstrating the performance difference between an early and late fusion model, led to an article in an international conference: Oliveira, Bruno, et al. “Explainable Multimodal Deep Learning for Heart Sounds and Electrocardiogram Classification” 2024 46th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC). IEEE, 2024.
- An accepted article in a national conference about the development of an explainable multimodal model for synchronous ECG and PCG classification, comparing Grad-CAM and SHAP techniques: Oliveira, Bruno, et al. “Explainable Multimodal Deep Learning for Heart Sounds and Electrocardiogram Classification: A comparison between SHAP and Grad-CAM” RECPAD Portuguese Conference on Pattern Recognition. (RECPAD 2024).

2. Background

This section provides essential background on the characteristics of physiological bio-signals, focusing on two common non-invasive cardiac signal acquisition methods: electrocardiograms (Chapter 2.1.1) and phonocardiograms (Chapter 2.1.2). It also explores the benefits of combining these two modalities (Chapter 2.1.3). Finally, an introduction to Convolutional Neural Networks (CNNs) and their intricacies is presented in Chapter 2.2.

2.1. Cardiac Physiology and Signal Interpretation

This subchapter offers a detailed examination of the heart's function, the techniques used to capture its electrical and mechanical activity, and the relationship between these signals and the heart's mechanical actions. The synchronization of ECG and PCG signals with specific cardiac events is essential for understanding the full scope of heart function and pathology.

2.1.1. Electrocardiogram Signals

Electrocardiograms (ECGs) are crucial diagnostic tools used to assess the electrical activity of the heart. By recording the electrical impulses that regulate heartbeats, ECGs provide insights into the heart's rhythm, rate, and overall function, making them essential for diagnosing a variety of cardiac conditions, including arrhythmias, myocardial infarctions, and other heart disorders [12].

A solid understanding of cardiac anatomy is fundamental to interpreting ECGs. The heart consists of four chambers: the left and right atria and the left and right ventricles. These chambers are separated by valves that ensure unidirectional blood flow. The sinoatrial (SA) node, located in the right atrium, acts as the heart's natural pacemaker, initiating electrical impulses that propagate through the atria to the atrioventricular (AV) node and then to the ventricles via the His-Purkinje system (Figure 1 to inspect for their locations and relationships with ECG). The anatomical structure of the heart influences the orientation of the electrical vectors recorded by ECGs. For instance, the depolarization wavefront travels from the atria to the ventricles, producing characteristic waves on the ECG trace. These anatomical and physiological aspects are crucial for understanding the origins of the different ECG components [13].

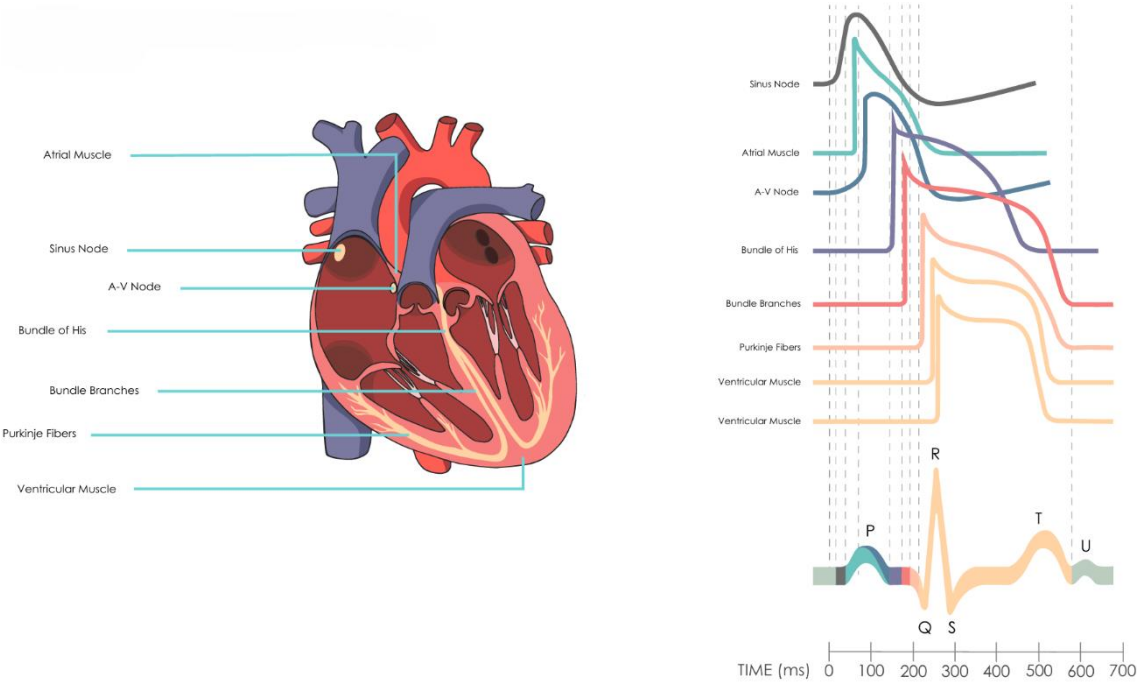


Figure 1 - Illustration of the heart's conduction system, key heartbeat complexes, and corresponding ECG waveform. Adapted from [14].

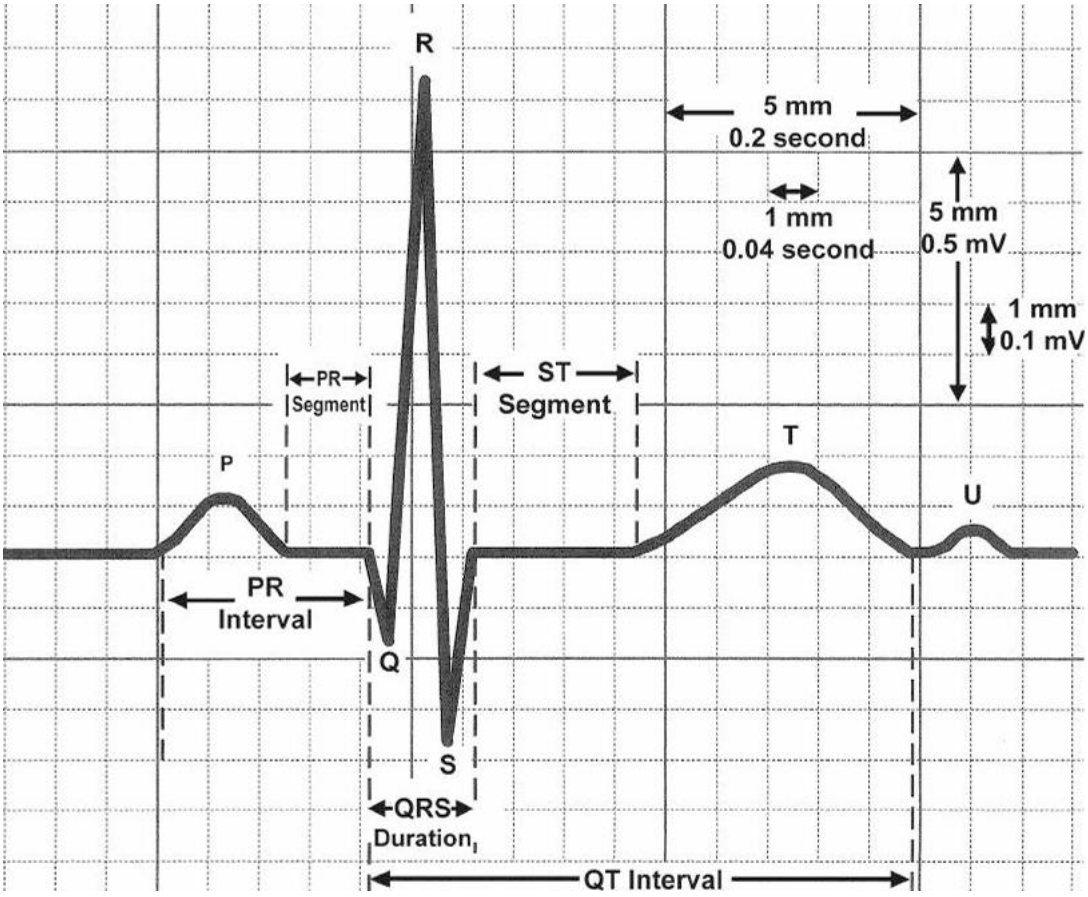


Figure 2 - Representation of ECG Waves, Intervals, and Segments. Adapted from [15].

The heart's electrophysiology is driven by the coordinated activity of ion channels within cardiac cells. These channels regulate the flow of ions like sodium, potassium, and calcium, which generate the electrical impulses necessary for heartbeats. The depolarization phase is initiated by the influx of sodium ions through fast sodium channels, leading to the rapid upstroke of the action potential. Depolarization of the atria and ventricles is reflected in the P wave and QRS complex, respectively. After depolarization, the heart's cells must return to their resting state, a process regulated by the efflux of potassium ions and the gradual closure of calcium channels, a process known as repolarization. This phase is represented on the ECG as the T wave. The precise timing and coordination of these electrical events are crucial for maintaining normal heart rhythm. Disruptions in these processes can result in arrhythmias, often detected through abnormalities in the ECG waveform [16].

The ECG waveform is composed of several key complexes that correspond to different phases of the cardiac cycle (see Figure 2 for more details). The most prominent components include:

- P Wave: Represents atrial depolarization initiated by the SA node. It is the first wave seen on the ECG and typically displaying a smooth, rounded appearance.
- QRS Complex: Corresponds to ventricular depolarization, a rapid process that results in a sharp, spiked waveform. The QRS complex is crucial for assessing ventricular function and identifying conditions like bundle branch blocks or ventricular hypertrophy.
- T Wave: Represents ventricular repolarization, the process by which the ventricles reset electrically to prepare for the next contraction. T wave abnormalities can indicate electrolyte imbalances, ischemia, or other cardiac issues.
- U Wave: A less common component that may appear after the T wave, thought to be related to repolarization of the Purkinje fibers or the late phases of ventricular repolarization [15].

Beyond these waves, several important intervals and segments on the ECG provide further diagnostic information (see Figure 2 for more details):

- P-R Interval: The P-R interval extends from the beginning of the P wave to the beginning of the QRS complex. It represents the time taken for the electrical impulse to travel from the SA node through the AV node to the ventricles. Normal

values range from 120 to 200 milliseconds. Prolonged or shortened P-R intervals can indicate conduction abnormalities such as atrioventricular block [17].

- Q-T Interval: The Q-T interval measures the time from the beginning of the Q wave to the end of the T wave. It reflects the total duration of ventricular depolarization and repolarization. The normal Q-T interval varies with heart rate, but typically it should be less than 440 milliseconds in men and 460 milliseconds in women. Prolongation of the Q-T interval can be associated with an increased risk of arrhythmias and sudden cardiac death [18].
- ST Segment: The ST segment is the flat section of the ECG between the end of the QRS complex and the beginning of the T wave. It represents the period when the ventricles are depolarized and are in the plateau phase of the action potential. Deviations from the baseline in the ST segment can indicate myocardial ischemia or infarction [19].

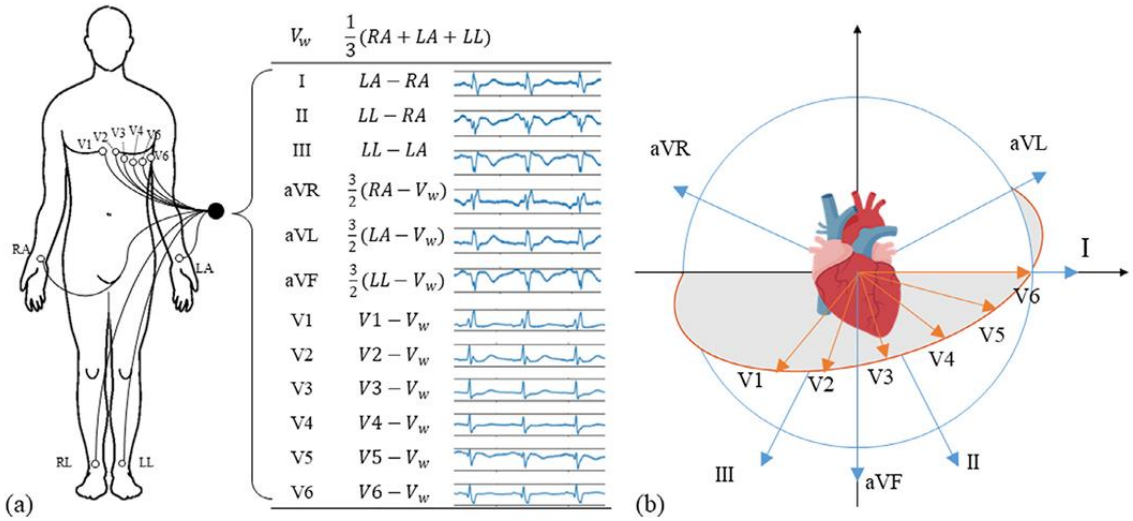


Figure 3 - 12-lead ECG system: a) Electrode placement for limb and precordial (chest) leads. b) ECG signal acquisition, capturing heart activity from multiple spatial angles. Adapted from [20].

Understanding the ECG's basic waveform is essential, but so is the context in which it is captured, particularly the ECG format itself. The 12-lead ECG is the most comprehensive and widely used system in clinical practice. It provides a multidimensional view of the heart's electrical activity by recording signals from 12 different leads or perspectives. This system includes three limb leads (Lead I, Lead II, and Lead III) and three augmented leads (aVR, avF, and aVL), which record electrical activity in the frontal plane, and six precordial leads (V1-V6) that capture electrical activity in the horizontal plane (see Figure 3). The 12-lead ECG's advantage lies in its ability to provide a detailed view of the heart's electrical function, aiding in the detection of complex

conditions like myocardial infarction, ischemia, and ventricular hypertrophy. However, it is more time-consuming and requires precise placement of electrodes, which may not be feasible in all situations [21]. In contrast, single-lead ECGs offer a simpler and more accessible alternative. These systems record the heart's electrical activity using only one lead, typically derived from two or three electrodes. Single-lead ECGs are particularly useful for long-term monitoring, especially in non-clinical environments where ease of use is crucial. Handheld devices can be used by patients themselves, making it easier to monitor conditions like arrhythmias without needing to visit a healthcare provider. While single-lead ECGs are less comprehensive than 12-lead ECGs and may miss certain conditions like myocardial infarction or left ventricular hypertrophy, their portability and ease of use make them invaluable for screening and ongoing monitoring [22]. Comparing these two systems, the 12-lead ECG is superior for a full diagnostic assessment, especially in acute settings where detailed information about the heart's electrical activity is fundamental. However, single-lead ECGs excel in scenarios where continuous monitoring is needed, or where access to full 12-lead equipment is impractical. The trade-offs between detail and convenience mean that both types of ECGs have their place in modern cardiology [23].

In addition to these considerations, understanding the timing of cardiac cycles is essential for accurate ECG interpretation. Systole begins with the onset of the QRS complex and ends with the T wave, representing the period of ventricular contraction and ejection of blood. Diastole follows, beginning after the T wave and ending with the next QRS complex, representing the period of ventricular relaxation and filling. These phases are critical for assessing the mechanical function of the heart in conjunction with its electrical activity [16].

ECGs are indispensable in clinical practice for their ability to provide real-time, non-invasive assessments of cardiac function. They are routinely used in various settings, from emergency departments to routine check-ups, and are pivotal in the management of patients with cardiovascular disease. The versatility of ECGs extends beyond initial diagnosis; they are also vital for monitoring ongoing treatment, detecting complications, and guiding therapeutic interventions [24].

2.1.2. Phonocardiogram signals/Heart sounds

Phonocardiograms (PCGs) are important diagnostic tools used to record and analyse the sounds produced by the heart during the cardiac cycle. These sounds, which

correspond to the mechanical events of the heart, provide valuable insights into cardiovascular health and can aid in diagnosing various heart conditions. They arise from the turbulence caused by the closure of heart valves and the movement of blood through the heart's chambers, with their integrity assessment based on the sounds produced during valve openings and closings [16].

The cardiac cycle consists of two main phases: systole and diastole. Systole is the phase when the heart contracts to pump blood into the arteries, while diastole is the relaxation phase during which the heart fills with blood. A standard PCG from a normal cardiac cycle captures two fundamental heart sounds, namely the first heart sound (S1) and the second heart sound (S2). Additionally, PCGs may detect other sounds such as the third heart sound (S3) and the fourth heart sound (S4), which are often associated with pathological conditions [25]. Figure 4 illustrates the typical locations of the four heart sounds.

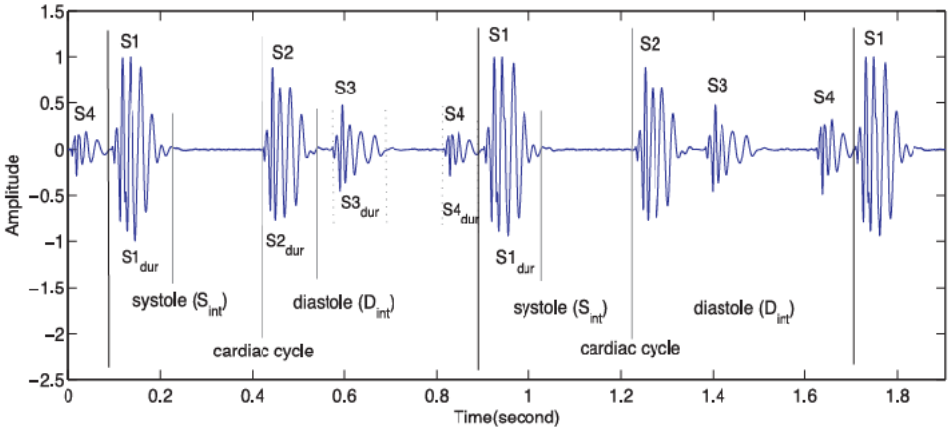


Figure 4 - PCG signal depicting heart sounds S1, S2, S3, and S4, along with the corresponding systolic and diastolic phases. Adapted from [26].

Here is a summary of the four main heart sounds, denoted as S1, S2, S3, and S4, each associated with specific events in the cardiac cycle:

- S1 (First Heart Sound): Known as the "lub" in the "lub-dub" sequence, S1 marks the beginning of systole and results from the closure of the atrioventricular (AV) valves (mitral and tricuspid valves) as the ventricles contract. The intensity and quality of S1 can vary based on valve position and ventricular contraction strength, with dominant frequencies between 20 and 175 Hz, and peaks ranging from 10 to 140 Hz. Abnormalities in S1 may indicate conditions such as mitral stenosis or AV node conduction delays [25].

- S2 (Second Heart Sound): Referred to as the "dub," S2 marks the end of systole and the start of diastole, produced by the closure of the semilunar valves (aortic and pulmonary valves) as the ventricles relax. S2 is typically shorter and higher pitched than S1 and can be split into A2 (aortic valve closure) and P2 (pulmonary valve closure) during inspiration. Its dominant frequencies range from 20 to 175 Hz, with peaks from 10 to 400 Hz. Abnormal splitting of S2 may indicate conditions like right bundle branch block or pulmonary hypertension [25].
- S3 (Third Heart Sound): Also known as the ventricular gallop, S3 occurs early in diastole during rapid ventricular filling. It is considered normal in children and young adults but may signal heart failure or volume overload in older adults. S3 is a low-frequency sound, typically between 20 and 70 Hz, and can indicate increased ventricular filling pressures or valvular regurgitation [27].
- S4 (Fourth Heart Sound): The S4 sound, or atrial gallop, occurs just before S1, during late diastole, when the atria contract to complete ventricular filling. S4 is usually pathological and associated with conditions such as left ventricular hypertrophy, aortic stenosis, or ischemic heart disease. It is also a low-frequency sound, observed between 20 and 70 Hz, and indicates decreased ventricular compliance and diastolic dysfunction [27].

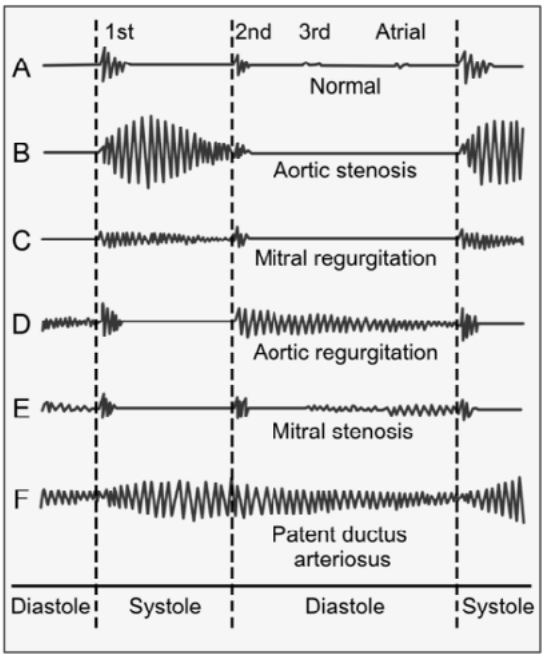


Figure 5 - PCGs signals displaying normal and various abnormal heart sound patterns. Adapted from [28].

In addition to these sounds, murmurs may be detected, characterized by blowing, whooshing, rasping, or swishing noises originating from or near the heart. Murmurs occur

when heart valves do not close properly or when blood flows through a narrowed valve. They exhibit a range of morphologies and are classified based on various factors, such as the phase of the cardiac cycle (systolic, diastolic, or continuous), their location, intensity, and the pitch, which can be low, high, or a combination of different frequencies. While some murmurs may be benign and not indicate heart disease, loud murmurs or irregular sound patterns can suggest underlying cardiac abnormalities [29]. Figure 5 provides a comparison between normal and abnormal heart sounds.

PCGs offer several advantages for cardiac assessment. They are non-invasive and can be used for continuous or intermittent monitoring without requiring complex equipment or procedures. PCGs are effective for detecting conditions such as murmurs, valve abnormalities, and cardiac anomalies, providing real-time information about heart function. They are also relatively cost-effective compared to other diagnostic tools like echocardiography and cardiac MRI, making them accessible for routine evaluations [2].

However, PCGs have limitations. They are sensitive to ambient noise and other sound interferences (like the internal organ's sound), which can affect recording clarity and accuracy. This sensitivity can make it challenging to differentiate between normal and abnormal sounds, particularly in noisy environments. Moreover, accurate interpretation of PCGs requires skill and training, as heart sounds can be complex, varied, subtle and occur closely in time, often at frequencies that are less perceptible to the human ear [29], [30]. Thus, while PCG is a valuable tool, its effectiveness can be compromised by environmental factors and the expertise required for accurate analysis. Despite these challenges, advancements in digital signal processing have enhanced PCG utility by enabling precise analysis and automatic detection of abnormal murmurs [31]. Modern algorithms can identify specific frequency patterns associated with various heart conditions, improving diagnostic accuracy. These advancements make PCGs valuable not only in clinical settings but also in remote and telemedicine applications where access to advanced imaging may be limited.

2.1.3. Relation between ECG and PCG signals

ECGs and PCGs are fundamental concurrent diagnostic tools that provide complementary information about cardiac function. While ECGs capture the heart's electrical activity, PCGs record the mechanical events associated with heart sounds, particularly those generated by the opening and closing of the heart valves, which in turn are triggered by electrical activation [32]. The integration of ECG and PCG signals

provides a more comprehensive understanding of the cardiac cycle, enhancing the diagnosis and monitoring of various cardiovascular conditions [33]. This approach has proven particularly valuable in diagnosing conditions like valvular heart disease, where both electrical and mechanical abnormalities may coexist [34].

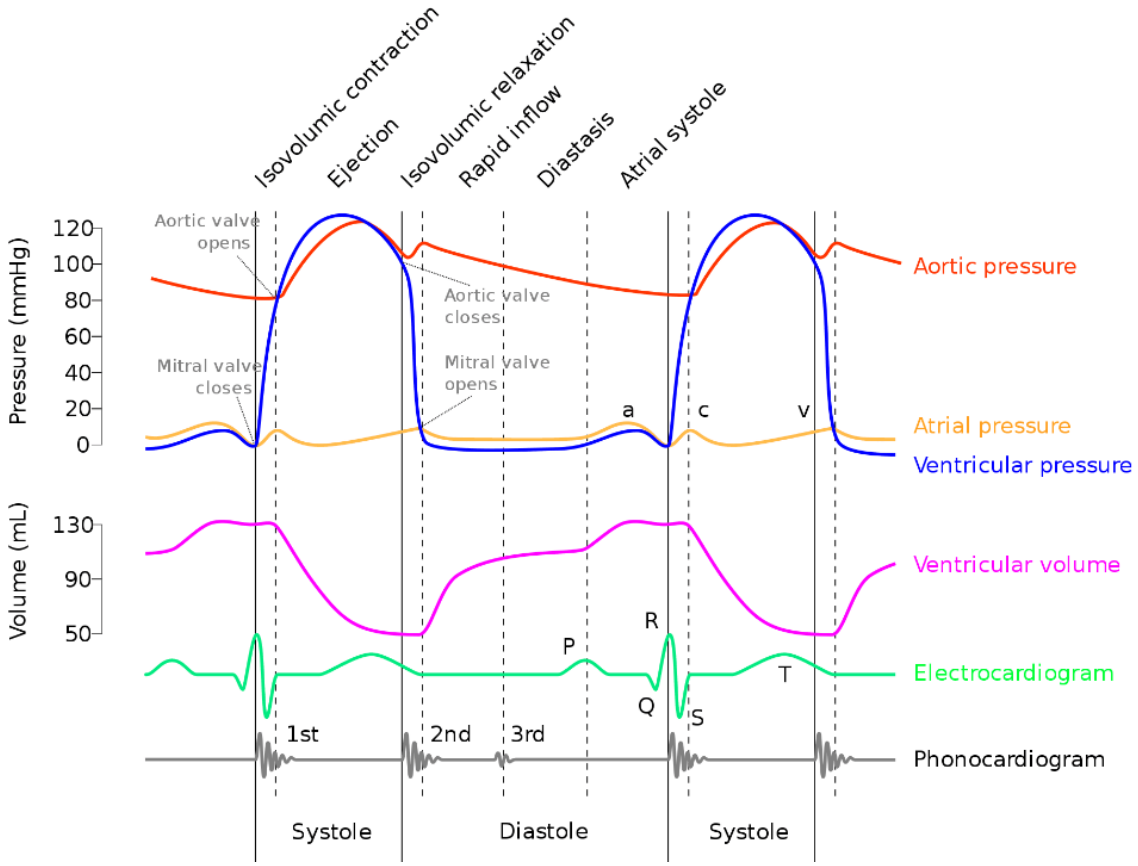


Figure 6 - Wiggers diagram. The green line represents the ECG, while the grey line depicts the corresponding heart sounds (PCG). Adapted from [35].

The heart operates as a coordinated system, where electrical and mechanical activities are closely linked. The electrical impulses recorded by the ECG, specifically the P wave, QRS complex, and T wave, correspond to mechanical events detected by the PCG, such as the first (S1) and second (S2) heart sounds. As shown in Figure 6, the QRS complex represents ventricular depolarization, which leads to ventricular contraction [29]. This electrical activity precedes the first heart sound (S1), recorded by the PCG, as there is a brief delay between electrical activation and the mechanical contraction of the heart [36]. The S1 sound is associated with the closure of the atrioventricular (mitral and tricuspid) valves at the onset of ventricular systole. The T wave indicates ventricular repolarization, leading to ventricular relaxation. The second heart sound (S2), occurring near the end of the T wave, is also detected by the PCG and

results from the closure of the semilunar (aortic and pulmonary) valves, marking the end of systole and the beginning of diastole [16], [29].

Understanding the synchronous relationship between these signals is crucial. The S1 sound is typically more or less aligned with the R peak in the QRS complex, while the S2 sound occurs shortly after the T wave. This temporal relationship reflects the electrical and mechanical coordination required for effective cardiac function. However, it is important to note that abnormalities in a PCG signal do not necessarily imply corresponding issues in the ECG. For example, defective heart valves may not be evident in an ECG but manifest as heart murmurs in the PCG [32]. Conversely, conduction disorders are easily detected in the ECG but are less apparent through heart sounds [37]. In summary, integrating the distinct features of heart sounds with the ECG provides a more comprehensive understanding of cardiac function, ultimately enhancing the accuracy of heart disease diagnosis.

2.2. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have emerged as a transformative architecture in deep learning, widely recognized for their effectiveness in tasks such as image recognition, natural language processing, and medical image analysis [38]. Their capability to autonomously learn hierarchical features from raw data has led to significant advancements across various domains, including signal processing [39]. This section provides an in-depth exploration of the essential components and concepts associated with CNNs, focusing on convolutional layers (1D and 2D), pooling layers, flattening layers, batch normalization layers, dense layers, and other critical aspects of their architecture and training process.

The convolutional layer is the foundational element of a CNN, responsible for extracting features from the input data. It applies a set of filters, each defined by a specific kernel size, to the input, resulting in a feature map that emphasizes patterns such as edges, textures, and shapes. The convolution operation involves sliding a window across the input tensor, performing element-wise multiplication with a learnable kernel, and summing the results to produce the feature map, as observed in Figure 7. Multiple filters in a convolutional layer generate corresponding feature maps, with the number of filters determining the quantity of these maps. As layers are added, the network captures progressively more complex and abstract features, with initial layers detecting basic characteristics and deeper layers recognizing intricate patterns [40]. Convolutional layers

are characterized by the number of filters, input channels, and filter sizes, and can be categorized based on the dimensionality of the input they process:

- 1D Convolutional Layers: Primarily used for processing sequential data, such as time series or textual data, where the input is a one-dimensional array. A 1D convolution applies a filter along the temporal dimension to capture patterns over time steps. The operation can be mathematically expressed as:

$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a) \tag{2.1}$$

  where $s(t)$ is the output at time step t, $x(a)$ is the input sequence, and $w(t-a)$ is the filter [40].

- 2D Convolutional Layers: Commonly used in image processing, where the input is a two-dimensional array representing the image. The filter slides over the height and width of the image to detect spatial features like edges and textures, creating a feature map. The 2D convolution is mathematically represented as:

$$S(i,j) = (I * K)(i,j) = \sum_{m}\sum_{n} I(m,n)K(i-m, j-n) \tag{2.2}$$

  where $S(i,j)$ is the output feature map at position (i,j), $I(m,n)$ is the input image, and $K(i-m, j-n)$ represents the convolutional filter [40].

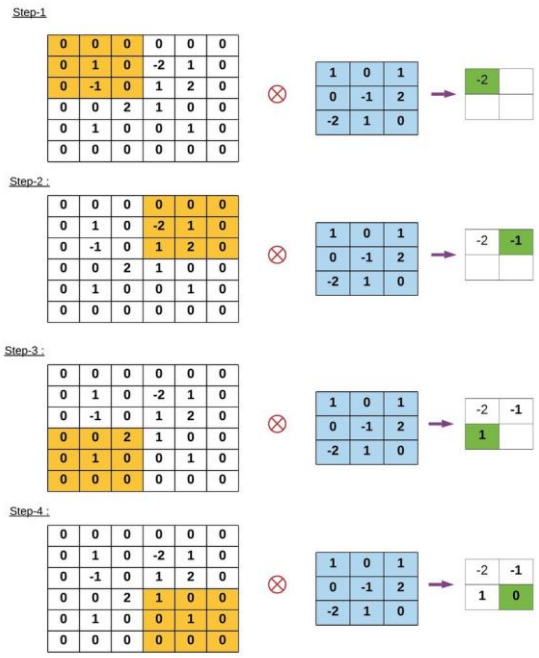


Figure 7 - Illustration of the computations at each convolution step: a 3 × 3 kernel (highlighted in blue) is applied to a matching 3 × 3 region (highlighted in yellow) of a 6 × 6 input image, which was generated by applying zero-padding to the original 4 × 4 input. The element-wise multiplication of the kernel and the region is summed to produce a corresponding value (highlighted in green) in the output feature map. Adapted from [41].

Convolutional layers are followed by activation functions that introduce non-linearity into the model, enabling it to learn complex representations. The Rectified Linear Unit (ReLU) is the most commonly used activation function after convolutional layers, defined as:

$$\mathrm{ReLU}(x) = \max(0, x) \tag{2.3}$$

This activation function accelerates convergence during training by mitigating the vanishing gradient problem, which can occur with other activation functions like sigmoid or tanh [41].

Pooling layers are used after convolutional layers to reduce the spatial dimensions of the feature maps, decreasing computational complexity and mitigating overfitting. Max pooling is one of the most common techniques, which retains the maximum value within a defined window (typically 2x2 for 2D data) and discards the rest. Mathematically, max pooling can be represented as:

$$y_{i,j} = \max_{(m,n)\in \mathrm{window}} x_{i\cdot s+m, j\cdot s+n} \tag{2.4}$$

where $y_{i,j}$ is the output value at position (i, j), x is the input feature map, s is the stride, which determines the step size of the pooling window, and the max operation is applied over a window size of $k \times k$. By reducing the resolution, pooling layers make the network more robust to variations in the input, such as small translations or distortions [40], [41].

After a series of convolutional and pooling layers, the high-level feature maps are typically flattened into a one-dimensional vector. This vector is then passed to fully connected layers (or dense layers), which resemble traditional neural networks. These layers integrate the extracted features to make final predictions. In binary classification tasks, the sigmoid activation function is often used, defined as:

$$\mathrm{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{2.5}$$

Dense layers play a critical role in integrating abstract features into meaningful predictions, which are subsequently evaluated using a loss function [41].

Batch normalization [42] is a technique employed to enhance the stability and performance of neural networks. It normalizes the input of each layer to have a mean of zero and a standard deviation of one, which accelerates training by reducing internal covariate shift and allowing the use of higher learning rates. Mathematically, batch normalization for a mini-batch is defined as:

$$\hat{x}^{(i)} = \frac{x^{(i)} - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{2.6}$$

$$y^{(i)} = \gamma \hat{x}^{(i)} + \beta \tag{2.7}$$

where $x^{(i)}$ is the input, μ and σ^2 are the mean and variance of the mini-batch, ϵ is a small constant to prevent division by zero, and γ and β are learnable parameters that scale and shift the normalized value.

The loss function measures the discrepancy between the predicted output and the actual label, guiding the optimization process. For binary classification, the binary cross-entropy loss is commonly used:

$$\text{Binary Cross-Entropy Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{2.8}$$

where y_i is the actual label, p_i is the predicted probability, and N is the number of samples.

To minimize the loss function and update the network's weights, CNNs are typically trained using the Adam optimizer, known for its efficiency and adaptive learning rate [43]. Adam combines the strengths of both AdaGrad [44], which is effective for sparse gradients but can cause learning to slow down significantly over time, and RMSProp [45], which addresses AdaGrad's diminishing learning rates but may struggle with handling noisy gradients. Adam balances these issues by maintaining per-parameter learning rates based on both the first and second moments of the gradients, leading to faster convergence and more stable updates. Therefore, the update rules for Adam are:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)\nabla L(\theta_t) \tag{2.9}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)\nabla L(\theta_t)^2 \tag{2.10}$$

$$\widehat{m_t} = \frac{m_t}{1 - \beta_1^t}, \quad \widehat{v_t} = \frac{v_t}{1 - \beta_2^t} \tag{2.11}$$

$$\theta_{t+1} = \theta_t - \frac{\eta \widehat{m_t}}{\sqrt{\widehat{v_t}} + \epsilon} \tag{2.12}$$

where m_t and v_t are estimates of the first and second moments, β_1 and β_2 are hyperparameters, η is the learning rate, and ϵ is a small constant to prevent division by zero.

The performance of a CNN is significantly influenced by hyperparameters like batch size, epochs, and learning rate. Batch size refers to the number of samples processed before updating the model's weights; larger batch sizes provide more stable gradient estimates but require more memory. The learning rate controls the step size in weight updates; a high learning rate can speed up convergence but may overshoot the loss function minimum, while a low learning rate ensures precise convergence but requires more training time. Epochs denote the number of complete passes through the training dataset, with multiple epochs refining the model's weights, although excessive epochs can lead to overfitting [40], [41].

Regularization techniques are crucial for preventing overfitting in CNNs. Dropout is a widely used technique where a fraction of neurons in a layer are randomly set to zero during training, compelling the network to learn more robust features. This reduces dependency on specific neurons, enhancing the network's generalization ability. The dropout rate, typically between 0.2 and 0.5, determines the fraction of neurons dropped. Another common regularization method is L2 regularization (ridge regression), which adds a penalty to the loss function proportional to the square of the weights, discouraging the model from learning excessively large weights and thus reducing overfitting. The modified cost function with L2 regularization is:

$$\text{Cost Function} = \text{loss} + \frac{\lambda}{2}||w||^2 \tag{2.13}$$

where loss is the original loss function, λ is the regularization parameter, and w represents the matrix norm of the network' weights [40], [41].

CNNs are a pivotal component of modern deep learning, offering powerful tools for feature extraction and classification. By understanding and optimizing the key components such as convolutional layers, batch normalization, and the Adam optimizer, researchers can develop models that achieve superior performance in a wide range of applications. Regularization techniques like dropout and L2 regularization further enhance model robustness, making CNNs indispensable for handling complex data.

3. Literature review

The main purpose of this chapter is to perform a search, analysis and discussion of the knowledge already produced related to the thesis' main objective. Furthermore this chapter is divided in 3 sections: the first one, Search and Review Methodology, explains the applied article search methodology, as well as the criteria used for the selection; the second one, Results, is a brief overview of the deep learning models used for ECG plus PCG (multimodal) classification, as well some examples of the models used in ECG only or PCG only, with the respective XAI techniques applied; the third and last section, Discussion, is the analysis and overall discussion of the results/information obtained with a table (Table VI) that summarizes the most important aspects of the chosen articles.

3.1. Search and Review Methodology

To ensure the reproducibility and relevance of this review, the online article databases chosen were PubMed, ScienceDirect, Scopus, and IEEE Xplore. The search was conducted using a PICO model, which stands for Population, Intervention, Comparison, and Outcome. This model was implemented in the queries applied across all databases. The first step was to define the appropriate keywords for each component of the PICO model:

- P (population): "heart sound", "PCG", "electrocardiogram", "ECG";
- I (intervention): "multimodal deep learning", "deep learning";
- C (comparison): not applicable in this situation;
- O (outcome): "explainable artificial intelligence", "XAI", "explainable AI".

Next, these keywords were combined to form different query combinations. After several attempts, which yielded varying numbers of articles, the most effective query was determined to be: ("heart sound" OR "PCG" OR "electrocardiogram" OR "ECG") AND ("deep learning" OR "multimodal deep learning") AND ("explainable artificial intelligence" OR "explainable AI" OR "XAI"). The resulting number of articles retrieved from each database is presented in Table I.

Table I - Article XAI query search results

Database	Number of articles
PubMed	22
Science Direct	375
Scopus	66
IEEE Xplore	25

Additionally, another query has built to include the multimodal models that take ECG and PCG as an input: (“heart sound” OR “PCG”) AND (“electrocardiogram” OR “ECG”) AND (“deep learning” OR “multimodal deep learning”). The resulting number of articles retrieved from each database for this query is presented in Table II.

Table II - Article multimodal query search results

Database	Number of articles
PubMed	6
Science Direct	306
Scopus	58
IEEE Xplore	41

The first search resulted in a total of 488 papers and the second search resulted in 411 papers. To identify the most relevant articles, a two-step selection methodology was employed:

- Superficial selection: Articles were initially filtered based on their title and abstract to determine their relevance to the thesis scope.
- Detailed selection: A thorough review of the full text was conducted to select articles with significant results or conclusions relevant to this work.

From this selecting process, a total of 20 articles were selected from all databases, 10 articles for each query. Tables III and IV showcase how many articles were considered from each database for the final selection:

Table III - Article XAI query final search results

Database	Number of articles
PubMed	0
Science Direct	1
Scopus	5
IEEE Xplore	4

Table IV - Article multimodal query final search results

Database	Number of articles
PubMed	1
Science Direct	2
Scopus	1
IEEE Xplore	6

Each article was carefully examined, and pertinent data were extracted and compiled according to the evaluated attributes and their purpose, as outlined in Table V.

Table V - Attributes evaluated and their purpose

Attributes	Purpose
Article	Reference of the article
Application/Objective	Brief description of the project's purpose (e.g., classification task)
Dataset	Identification of the dataset and its characteristics
Methodology/Algorithm applied	Summary of the key characteristics of the model architecture
XAI technique	Description of the XAI technique or algorithm used (if applicable)
Performance	Synopsis of the metrics calculated in the study

This systematic approach ensured a comprehensive review of existing literature, focusing on the integration of deep learning and explainable AI techniques for multimodal ECG and PCG classification. By meticulously selecting and analysing these articles, the review aims to provide a solid foundation for understanding the current state of research in this area and identifying gaps that future work can address.

3.2. Results

In order to organize this section, the results were divided into two subgroups, one for each query. We begin first with the multimodal deep learning models that use PCG and ECG as input, and then we discuss about the XAI techniques applied to ECG or PCG deep learning models (since there is only one work regarding the application of XAI using a multimodal model that have ECG and PCG as input).

3.2.2. Multimodal deep learning models

Few studies have investigated the potential performance enhancements gained by integrating PCG and ECG signals in a deep learning framework for simultaneous analysis, as earlier research predominantly employed traditional machine learning methods like Ensemble Classifiers, Support Vector Machines (SVM) and Random Forest, as seen in [46], [47], [48]. Given that these signals are synergistic and may each provide unique information about cardiac function, combining them could theoretically enhance the accuracy and efficacy of automated disease screening systems. Studies with only one classification metric (e.g., a single metric like accuracy in [49]) or with non-reliable results were excluded. In the following paragraphs, the reviewed studies that focus specifically on multimodal deep learning classification models will be summarized.

Han li et al. proposed a dual-input neural network that integrates feature extraction and deep learning techniques to enhance the diagnosis of coronary artery disease (CAD) [9]. The study employs a private dataset comprising simultaneously recorded ECG and PCG signals (5 minutes, with sample rate of 1kHz) collected from 195 subjects (135 with CAD and 60 non-CAD). The proposed architecture consists of a dual-input neural network that processes extracted features from both signals, raw ECG signals and decomposed PCG signals. The first input includes multi-domain features extracted from the ECG and PCG signals (selected by Information Gain Ratio), encompassing time, frequency, energy, entropy and time-frequency domains. The second input is a five-channel concatenation consisting of raw ECG signals and a four-scale wavelet decomposition of PCG signals. The network combines two fully connected layers, that takes the first input, with a CNN, composed of 10 convolutional identical blocks (each with 3 convolutional layers), and a bidirectional gated recurrent unit (Bi-GRU) with an attention mechanism that takes the second input. The output of each model part is then concatenated (see Figure 8). Performance results indicate that the dual-input approach with the selected features significantly outperforms single-modality

in both fully connected and deep learning approaches. The proposed method achieved a classification sensitivity and specificity of 0.985 and 0.892 respectively when analysing the combined ECG and PCG recordings cropped to 15 seconds. The study highlights the importance of integrating multiple signal types to improve diagnostic accuracy and suggests further exploration of feature relevance and data volume for enhanced classification performance in future research.

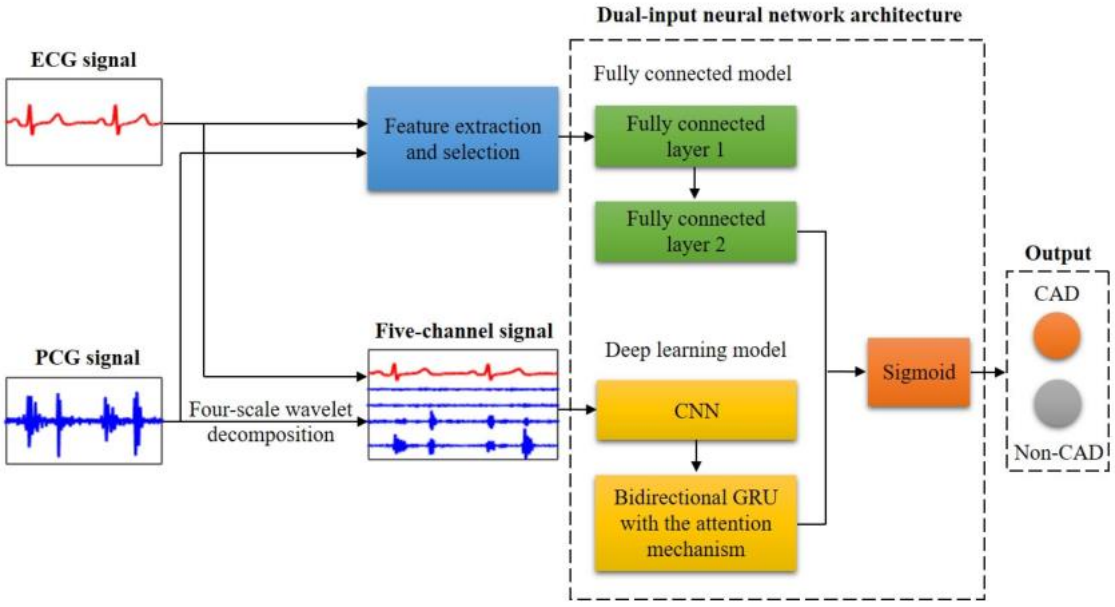


Figure 8 - Proposed Dual Input neural network by Han Li et al. Adapted from [9].

Ramith Hettiarachchi et al. developed a novel approach for screening pre-existing heart diseases by utilizing synchronized ECG signals and PCG waveforms [8]. The best model from the study employs the PHY16 dataset, which consists of recordings from the PhysioNet Challenge 2016, containing simultaneous PCG and ECG data (from training set A) from a limited number of patients [50]. The individual records were divided in 3.5 seconds, applying Continuous Wavelet Transform (CWT) to each segment to generate the scalogram. The authors propose a dual-CNN that consists of two distinct CNNs trained on individually acquired PCG and ECG waveforms, which then transfer the learned features to a single integrated CNN. Each single modality CNN branch is comprised of three sequential down sampling blocks, each consisting of a convolutional layer, max-pooling layer, and a ReLU activation function. Following these blocks, a convolutional layer and a flattening layer were implemented. The outputs from the flattening layers of both network architectures were then concatenated, after which the combined features were processed through a shared multilayer perceptron (see Figure 9). This approach enabled the model to effectively capture complex abnormalities in

heart conditions through automatic feature extraction, rather than relying on handcrafted features, which may not generalize well across diverse datasets. Regarding performance results, the proposed method achieved a high sensitivity of 0.947 and a specificity of 0.750, using dilated convolution and dropout techniques. These results indicate that the dual-modality approach (using both PCG and ECG) outperforms traditional single-modality methods, as it can better capture the nuances of cardiac abnormalities. Also, when using transfer learning weights from single modality models (one CNN branch), trained with signals from other datasets, and apply them on the hybrid model, it can increase specificity and Area Under the Curve (AUC) scores at a marginal cost of sensitivity and accuracy. The study highlights that the CNN's ability to learn features from both modalities leads to improved classification performance compared to methods that rely solely on one type of data.

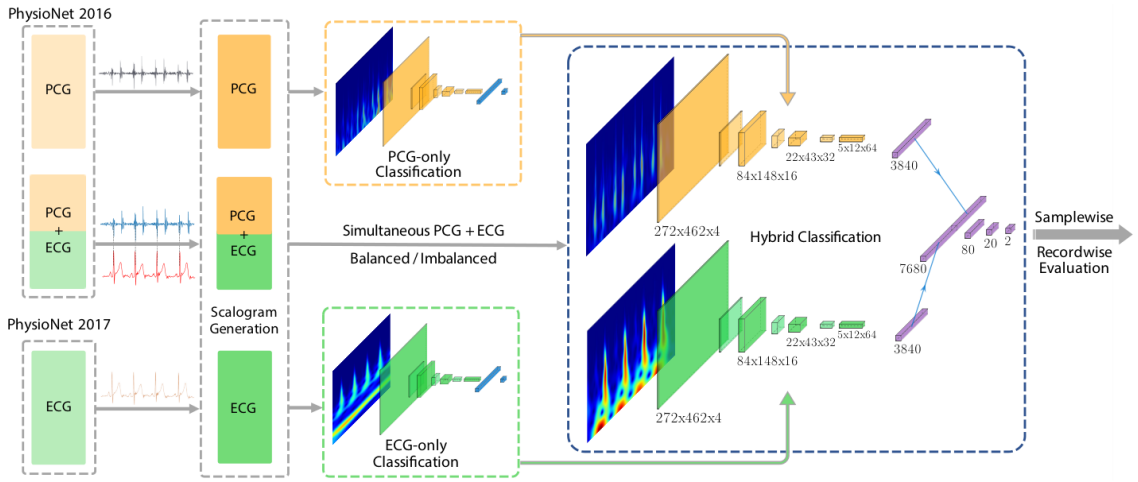


Figure 9 - Proposed network architecture by Ramith Hettiarachchi et al. Adapted from [8].

Pengpai Li et al., using the PHY16 dataset, developed an efficient multi-modal machine learning method for predicting cardiovascular diseases (CVDs) by integrating ECG and PCG features [51]. First, since the dataset is unbalanced, they segmented the positive labelled raw signals into a length of 8 seconds, while the window size of 3 seconds is applied to negative labelled raw signals. After that, they employed two convolutional neural networks: CL-ECG-Net for ECG signal encoding and CL-PCG-Net for PCG signal encoding. The CL-ECG-Net is composed of 8 convolutional layers and 8 max pooling layers, with each of them followed by a batch normalization layer and a ReLU layer. After the fully convolutional neural network, two Long Short-Term Memory (LSTM) layers are added. For the CL-PCG-Net, each PCG signal is decomposed into four frequency bands: 25–45Hz, 45–80Hz, 80–200Hz and 200–400Hz. Each band is fed

into to four convolutional channels (composed by convolutional layers and max pooling layers), followed by a feature fusion layer of the 4 channels and two LSTM layers. Thus, feature extraction is performed using deep-coding techniques. Finally, a genetic algorithm (GA) is employed for feature selection (generating feature subsets), succeeded by a SVM classifier to train each subset and give the prediction (see Figure 10). The best classification performance achieved with the combined features resulted in a sensitivity of 0.903, specificity of 0.845, F-score of 0.874, accuracy of 0.873, and an AUC of 0.936, with the multi modularity results being better than the single modularity results.

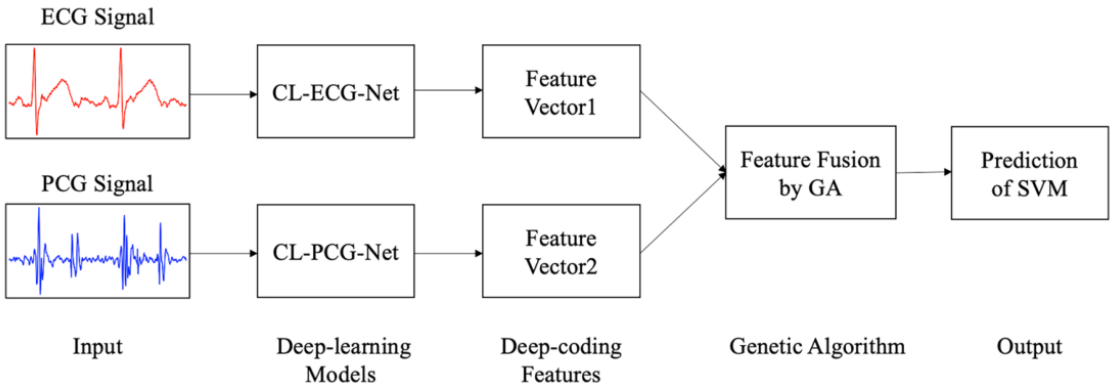


Figure 10 - Proposed framework by Pengpai Li et al. Adapted from [51].

Later, Han li et al., in 2021, using the same private dataset with the same objective as the previous study (detection of CAD), proposed a new feature extraction by integrating multi-domain (time, frequency, and time-frequency domains) deep features from ECG and PCG signals instead of hand-crafted and time-domain deep features, removing the requirement for feature engineering based on expert knowledge [52]. They proposed a new multi-input CNN architecture that integrates one 1D CNN model and two 2D CNN models. Raw pre-processed ECG and PCG signals are passed through 1D convolutional layers to initially capture time-domain features. Then, feature maps from the two convolutional layers are combined and used as input for the 1D CNN model, which is composed of ten identical convolutional blocks, followed by a Bi-GRU with an attention mechanism. For the analysis of ECG and PCG spectrum images generated via the GAF technique (explained in [53]), two 2D convolutional layers are employed to automatically extract features from the frequency domain. Afterwards, the outputs of these layers are then combined and input into a 2D CNN, consisting of three convolutional blocks followed by two fully connected layers. In what concerns the time-frequency domain features, they generated images by applying S transform on ECG and

Mel-Frequency Cepstral Coefficients (MFCCs) on PCG signals, implementing the same architecture as that used on ECG and PCG spectrum images. Lastly, the outputs of all three 1D/2D CNN models are concatenated, employing a sigmoid layer as the output layer (see Figure 11). This framework achieved an accuracy of 0.965, a sensitivity of 0.994, and a specificity of 0.901. In contrast, the performance metrics for single modalities were lower, highlighting the effectiveness of integrating multi-domain features for CAD detection.

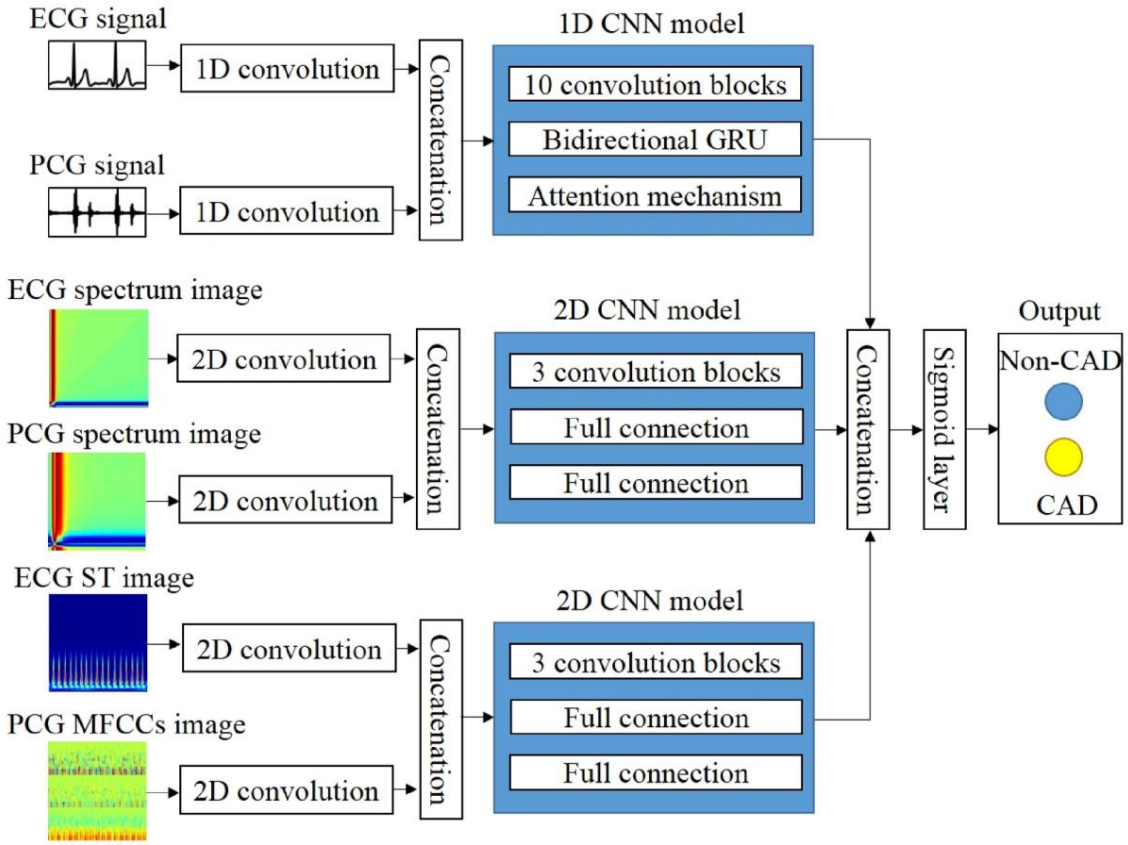


Figure 11 - Proposed newest neural network architecture by Han Li et al. Adapted from [52].

Monjur Morsh ed et al., to develop an automatic classification approach for heart valve defects (HVDs), employed synchronously recorded ECG and PCG signals from the PHY16 dataset [54]. They developed a CNN architecture designed to extract local features from the pre-processed and data augmented 1D ECG and PCG signals. The network consists of four 1D convolutional blocks in each branch (ECG and PCG), to extract initial time-domain features, all of them succeeded by batch normalization and dropout layers to prevent overfitting. At the end of each branch, a flatten layer is added followed by a dense, a batch normalization and a dropout layers (see Figure 12a). Subsequently, the output features from the two 1D CNNs are flattened, and the fully

connected layers are employed, using a SoftMax layer as output (see Figure 12b). In terms of performance results, the proposed model achieves an overall F1-score of 0.950 when using combined features from both ECG and PCG signals, which is approximately 5% and 2.5% higher than the accuracies obtained using only ECG or PCG signals, respectively. Grad-Cam++ is also applied to 2nd, 3rd and 4th convolutional layers to observe feature discrimination.

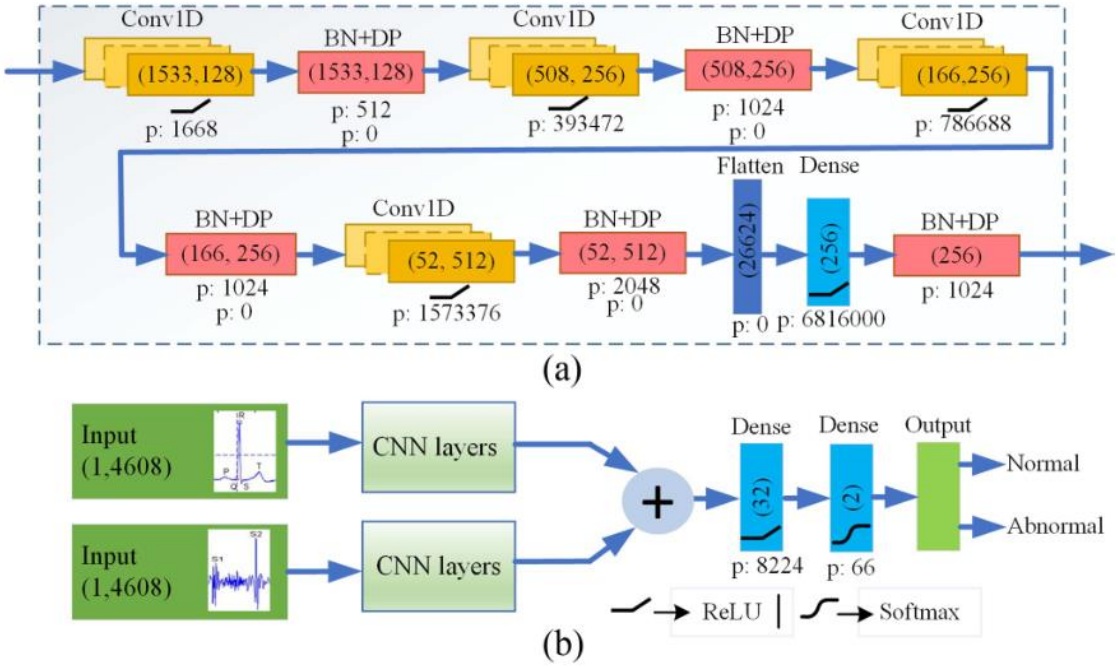


Figure 12 - Proposed neural network architecture by Monjur Morshed et al. Adapted from [54].

Haozhan Han et al. developed a multimodal deep neural network, referred to as MCT, for the classification of ECG and PCG signals, combining CNN and Transformer models to leverage both local and global feature extraction capabilities [55]. The datasets used include the EPHNOGRAM dataset [56] and a real-world private synchronized ECG and PCG dataset for disease classification. The EPHNOGRAM dataset includes 69 synchronized ECG and PCG recordings collected from 24 healthy adults across different exercise conditions, such as resting, walking, running, and cycling, with each recording duration varying between 30 seconds or 30 minutes, mainly applied to compare the heart-rate time-series and variabilities of the subjects during a stress-test. The real-world dataset, collected by the authors, is composed of 123 patients with various signal lengths, mainly focused on cardiovascular disease diagnosis. The model employs a novel attention mechanism for multimodal feature fusion, dynamically adjusting weights on intra- and inter-modal features to emphasize important information on 5 seconds

segmented raw signals (see Figure 13). First, the features are extracted by ResNet1d wang [57] and segregated into multiple patches. Then global average pooling (GAP) is performed, and each branch is fed into MCT blocks, consisting of CNN blocks with various CNN kernels, layer normalization and a Gaussian Error Linear Unit (GeLU) activation function. Afterwards, there is a fusion block were both modalities are concatenated and the relevance scores calculated. Finally, following three MCT blocks, the resulting tokens from the fusion blocks are concatenated to form the final feature vector, which is then passed through a classifier to generate the final prediction. Performance results indicate that the MCT model significantly outperforms previous models in both single and multimodal settings. On the EPHNOGRAM dataset, results in a F1 score of 0.987 and a recall of 0.988, showing improvements of over 0.05% and nearly 0.1%, respectively, over past methods. In the real-world dataset, the F1 score improves by nearly 3%, and both recall and accuracy experience approximately 6% enhancement over other methods.

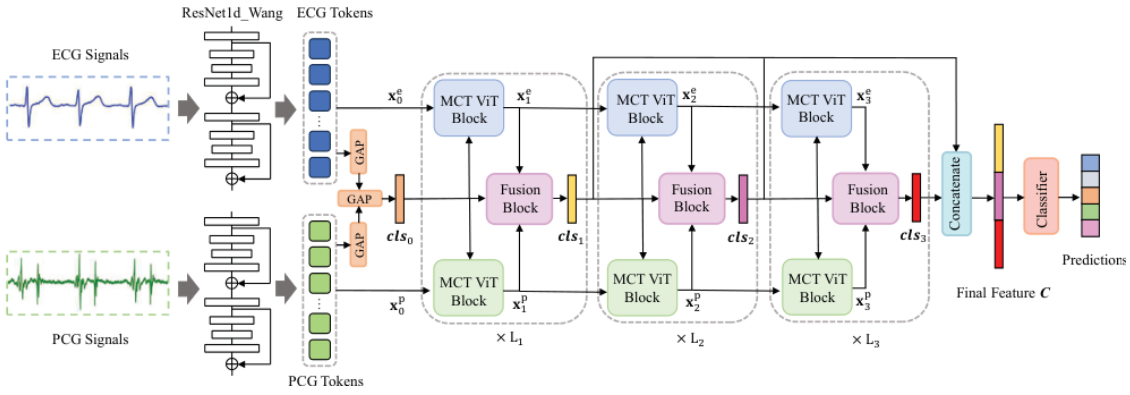


Figure 13 - Proposed MCT model by Haozhan Han et al. Adapted from [55].

Yuhan Chang et al. developed a multimodal model for detecting cardiac abnormalities by leveraging both ECG and PCG signals in a private collected dataset [58]. The dataset comprises 41 patient samples and 80 healthy control samples, divided in three classes (normal signal, common heart disease[1] and special heart disease[1]) to detect cardiac abnormalities. The ECG data has 8 leads sampled at 1000 Hz, while the PCG data includes 5 auscultation spots sampled at 8000 Hz. The proposed model integrates a CNN and a Transformer architecture. The CNN module is responsible for extracting features from the multi-lead ECG and PCG signals, whereas the Transformer module is utilized to fuse the multimodal information, enhancing the model's ability to

[1] In the original study, the authors did not provide any further details about the specific heart diseases classified as “common” or “special”.

capture long-range dependencies and modality-specific features through an attention mechanism channel. In detail, each modality goes through a feature extraction in independent CNN modules, composed of a Squeeze-and-Excitation (SE) module (with a GAP layer) that calculates a weight value for each feature map (using it in a scaling method), giving origin to attention weights for each channel. Then the result is fed into a series of convolutional and batch normalization layers. After that, a patch embedding procedure takes place using the feature map in a new convolutional layer, followed by a learnable positional embedding and a learnable modal-type embedding. Altogether is then inputted into a series of Transformer blocks called Vision Transformers (ViT). Each ViT block is comprised of Multi Head Self-Attention (MHSA) and Multi-Layer Perceptron (MLP) block (with 2 fully connected layers), using a GeLU activation function. Finally, they concatenate the features along the first dimension and establish a learnable global token, passing it to a classifier (see Figure 14). Performance results indicate that the proposed model significantly outperforms baseline models (like ResNet1dwang and Bi-LSTM), achieving more than a 2% enhancement in accuracy and recall, and over a 7% enhancement in the F1 score on the real-world data. Likewise, the multi modularity metrics surpasses the single modularity metrics, especially in F1-score.

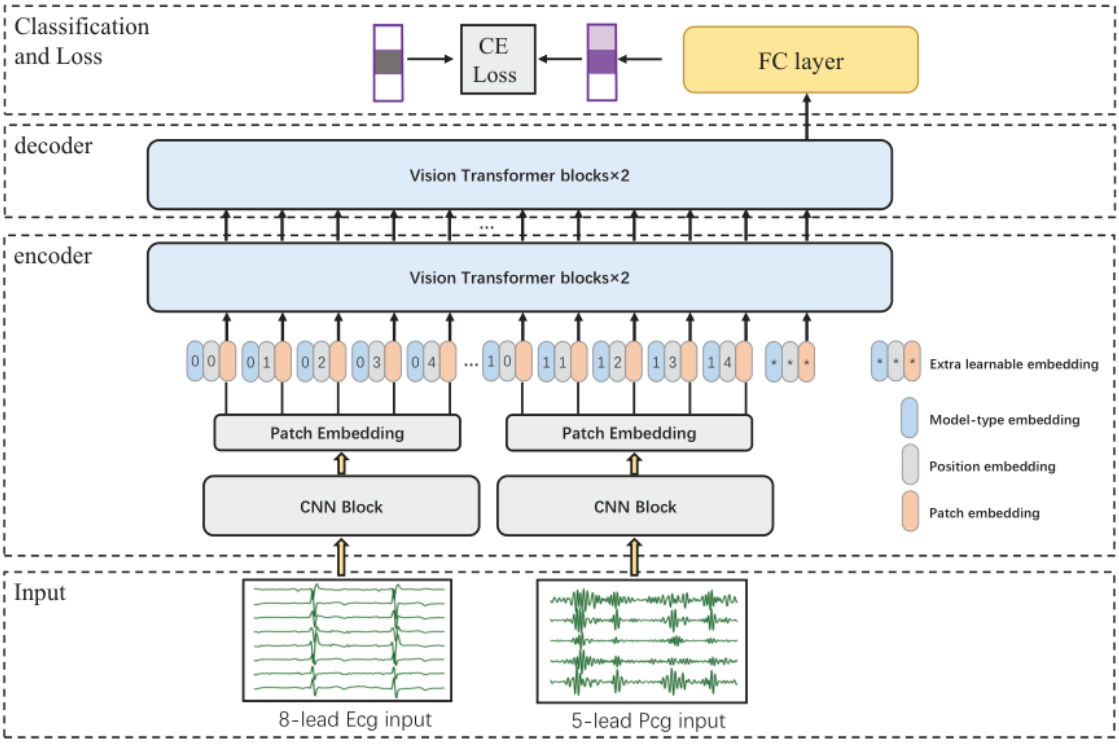


Figure 14 - Proposed model framework by Yuhan Chang et al. Adapted from [58].

Haobo Zhang et al. presents a novel approach for CVD detection using a co-learning-assisted progressive dense fusion network (CPDNet) that integrates synchronous ECG and PCG signals, addressing challenges such as training imbalance and missing-modality scenarios in clinical practice [59]. The study utilized two datasets, a public dataset from the PHY16 and a private dataset, composed of 212 pairs of recordings (ECG + PCG) from 53 individuals diagnosed with valvular diseases (positive cases) and 380 pairs of recordings (ECG + PCG) from 97 individuals without valvular diseases (negative cases), all with a record duration of 30 seconds. In valvular diseases is included aortic regurgitation, mitral stenosis, mitral regurgitation, and in non-valvular diseases is incorporated the normal types and other heart diseases such as cardiomyopathy. The CPDNet architecture consists of three branches: modality-specific encoders for ECG and PCG signals (to obtain multi-level features), and a progressive dense fusion encoder that aggregates features at multiple levels, employing cross-modality region-aware (CMR) and multi-scale feature optimization (MFO) modules to enhance multi-level feature extraction (from low to high level) and fuse them. In the end there is a decision-level fusion carried out on the output of the three branches. The modality-specific encoders consist of a spatial–temporal feature extraction module (with LSTM and CNN branches), a residual optimization module (with convolutional, max pooling, batch normalization and dropout layers) and a classification block (with two LSTM layers and a SoftMax output layer). Regarding the progressive dense fusion encoder, it gradually merges the low-level to high-level features (level by level), by a dense feature aggregation (DFA) method and refined step by step by applying CMR and MFO modules in alternating and repeated cycles. The CMR module is designed to dynamically assess the contributions of various modalities and signal regions, selectively emphasizing the most discriminative features, generating a separate weight map for each modality. On the other hand, the MFO module is created to improve the network's ability to represent multi-scale target regions within multi-modal signals, using progressive grouped convolutions with fewer parameters to achieve different receptive fields in a refined manner. To address training imbalance, the authors introduced a co-learning strategy to support the progressive dense fusion approach through intra-modality and joint loss functions. The intra-modality loss ensures that modality-specific encoders extract discriminative features effectively, while the joint loss promotes collaboration among these encoders and aids in fully integrating complementary information. Moreover, decision-level fusion is combined with feature-level fusion to leverage the data from the encoders, averaging output probability of the three encoders to obtain the final prediction (see Figure 15). The CPDNet surpassed state-of-the-art

multi-modality methods, achieving higher AUC scores across both datasets. It also excelled in scenarios with missing modalities (either ECG or PCG), demonstrating superior performance over single-modality models. This superiority was evident through significant improvements in sensitivity, specificity, and AUC metrics across both public and private datasets.

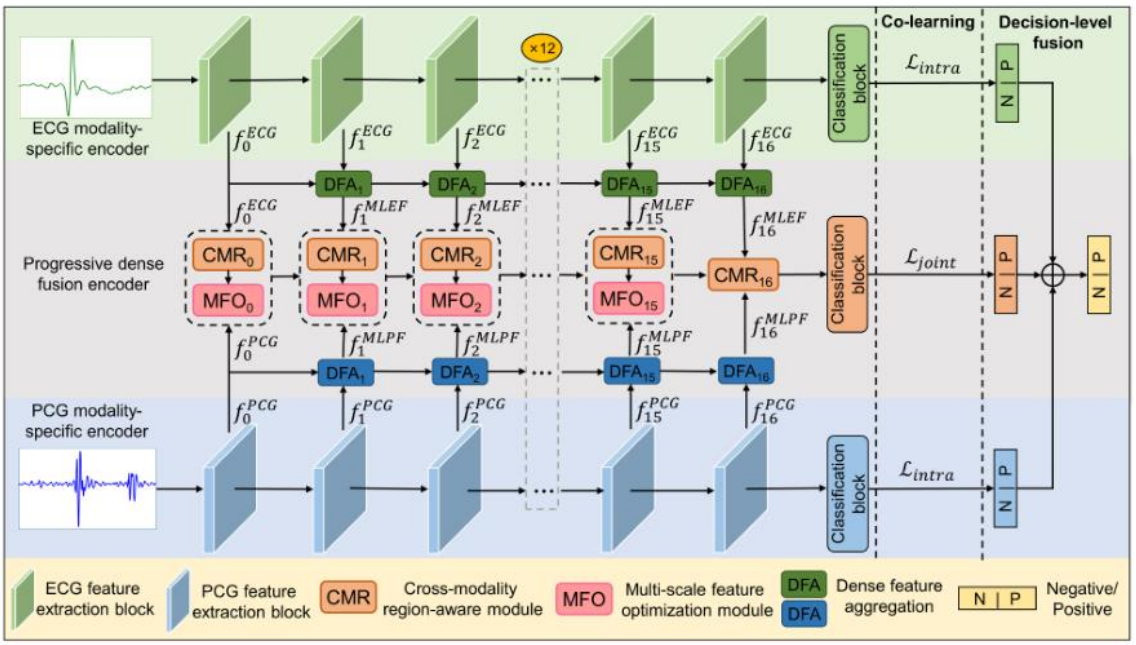


Figure 15 - Proposed CPDNet architecture by Haobo Zhang et al. Adapted from [59].

Jiayuan Zhu et al., using the PHY16 dataset, introduced a multi-branch residual network architecture (MR-Net) to detect cardiovascular diseases, employing a multi-modal approach that combines ECG and PCG signals [60]. The proposed model comprises 4 residual branches (with a convolutional, three residual blocks and max pooling layers) together with a SE attention module (SE-ResNet, to extract deep features for each modality). Afterwards, the resulting features are concatenated and fed to a SE-LSTM-Net (with a SE-Block, pooling and Bi-LSTM layers), followed by a fully connected layer and a SoftMax activation layer (all this trained separately). The resulting features are then fused and selected using Principal Component Analysis (PCA) before being classified using a SVM classifier (see Figure 16). The results indicate that the multi-modal method achieved a F1-score of 0.931 and an AUC value of 0.967, significantly outperforming both single-modal methods and prior multi-modal studies.

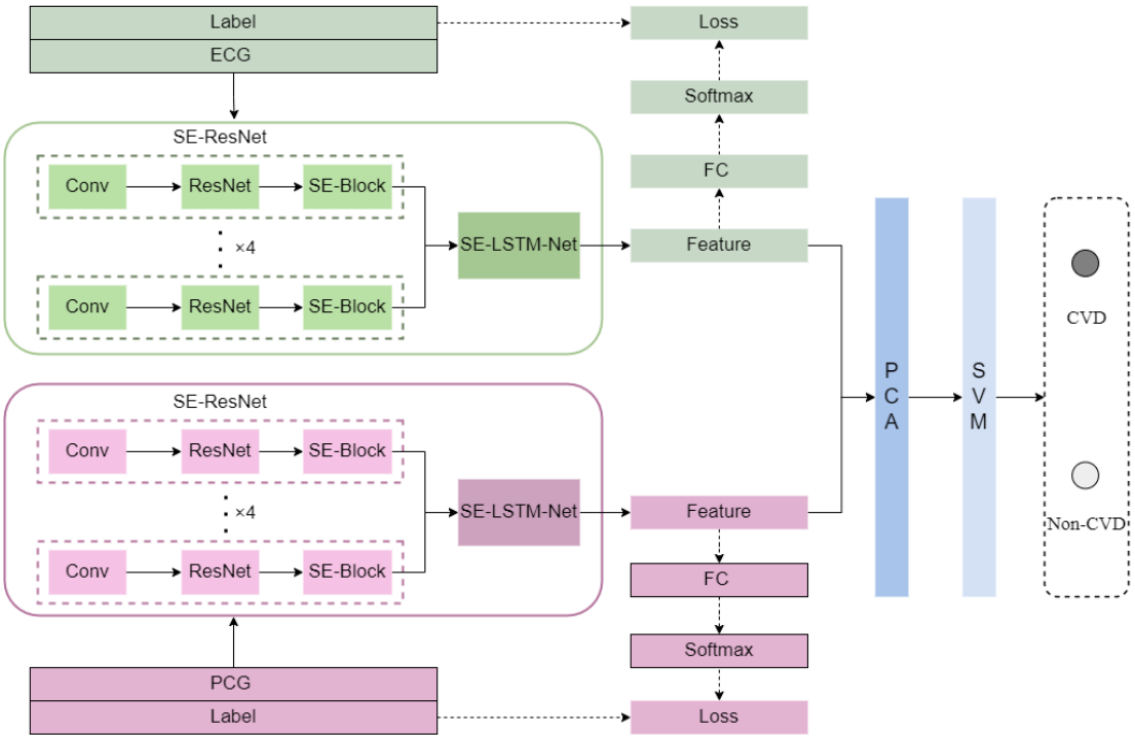


Figure 16 - Proposed MR-Net framework by Jiayuan Zhu et al. Adapted from [60].

Lastly, Mustafa Fuad Rifet Ibrahim et al. developed an efficient end-to-end deep learning model for the detection of cardiovascular abnormalities, to be applied in a battery-operated healthcare sensor patch, with the objective to reduce as much as possible the computational cost and memory requirements, while providing a state-of-the-art performance [61]. The study utilizes the PHY16 dataset and uses interpolation of missing values, down sampling, normalization and 3 seconds windows with 2 seconds overlap as preprocessing tasks. The proposed model encompasses four parts: an affine transform layer to fuse and assign different weights for each signal portion from both modalities; a 2D convolutional layer with a stride 2 to reduce dimensionality and computational costs; four bottleneck blocks from MobileNetV3-small (using 1D convolutions instead of 2D convolutions), to maintain low dimensionality and high expressiveness; a final fully connected layer with a SoftMax activation function to compute both normal and abnormal probabilities. Lastly, to combine the segment wise scores into the all-signal classification, the authors implemented a majority voting over all segments for the same record (see Figure 17). In case of a classification of the record in abnormal, it is then communicated to the physician for further evaluation. The model exhibits competitive performance by reducing the parameter count and floating-point operations (FLOPs) by more than two orders of magnitude compared to leading models. Despite these reductions, it maintains comparable performance to current state-of-the-

art models, achieving high AUC and F1 scores. This makes it well-suited for deployment on resource-constrained edge devices, such as sensor patches.

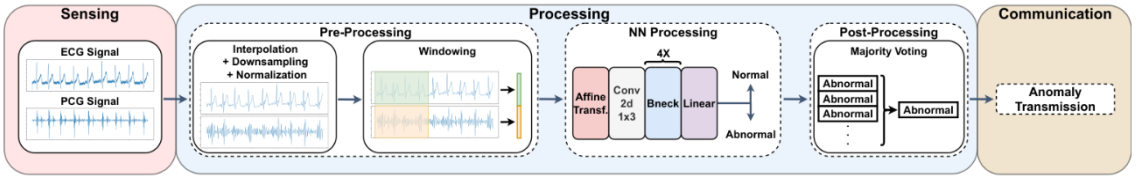


Figure 17 - Proposed model architecture by Mustafa Fuad Rifet Ibrahim et al. Adapted from [61].

3.2.3. Explainable Artificial Intelligence

There are a substantial number of articles employing commonly used XAI techniques; however, aside from the study by Monjur Morshed et al. mentioned earlier, most research appears to concentrate solely on single modalities (either ECG or PCG). Furthermore, studies that apply techniques out of the scope of this work, like federated learning or proof of concepts, and studies with non-reliable results (e.g. with only one non representative classification metric like accuracy), were excluded (except for one PCG article). Likewise, studies that only apply XAI in private datasets without external validation on a public dataset, were also excluded. These extra steps were considered due to a high number of articles, in order to select only 5 XAI articles for each modality. It is also important to refer that there are a substantial number of XAI articles in ECG signals, contrasting with the low XAI articles count for PCG signals. The following studies reviewed in the next paragraphs are exclusively focused on XAI models that depicts ECG (first five paragraphs) or PCG (last 5 paragraphs) abnormality classification.

Ganeshkumar M. et al. developed a robust method for multilabel classification of ECG signals, specifically focusing on identifying various heart diseases, including ST-segment elevation and depression, which are critical indicators of myocardial infarction [62]. The study used the CPSC2018 dataset, which comprises 6877 ECG recordings from 11 hospitals in China, focused on cardiac arrhythmias, covering conditions like normal sinus rhythm (SNR), atrial fibrillation (AF), first-degree atrioventricular block (IAVB), left bundle branch block (LBBB), right bundle branch block (RBBB), premature atrial contraction (PAC), premature ventricular contraction (PVC), ST-segment depression (STD), and ST-segment elevation (STE), with recordings ranging from 6 to 60 seconds [63]. The proposed methodology employs a modified VGG-16 architecture (with one convolution plus max pooling block removed to reduce the number of trainable

parameters), where feature extraction is enhanced through preprocessing steps like baseline wander and power line interference removal, followed by beat segmentation and period normalization. The resulting performance metrics encompass a subset accuracy of 0.962, precision of 0.986, recall of 0.949 and a F1-score of 0.967, demonstrating the method's efficacy in accurately diagnosing heart diseases. Moreover, the model's performance is validated using XAI techniques, specifically Grad-CAM, which confirms that the CNN effectively learns relevant features for classification. Figure 18 demonstrates an example of an ECG containing LBBB and its respective average activations from the study. A concomitant high activation in the broad QRS complex (with a clumsy R-wave) and in the inverted T wave, characteristic of the LBBB, demonstrates that the model is indeed learning the abnormal heart conditions.

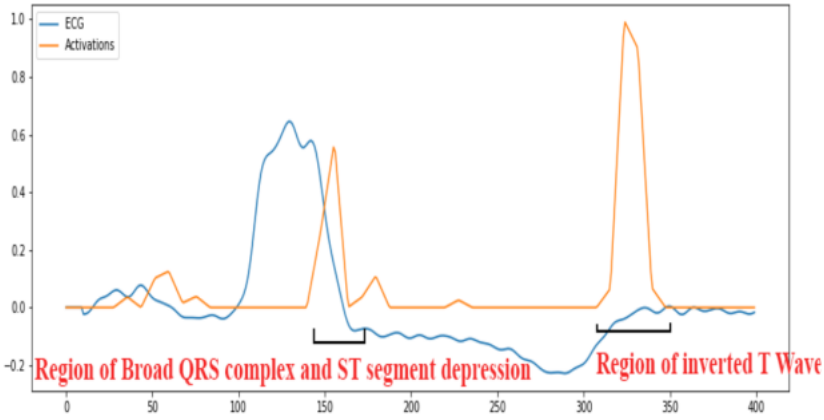


Figure 18 - LBBB ECG sample together with the average activations within the beat range. Adapted from [62].

Atul Anand et al. created an explainable AI decision model for the analysis of ECG data to detect cardiac disorders [64]. The primary objective is to create an efficient and interpretable AI model that clinicians can trust for diagnosing heart diseases. The researchers used two publicly available datasets: PTB-XL (21837 10 seconds length samples from 18885 patients, including healthy patients and various heart conditions like myocardial infarction, arrhythmia and ventricular hypertrophy) [65], and the CPSC2018 dataset. The proposed model, ST-CNN-GAP-5, employs a Spatio-temporal CNN with global average pooling instead of max pooling in the last convolutional layer (to reduce the number of trainable parameters), which was followed by a SHAP technique for interpretability enhancement, selecting the top 500 SHAP values to provide visual explanations on each signal. The tested model achieved an AUC of 0.934 on the PTB-XL dataset and 0.995 on the CPSC2018 dataset, demonstrating competitive performance. The model has also demonstrated ability to highlight relevant ECG wave

alterations for clinical diagnostics, like the identification of the slurred S wave with a total duration greater than the R wave in Lead I, in a patient that has RBBB (see Figure 19).

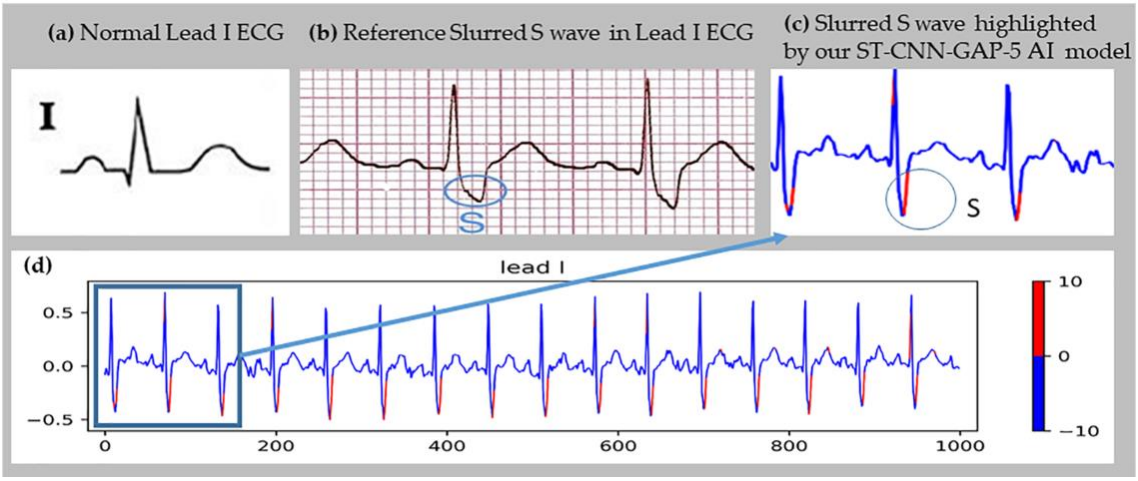


Figure 19 - RBBB ECG sample with overlaid SHAP values, where red values indicate a positive contribution on the predicted class and blue values indicate a negative contribution. Adapted from [64].

Khiem H. Lea et al. built a deep learning system for accurately identifying multiple cardiovascular abnormalities using only three ECG leads (I, II, and V1), designed to make ECG more accessible through portable or wearable devices [66]. Two large-scale ECG datasets are employed: the Chapman dataset that includes 10646 12-lead ECG samples with 10 seconds duration and a sampling rate of 500 Hz, covering 11 common rhythms and 67 additional cardiovascular conditions like atrial fibrillation and supraventricular tachycardia [67]; and the CPSC2018 dataset. The network's architecture is composed of three redesigned One-dimensional Squeeze-and-Excitation Residual Networks with 18 main layers (1D-SEResNet18) for feature extraction from each lead, followed by a Lead-wise Attention module for robust feature aggregation. Performance results show F1 scores of 0.972 on Chapman and 0.800 on CPSC2018, which is in line with current state-of-the-art methods, while maintaining computational and storage efficiency. The system also incorporates a modified Grad-CAM technique (Lead wise Grad-CAM) for explainability, using importance scores collected from the Lead-wise Attention module that displays the contribution of each backbone's features (i.e. each signal lead) to the final prediction, by multiplying each importance with each generated CAM and normalizing it. One such example is in Figure 20, where we can observe the importance scores of each lead (with Lead I and V1 showcasing higher scores), together with the activation maps of the signal parts that contributes the most for the prediction of LBBB (larger QRS waves in Lead I and deep S waves in Lead V1).

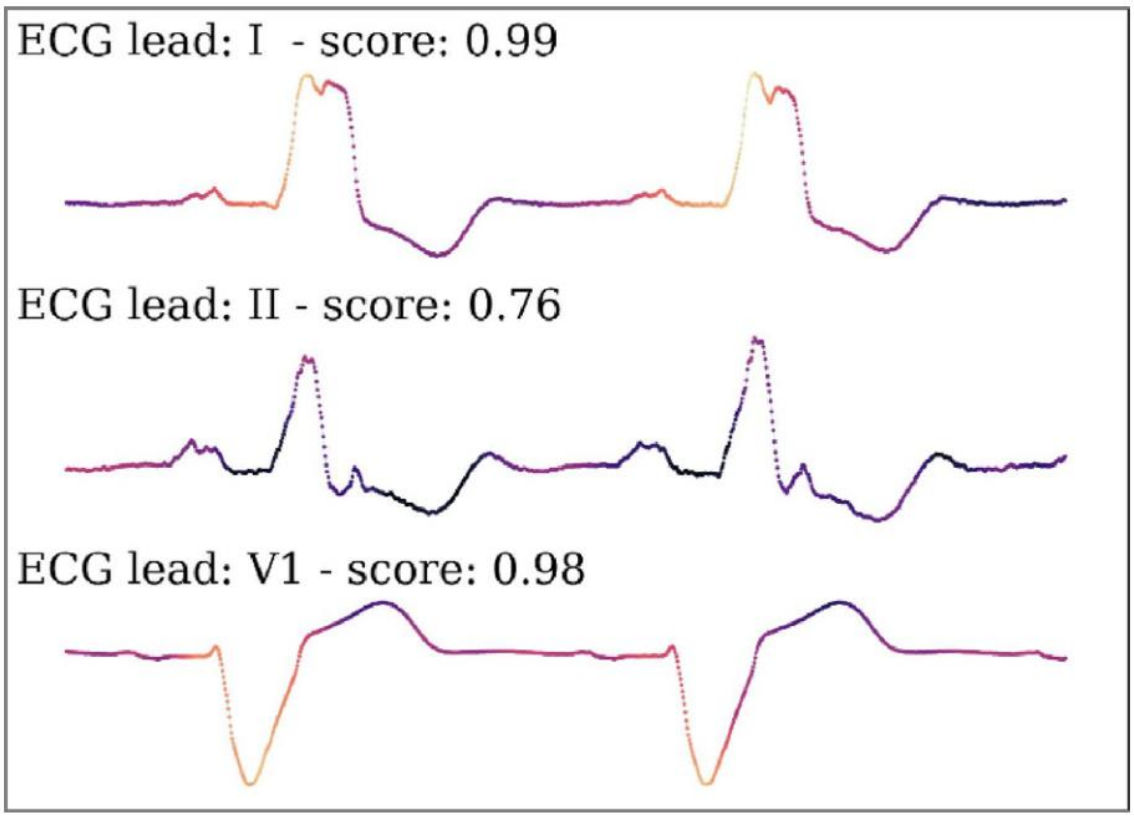


Figure 20 - LBBB ECG samples from lead I, II and V1, with overlaid activation map and lead importance scores. Adapted from [66].

Negin Alamatsaz et al. developed a lightweight deep learning approach to detect eight different cardiac arrhythmias and normal sinus rhythms using ECG signals [68]. In this work the authors employed ECG signals from two prominent databases: the MIT-BIH arrhythmia database, containing 48 half-hour 2 channels ECG recordings from 47 subjects at a sampling frequency of 360 Hz, with conditions such as supraventricular arrhythmias and conduction defects [69]; and the long-term AF database (LTAF), that comprises 84 two lead ECG recordings of subjects with paroxysmal or sustained atrial fibrillation, with signal duration between 24 and 25 hours and a sampling rate of 128 Hz [70]. The study introduced an 11-layer network that combines CNN (three convolutional blocks composed by a 1D convolutional, ReLU, max pooling and dropout layers) and LSTM architectures for effective feature extraction, followed by fully connected layers and a SoftMax output layer. In terms of performance results, the proposed model achieved an accuracy, sensitivity and specificity of 0.982, 0.861 and 0.975 respectively, indicating a high level of reliability in classifying different arrhythmia types. The model's performance was further analysed using SHAP, which provided insights into the contributions of individual ECG samples to the model's predictions, leveraging the top 250 SHAP values for each arrythmia condition. The use of SHAP not only helped in

understanding the model's decision-making process but also highlighted the most significant features influencing predictions, as seen in the AF condition presented in Figure 21, with the absence of P waves being highlighted, as well as the T waves for the detection of irregular ventricular responses.

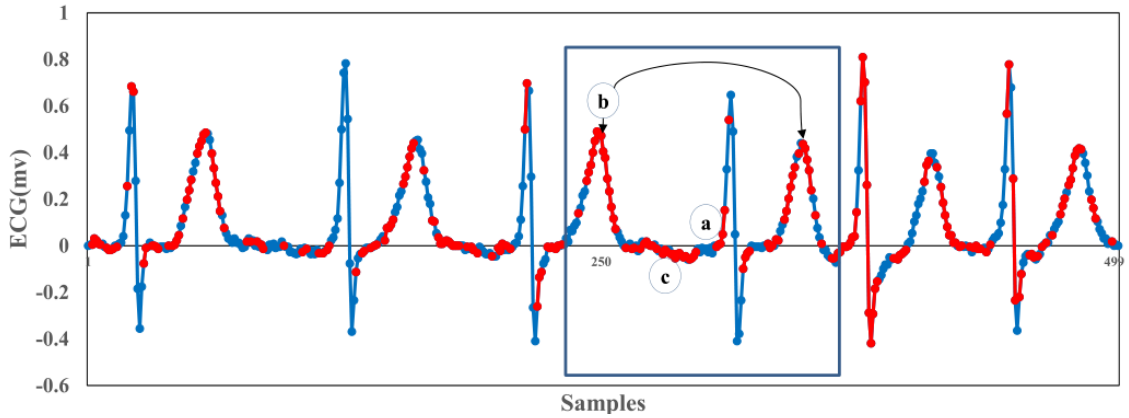


Figure 21 - AF ECG sample with overlaid SHAP values. Red dots represent the most important features, and the blue dots the less important features. Absence of P waves, Irregular ventricular response and Irregular baselines are identified by the letters a, b and c respectively. Adapted from [68].

Jaya Ojha et al. adapted the model developed by Atul Anand et al., known as the ST-CNN-5 model, to enhance the interpretability and performance of ECG data classification using different XAI methods on the PTB-XL dataset [71]. They adjusted the model's architecture by adding skip connections before the ReLU layer, implementing early stopping, and increasing the dropout rate to prevent overfitting. Additionally, they used ReduceLROnPlateau to dynamically adjust the learning rate and aid in model convergence. The proposed model, “ST-CNN-5 new”, achieved an overall good performance, achieving a slight improvement in accuracy (0.891), although the specificity (0.934), precision (0.798) and AUC (0.932) were just in line with the state-of-the-art models, exhibiting also lower recall (0.693) when compared to other models. The interpretability of the model is assessed using three XAI techniques: SHAP, Grad-CAM, and LIME. Among these, SHAP stands out for its ability to emphasize key ECG features, whereas Grad-CAM has difficulty capturing detailed patterns, and LIME seems to struggle with non-linear relationships. Using an example from the article (Figure 22), one of the characteristics of LBBB is a deep S wave in Lead V1, observed in the SHAP and Grad-CAM visual explanation, even though Grad-CAM seems to give a greater importance to the PQ segment. On the other hand, LIME does not capture any relevant features, and since its generated by RecurrentTabularExplainer, maybe it’s not adequate for use in signal data. Overall, the work emphasizes the importance of integrating interpretable AI models into clinical workflows to enhance trust and understanding

among healthcare professionals, employing various XAI methods to provide different insights.

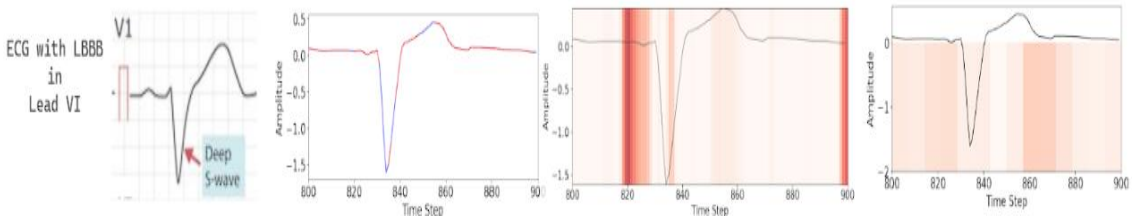


Figure 22 - LBBB ECG sample from PTB-XL database. From left to right: Characteristic feature of LBBB in lead V1, overlaid SHAP explanation (red and blue colour represents the most and less important features respectively), Grad-CAM heatmap and LIME heatmap on top of the ECG signal (darker areas are the most important features). Adapted from [71].

Switching now to XAI studies on PCG data, Theekshana Dissanayake et al. developed a deep learning model for the classification of heart sounds, particularly focusing on abnormal heart conditions, testing if the segmentation of the signals has a positive impact on the model performance or not [72]. One of the datasets utilized in this study, PHY16 dataset, consists of 3.240 heart sound recordings from various clinical and non-clinical environments, collected from 764 patients, ranging from 5 seconds to slightly over 120 seconds, labelled as either normal (2575 signals) or abnormal (665 signals). The patients had various health conditions, primarily heart valve defects and CAD, but such annotations are not included in the dataset [50]. The authors further evaluated the model using the murmur database developed by Yaseen et al., which includes 200 normal heart sounds and 800 abnormal murmur sounds, with 200 samples for each condition: Aortic Stenosis (AS), Mitral Regurgitation (MR), Mitral Stenosis (MS), and Mitral Valve Prolapse (MVP) [73]. The top-performing model architecture contains three 2D convolutional layers, two max pooling layers and two fully connected layers, making it a simple yet efficient design. It is also important to note that the tested segmentation steps (using a pretrained LSTM) did not improve model performance. Feature extraction was performed using pre-extracted representations from MFCCs, which are crucial for interpreting the temporal nature of the heart sounds. Performance results indicate that the model achieved an accuracy of 0.998 (±0.002), with a sensitivity of 0.998 (±0.001) and specificity of 0.997 (±0.002) on the PHY16 dataset. On the murmur database, the model attained an accuracy of 0.999 (±0.002), a sensitivity of 0.999 (±0.003) and specificity of 0.999 (±0.003), showing that the model is stable and robust enough to perform well with unseen data. Furthermore, the study also employed the SHAP algorithm and Occlusion Maps: SHAP to evaluate the contribution of each feature in relation to the final output; Occlusion Maps to observe the prediction variation if a certain

region/feature is taken out in the input, to find the importance of the taken feature. In Figure 23 we can inspect an example from the article, using an abnormal patient, where the Shapley values indicate that features in the vicinity of the S1 and S2 fundamental sounds has a greater positive or negative contribution to the final prediction. In line with the SHAP results, we can also observe that, in Occlusion Maps, even when the S1 location features are hidden, the model can still produce accurate predictions, since in the majority of explanations, the calculated probabilities exceed 0.9 (decision boundary for an abnormal classification).

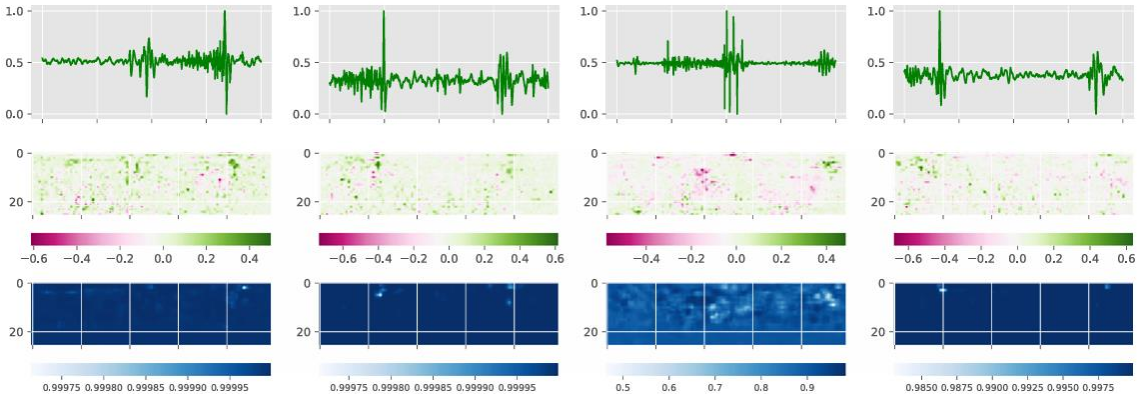
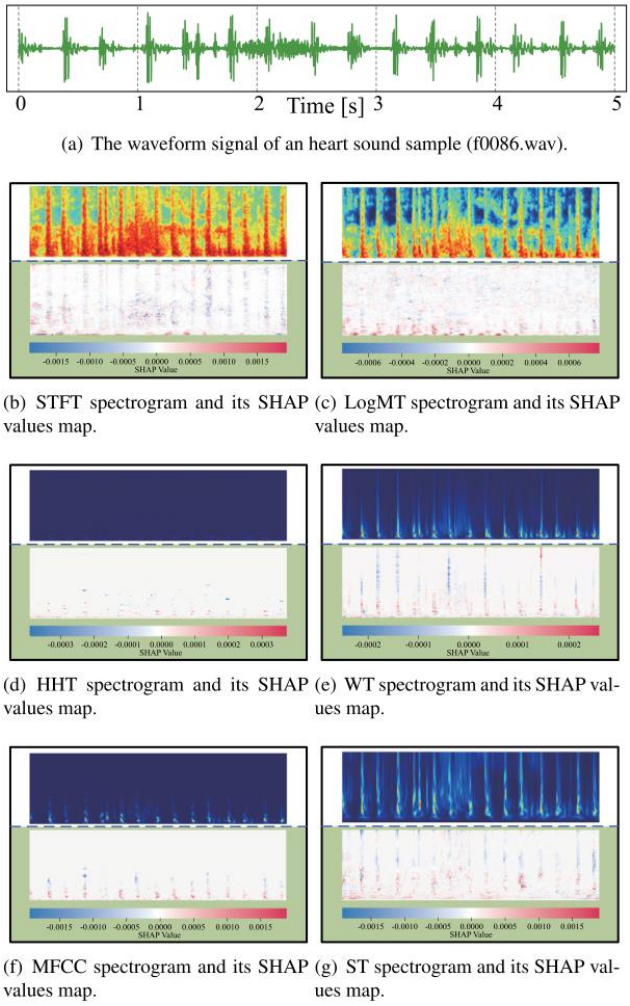


Figure 23 - Four abnormal PCG samples from PHY16 database. From top to bottom: ECG signal, Shapley values and Occlusion maps (both techniques applied to the MFCC of the signal). In SHAP, negative contributions are in pink and positive contributions are in green. Occlusion Maps displays the model output (between 0.0 and 1.0, normal-abnormal) if the input region is masked. Adapted from [72].

Zhihua Wang et al. aimed to enhance heart sound abnormality detection by combining time-frequency representations with pre-trained deep learning models, specifically focusing on a subject-independent database, in this case the PHY16 database, and eliminating the need for pre-processing [74]. The study employs a deep learning architecture based on the VGG19 model, which is fine-tuned specifically for this task. To achieve this, the authors substituted all layers following the final max pooling layer with a flattening layer, two fully connected layers with 128 and 32 neurons respectively, and a SoftMax layer for classification into two categories. Additionally, they incorporated regularization terms into the fully connected layers to mitigate overfitting. During the retraining phase on the extracted images, the convolutional layer parameters are kept unchanged, with only the parameters of the subsequent layers being updated. Feature extraction is carried out using various time-frequency analysis techniques, such as continuous wavelet transform (WT), Log-Mel Transformation (LogMT), MFCC, short-time Fourier transformation (STFT), Hilbert–Huang transformation (HHT) and Stockwell transform (ST), which convert heart sounds into 2D images for classification. Performance evaluations across six data division schemes reveal that the ST provides

the highest mean accuracy (MAcc) of 0.652 in data division scheme 6, along with a specificity of 0.529 and sensitivity of 0.775. Additionally, the study demonstrates that the average MAcc across different data division schemes is 0.609 (with 0.663 of sensitivity and 0.555 of specificity), underscoring the robustness of the method. This study also applied SHAP on each of the time-frequency representations for a normally labelled patient. It concluded that the SHAP maps for ST, WT, and MFCC are clearer and less affected by environmental noise during acquisition, providing a more accurate representation of heart sounds. The highest SHAP values are predominantly found in the S1 and S2 heart sounds, with minimal contribution from systole and diastole.



(a) The waveform signal of an heart sound sample (f0086.wav).

(b) STFT spectrogram and its SHAP values map.

(c) LogMT spectrogram and its SHAP values map.

(d) HHT spectrogram and its SHAP values map.

(e) WT spectrogram and its SHAP values map.

(f) MFCC spectrogram and its SHAP values map.

(g) ST spectrogram and its SHAP values map.

Figure 24 - Six time-frequency representation of a normal PCG sample from PHY16 database. The signal waveform is depicted in (a), with spectrograms at the top and SHAP value maps at the bottom of each representation. In the SHAP value maps, blue indicates negative contributions, white indicates zero contribution, and red indicates positive feature contributions to the model's prediction. Adapted from [74].

Additionally, it was observed that higher positive SHAP values are concentrated in the lower frequency range, while higher negative SHAP values are in the higher frequency

range (see Figure 24). Overall, this work highlights the potential of deep learning and time-frequency analysis in enhancing the detection of heart sound abnormalities, although the performance is still not good for clinical application.

Anandita Bhardwaj et al. developed a deep learning model to classify various types of valvular heart disease using PCG signals, aiming to enhance clinical decision-making through improved accuracy and interpretability [75]. The study utilized the murmur dataset and PHY16, employing the Morlet CWT as the preferred time-frequency representation of the PCG signals. The proposed model is a 2D CNN architecture optimized for feature extraction from these time-frequency representations. It consists of a normalization layer, two 2D convolutional layers (each followed by a ReLU layer), a max pooling layer, a fully connected layer, and concludes with a SoftMax output layer with cross-entropy loss for each of the five output classes. The model achieved a mean accuracy, precision, and recall of 0.983 ± 0.010, 0.958 ± 0.026, and 0.960 ± 0.024, respectively, during 5-fold cross-validation on the Murmur database, and a mean accuracy of 0.931 on the PHY16 database. For explainable AI techniques, the authors applied Occlusion Maps (for local explanation) and Deep Dream Images (DDI, for global interpretation) across all five PCG classes. DDI adjusts the input image to maximize the activation of certain neurons for a selected class, iterating this process a fixed number of times to amplify detected patterns and maximize the gradients added between the loss and the initial image, thereby generating a class-specific image [76]. In Figure 25, a Morlet CWT of an aortic stenosis patient is shown, followed by four occlusion maps. High-intensity values are observed in regions corresponding to the first cardiac cycle in time (horizontal axis) and similar frequency ranges (vertical axis). Notably, none of the occlusion maps contain negative intensity values, demonstrating the network's capability to avoid misclassifications. The model examines different regions of the CWT in each class, further assuring the network's robustness and reliability in identifying relevant features across different classes.

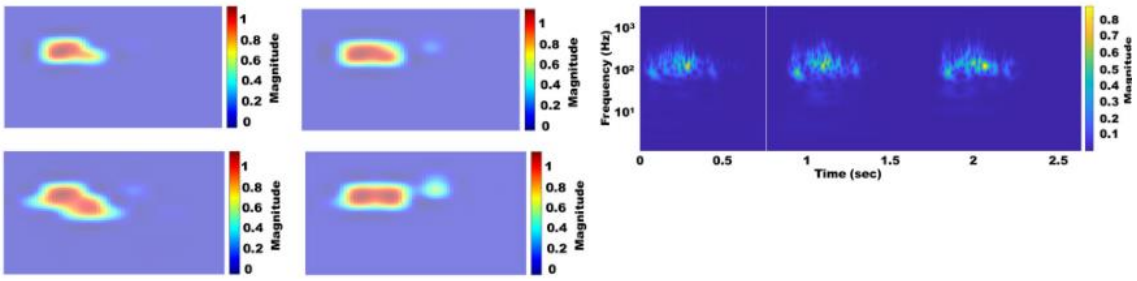


Figure 25 - Occlusion maps of an aortic stenosis PCG sample from the Murmur database. The left panel displays four occlusion maps for the class, while the right panel shows a Morlet CWT for the same class. Adapted from [75].

Chelluri Divakar et al. developed an XAI framework for diagnosing valvular heart diseases using PCG signals [77]. The dataset used was the Murmur database, which was later augmented to double its original size, ensuring balanced representation across all five classes. The architecture combines a 2D CNN module, consisting of twelve 2D convolutional layers, each followed by batch normalization, ReLU activation, max or average pooling layers, and a final dropout layer, with an LSTM framework, five fully connected layers, and a SoftMax output layer, creating a CNN-LSTM network. Features were extracted using Mel-spectrograms generated from PCG signals. The CNN-LSTM model achieved an accuracy of 0.975 on a test set of 400 images, with individual class accuracies of 0.988 for AS and MR, 0.962 for MS and MVP, and 0.975 for the normal class. The model's interpretability was enhanced using the Grad-CAM technique applied to the final convolutional layer, which visualized areas of focus in the input Mel-spectrogram images during classification, for both correct and incorrect classifications. Figure 26 illustrates its application on correctly classified signals, showing that the model primarily focuses on low-frequency patterns, enabling accurate categorization for each specific class and further enhancing transparency.

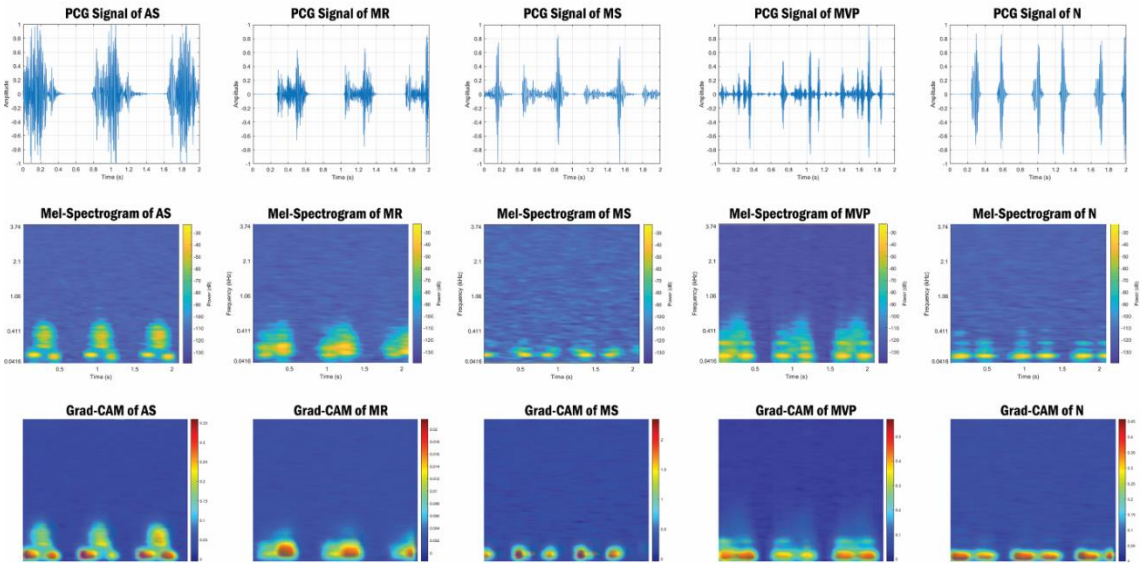


Figure 26 - Examples of Grad-CAM maps for all five correctly classified PCG classes from the Murmur database. The top row displays the original signal representations for each of the five classes, followed by their corresponding Mel-spectrograms in the middle row, with the Grad-CAM maps shown on the bottom row. Adapted from [77].

In the final XAI-related work, Shuaizhong Li et al. introduced an end-to-end heart sound classification approach designed to assist in diagnosing congenital heart disease (CHD) without the need for manual feature extraction [78]. The study utilized two datasets, a private dataset, referred to as Dataset A, which included 5,000 heart sound recordings, each lasting 20 seconds, collected from 1,000 patients (five-point signal

auscultation at a 5000 Hz sampling rate). This dataset comprised 2,500 abnormal samples (including conditions like atrial septal defects, ventricular septal defects, and patent ductus arteriosus) and 2,500 normal samples, all from adolescents aged 2 to 18, with recordings focused on the mitral valve, pulmonary valve, aortic valve, aortic valve second, and tricuspid valve areas. The second dataset, Dataset B, was the PHY16 dataset. The researchers developed a novel architecture combining a multi-scale dense network (MDN) with a multi-head recurrent neural network (MARNN). The MDN utilized multi-scale convolution with varying receptive fields to integrate information from different scales, enhancing the capture of both global features and local details, with the DenseNet architecture adapted to one dimension for compatibility with 1D signals. The MARNN incorporated a multi-head self-attention mechanism (MA) to extract features and temporal correlations from the input data, positioned between two layers of Bi-GRUs that extracted sequence correlation features. The Bi-GRUs captured the global features of the sequence, while the MA mechanism amplified differences between key and redundant features by assigning different attention weights. Feature extraction involved segmenting the raw 2-second heart sound recordings with a 1-second overlap, which were then inputted into the MDN-MARNN model. The results demonstrated that the proposed method achieved an F-beta score (similar to the F1 score, but with a beta parameter adjusting the weight of specificity relative to sensitivity) of 0.951 and an accuracy of 0.954 on Dataset A. On the PHY16 dataset, the method produced an F-beta score of 0.938 and an accuracy of 0.930, highlighting the model's robustness on unseen data. The SHAP algorithm was employed for explainability, with Figure 27 illustrating the visualization and interpretation of the model's prediction on a True Positive case. The visualization showed that the negative (blue) and positive (red) contributions were concentrated in the S1 and S2 regions, indicating that the model effectively learned the features without the need for extensive preprocessing, apart from basic segmentation.

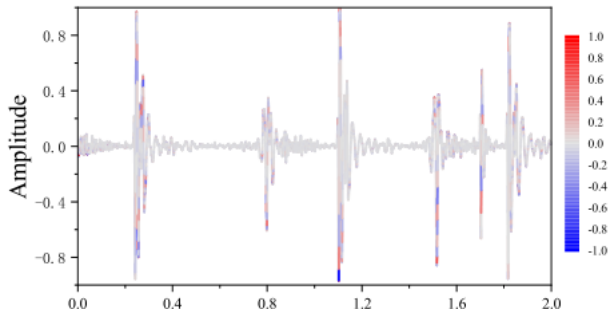


Figure 27 - Example of overlaid SHAP values on a PCG signal, on a True Positive (abnormal) case. Red colour indicates a positive contribution, and the blue colour represents a negative contribution. Adapted from [78].

3.3. Discussion

Given that the objective of this research is to apply a multimodal model for classifying synchronous ECG and PCG data, followed by the application of XAI methods to explain and visualize the classification process, a comprehensive review of relevant literature was conducted. The search results for deep learning multimodal models revealed that there are only a few studies utilizing synchronized ECG and PCG data, and nearly all of these have been discussed in this section. One of the primary limitations identified in these studies is the scarcity of publicly available synchronized ECG and PCG datasets; the only notable datasets are PHY16 and EPHNOGRAM. However, the EPHNOGRAM dataset is primarily focused on comparing heart rate variability during physical stress tests, making it unsuitable for our work. Moreover, few studies provide external validation of model performance, with notable exceptions being the works by Haozhan Han et al., who tested their model on both the EPHNOGRAM and a private dataset, and Haobo Zhang et al., who validated their model on the PHY16 dataset and confirmed its stability using a private dataset. The lack of extensive synchronous ECG and PCG datasets poses a challenge for effective model generalization and robustness testing. Additionally, another limitation is the division between complex signal preprocessing techniques and/or intricate models. Complex signal preprocessing can be problematic as, without proper data generalization and a large dataset, it is difficult to predict if the pipeline will generalize well. Complex models also tend to incur high computational costs, making them unsuitable for devices with limited processing power, such as battery-operated multimodal stethoscopes. Among the reviewed studies, CNN, LSTM, and Attention modules are commonly employed in network architectures. However, since LSTM and Attention modules require more computational resources [79], this work will focus on 1D and 2D CNN architectures with simpler preprocessing techniques.

In terms of XAI techniques, the literature review shows that more studies focus on applying these techniques to ECG signals than to PCG signals, with only one study by Monjur Morshed et al. applying the technique to both synchronous signals, though without significant depth. The most commonly used XAI techniques identified in this review are SHAP, which is model-agnostic, and Grad-CAM, which is specifically made for CNNs. Studies reveals that SHAP provides more fine-grained details compared to Grad-CAM. LIME is also considered a state-of-the-art technique; however, in the work by Jaya Ojha et al., they used RecurrentTabularExplainer, which is not designed for time series explanations, and currently, there is no official LIME package for time series or bio-signals. Therefore, this study will compare the Grad-CAM technique with SHAP to

evaluate the differences between the two. A summary of the most important aspects of the reviewed articles is presented in Table VI.

Table VI - Relevant information from the selected final articles. Acc, Sen, Spe and Prec refer to Accuracy, Sensitivity, Specificity and Precision, respectively. Note: Sensitivity and Recall are different terms for the same metric

Article	Application/ Objective	Dataset	Methodology/ Algorithm applied	XAI technique	Performance
[9]	CAD detection	Private dataset: 195 subjects (135 with CAD and 60 non-CAD)	Fully connected layers, CNN, Bi-GRU with Attention	Not applicable	Acc: 0.956±0.010 Sen: 0.985±0.012 Spe: 0.892±0.031
[8]	Binary classification (normal abnormal signals)	PHY16 dataset training folder A (described in materials)	CNN, Multilayer Perceptron	Not applicable	Acc: 0.904 Sen: 0.947 Spe: 0.750 AUC: 0.911
[51]	Binary classification (normal abnormal signals)	PHY16 dataset training folder A (described in materials)	CNN, LSTM, GA for feature selection, SVM classifier	Not applicable	Acc: 0.873 Sen: 0.903 Spe: 0.845 AUC: 0.936 F1-score: 0.874
[52]	CAD detection	Private real-world dataset: 195 subjects (135 with CAD and 60 non-CAD)	1D CNN, Bi-GRU with Attention, 2D CNN	Not applicable	Acc: 0.965±0.012 Sen: 0.994±0.008 Spe: 0.901±0.036
[54]	Binary classification (normal abnormal signals)	PHY16 dataset training folder A (described in materials)	1D CNN	Grad-Cam++	Acc: 0.951 Recall: 0.951 Prec: 0.951 Spe: 0.909 AUC: 0.990 F1-score: 0.950
[55]	Classification task	EPHNOGRAM dataset and Private real-world	ResNet1d wang, GAP layer, MCT blocks, Tokens from each	Not applicable	(EPHNOGRAM/private dataset) Acc: 0.983/ 0.892

		dataset with 123 patients	block are concatenated and passed through a classifier		Recall: 0.988/ 0.811 F1-score: 0.987/ 0.814
[58]	Classification in normal signal, common heart disease and special heart disease	Private real-world dataset with 41 abnormal patients and 80 healthy control samples divided in 3 classes	SE module (with a GAP layer), CNN, ViT blocks	Not applicable	Acc: 0.972 Recall: 0.972 F1-score: 0.971
[59]	Detection of cardiovascular diseases	PHY16 dataset training folder A (described in materials) and a Private real-world dataset: 212 pairs of ECGs and PCGs with valvular diseases and 380 pairs of ECGs and PCGs without valvular diseases	Modality specific encoders (with CNN and LSTM), Progressive dense fusion encoder, Co-learning strategy to support the progressive dense fusion approach, Combination of decision-level with feature-level fusion to leverage the data, Averaging output probability of the 3 encoders to obtain prediction	Not applicable	(PHY16 dataset/ private dataset) Acc: 0.944±0.030/ 0.855±0.030 Recall: 0.949±0.051/ 0.830±0.058 Spe: 0.940±0.050/ 0.879±0.050 AUC: 0.973±0.026/ 0.920±0.025
[60]	Binary classification (normal	PHY16 dataset training folder	CNN, Residual blocks,	Not applicable	Acc: 0.931 Recall: 0.929 Spe: 0.933 AUC: 0.967

	abnormal signals)	A (described in materials)	SE attention module (SE-ResNet), SE-LSTM-Net (with a SE-Block, pooling and Bi-LSTM layers) Fully connected layers, PCA (for fusion and feature selection), SVM classifier		F1-score: 0.931
[61]	Binary classification (normal abnormal signals), reducing the computational cost and memory requirements	PHY16 dataset training folder A (described in materials)	Affine transform layer, 2D CNN, Bottleneck blocks from MobileNetV3-small (using 1D instead of 2D convolutions), Fully connected layer, Majority voting for classification	Not applicable	Acc: 0.970±0.014 Recall: 0.975±0.018 Spe: 0.965±0.022 Prec: 0.966±0.021 F1-score: 0.970±0.014 AUC: 0.986±0.005
[62]	Multiclass Classification 12-lead ECG	CPSC2018 dataset, with 6,877 12-ECG recordings, composed by 9 different cardiac arrhythmias	Modified VGG-16	Grad-CAM	Subset Acc: 0.962 Recall: 0.949 Prec: 0.986 F1-score: 0.967
[64]	Multiclass Classification 12-lead ECG	PTB-XL (21837 samples from 18885	ST-CNN-GAP-5 (with GAP instead of Max pooling in the last	SHAP (top 500 values)	(PTB-XL/ CPSC2018 at 500 Hz) Acc: 0.897/ 0.959

		patients, including healthy patients and various heart conditions) plus CPSC2018 dataset	convolutional layer)		Macro Recall: -/ 0.953 Macro Prec: -/ 0.954 Micro F1-score: 0.793 Macro F1-score: -/ 0.954 Macro AUC: 0.934/ 0.995 Macro AUPRC: 0.834/ 0.985
[66]	Multiclass Classification 3 selected ECG-leads	Chapman dataset: 10,646 12-lead ECG samples covering 11 common rhythms and 67 additional cardiovascular conditions, plus CPSC2018 dataset	1D-SEResNet backbones for feature extraction, Lead-wise Attention module for robust feature aggregation	Lead wise Grad-CAM	(Chapman/ CPSC2018) Acc: 0.987/ 0.963 Recall: 0.970/ 0.786 Prec: 0.974/ 0.821 F1-score: 0.972/ 0.800
[68]	Multiclass arrythmia Classification 2 lead ECGs	MIT-BIH arrhythmia database (2 channels ECG from 47 subjects) and the LTAF database (84 two lead ECG of subjects with paroxysmal or sustained	11-layer network that combines 1D CNN and LSTM architectures	SHAP (top 250 values)	Acc: 0.982 Recall: 0.861 Spe: 0.975

		atrial fibrillation)			
[71]	Multiclass Classification 12-lead ECG	PTB-XL	ST-CNN-5 new, same as ST-CNN-GAP-5 but with added skip connections, early stopping, increasing dropout rate and usage of ReduceLROnPlateau	SHAP, Grad-CAM and LIME	Acc: 0.891 Recall: 0.693 Prec: 0.798 Spe: 0.934 AUC: 0.932
[72]	Binary classification (normal abnormal signals)	PHY16 dataset (3.240 heart sound recordings, with normal/abnormal labels) plus Murmur database (200 normal heart sounds and 800 abnormal murmur sounds)	2D CNN, Fully connected layers	SHAP Occlusion Maps	(PHY16/ Murmur database) Acc: 0.998±0.002/ 0.999±0.002 Recall: 0.998±0.001/ 0.999±0.003 Spe: 0.997±0.002/ 0.999±0.003
[74]	Binary classification (normal abnormal signals)	PHY16 dataset	Modified VGG19 model	SHAP	Mean Acc: 0.609 Recall: 0.663 Spe: 0.555
[75]	Multiclass Classification on PCG signals and	Murmur database (multiclass classification) plus PHY16	2D CNN, Fully connected layer	Occlusion maps (local explanation)	(Murmur database/ PHY16) Acc: 0.983±0.010/ 0.931

	binary classification	dataset (binary classification)		Deep Dream Images (global explanation)	Recall: 0.960±0.024 Prec: 0.958±0.026
[77]	Multiclass Classification on PCG signals	Murmur database	2D-CNN-LSTM architecture	Grad-CAM	Acc: 0.975
[78]	Binary classification (normal abnormal signals)	Private Dataset with 5,000 heart sound recordings, from 1,000 patients with 2,500 abnormal and 2,500 normal samples	MDN-MARNN model, multi-scale dense network with a multi-head recurrent neural network (with self-attention and Bi-GRU), adapted for compatibility with 1D signals	SHAP	(Private dataset/ PHY16 dataset) Acc: 0.954/0.930 Recall: 0.940/ 0.945 Spe: 0.968/0.926 F-beta score: 0.951/0.938

4. Methods and Materials

This chapter details the methods and materials employed in this study. Section 3.1 begins with a description of the dataset used, specifically the PHY16 training set A. Section 3.2 then explains the pre-processing pipeline implemented for both ECG and PCG signals. Following this, Section 3.3 outlines the architectures of the models used. In Section 3.4, the training and evaluation processes are discussed. Finally, Section 3.5 provides an explanation of the selected XAI algorithms, Grad-CAM and SHAP.

4.1. Materials

Publicly available datasets for joint signal analysis are currently quite scarce. In 2016, PhysioNet introduced a dataset focused on CVDs, specifically the training set "A," which contains both synchronous ECG and PCG signals. Of the 409 PCG recordings included, four were excluded due to the absence of corresponding ECG data. These recordings were obtained from 121 subjects across nine different positions. The subjects were divided into five categories: normal control (117 recordings from 38 subjects), MVP murmurs (134 recordings from 37 patients), benign murmurs (118 recordings from 34 patients), aortic disease (AD) (17 recordings from 5 patients), and miscellaneous pathological conditions (MPC) (23 recordings from 7 patients).

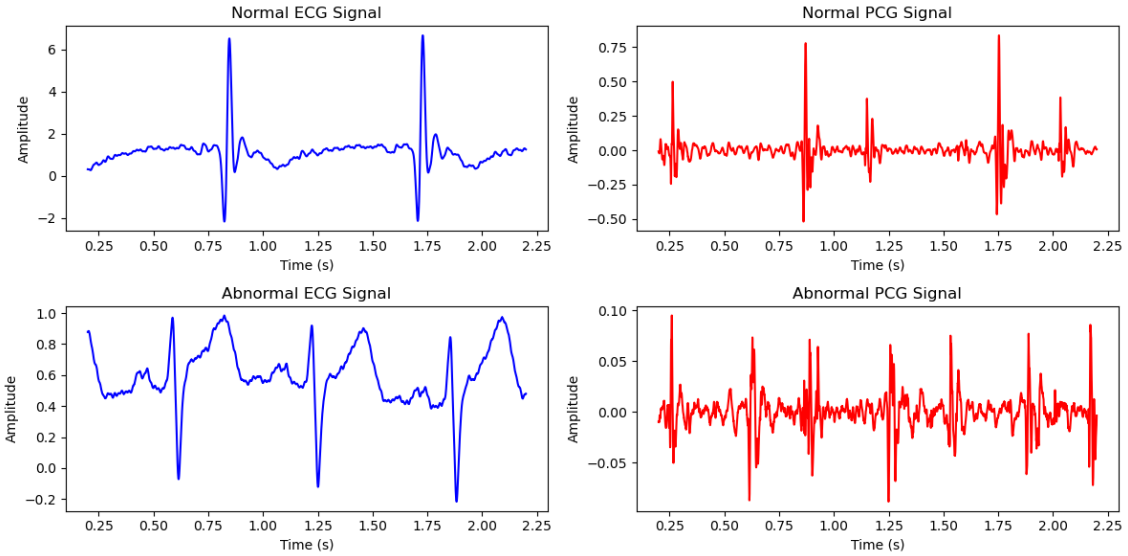


Figure 28 - Comparison between synchronous normal ECG and PCG signals with synchronous abnormal ECG and PCG signals. Data taken from PHY16 training set “A”.

Since the records have only binary labels (only distinguished as normal/abnormal), it is not possible to link signals to specific conditions or to identify which record belongs to which patient. Thus, all recordings are treated as a single lead per patient. The final dataset comprises 405 signals, all sampled at 2000 Hz, with durations ranging from 9 to 37 seconds. It includes 113 normal signals (about 28%) and 292 abnormal signals (about 72%), resulting in a highly imbalanced dataset [50].

4.2. Preprocessing

4.2.1. Heart sounds (PCG)

Each PCG signal was segmented into non-overlapping 2-second intervals, followed by the application of a 4th-order Butterworth bandpass filter, incorporating a high-pass filter with a cutoff frequency of 25 Hz and a low-pass filter with a cutoff frequency of 400 Hz. A spike elimination procedure was then conducted by dividing the recording into 500-millisecond windows, where the maximum absolute amplitude (MAA) was determined for each window. If an MAA exceeded three times the median MAA value, the following steps were performed: (1) the window with the highest MAA was identified; (2) the peak of the spike, indicated by the MAA, was located; (3) the spike's start was marked at the last zero-crossing point before the peak; (4) the end of the spike was marked at the first zero-crossing point after the peak; (5) the identified noise spike was replaced with zeros, and the process resumed at step 2. If no MAA exceeded the threshold, the procedure was completed [80].

After that task, spectral characteristics of the PCG signals were extracted using static, delta, and delta-delta Mel-frequency cepstral coefficients (MFCCs). The cepstral transformation was employed to detect periodic patterns and separate signal components that overlap in the frequency domain. To compute MFCCs, the Discrete Cosine Transform (DCT) was applied to the Fourier transform power coefficients, which had been mapped onto the Mel scale using triangular bandpass filters [81]. The Mel scale approximates a linear relationship for frequencies below 1 kHz and a logarithmic relationship for frequencies above 1 kHz, mirroring human auditory perception [82]. This representation is particularly beneficial for analysing PCG signals, as heart sounds and murmurs predominantly occur in the lower frequency ranges. The use of MFCCs effectively reduces the dimensionality of PCG audio signals while preserving critical information for pattern recognition.

The PCG signals were first processed using Mel spectrograms, applying a Hann window to each frame, calculated as follows:

$$w(n) = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi n}{N-1}\right)\right] \tag{4.1}$$

where N is the length of the window and n is the sample index [83]. During this phase, various window lengths (128 ms, 64 ms, and 32 ms) and hop lengths (64 ms and 16 ms) were tested for MFCC computation. Ultimately, a window length of 32 milliseconds and a hop length of 16 milliseconds were selected to achieve an optimal balance between time and frequency resolution.

Next, the power spectrum was mapped onto the Mel scale, which is given by:

$$f_m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{4.2}$$

where f represents the standard frequency scale. The power values of the Mel spectrogram were subsequently converted to a decibel (dB) scale to enhance visualization and interpretation. This conversion is performed using:

$$S_{dB}(x) = 10 \cdot \log_{10}\left(\frac{S(x)}{S_{\text{ref}}}\right) \tag{4.3}$$

where $S(x)$ is the power value at a particular frequency or time index and S_{ref} is a reference power level, in this case, set to the maximum value of the Mel spectrogram to normalize the spectrogram values by converting the maximum value of the input spectrogram to 0 dB [84].

To extract cepstral features, the Mel spectrogram in dB scale was subjected to the Discrete Cosine Transform (DCT) Type-2, resulting in MFCCs, calculated as:

$$C(m) = \sum_{i=1}^{N} cos\left(\frac{m(i-0.5)\pi}{N}\right) E_i, \quad \text{for } m = 1,2,\dots,L \tag{4.4}$$

where $C(m)$ is the m$^{\text{th}}$ Mel-scale cepstral coefficient, N is the number of triangular bandpass filters, L is the number of Mel-scale cepstral coefficients (in this application, 13 were used), and E_i is the log energy output from the i^{th} triangular band pass filter [85].

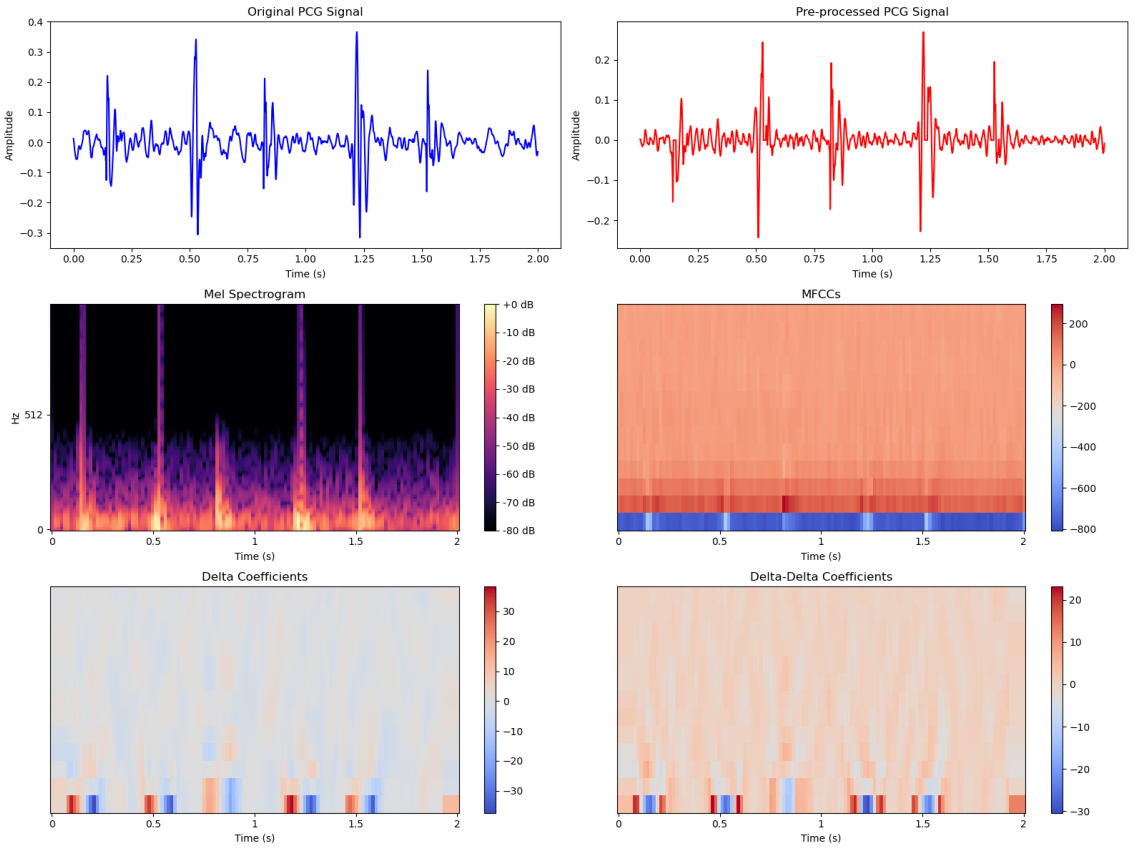


Figure 29 - Representation of the original normal PCG signal, together with the pre-processed signal, Mel spectrogram, MFCCs, Delta, and Delta-Delta coefficients.

Additionally, to capture the dynamic aspects of the PCG signals, the first and second-order differences of the MFCCs were calculated, known as delta and delta-delta coefficients. These derivatives provide insights into the temporal evolution of the spectral features [86]. The combined use of MFCCs, delta, and delta-delta coefficients forms a comprehensive feature set that effectively represents both the static and dynamic characteristics of the PCG signals. Figures 29 and 30 illustrate the evolution of normal/abnormal PCG signals preprocessing tasks. The pre-processed PCG signal corresponds to the signal after filtering and spike elimination, but before MFCCs calculation. In normal signals, the S1 and S2 peaks are clearly identifiable, with corresponding high-energy peaks represented in the Mel spectrogram, MFCCs, delta, and delta-delta coefficients.

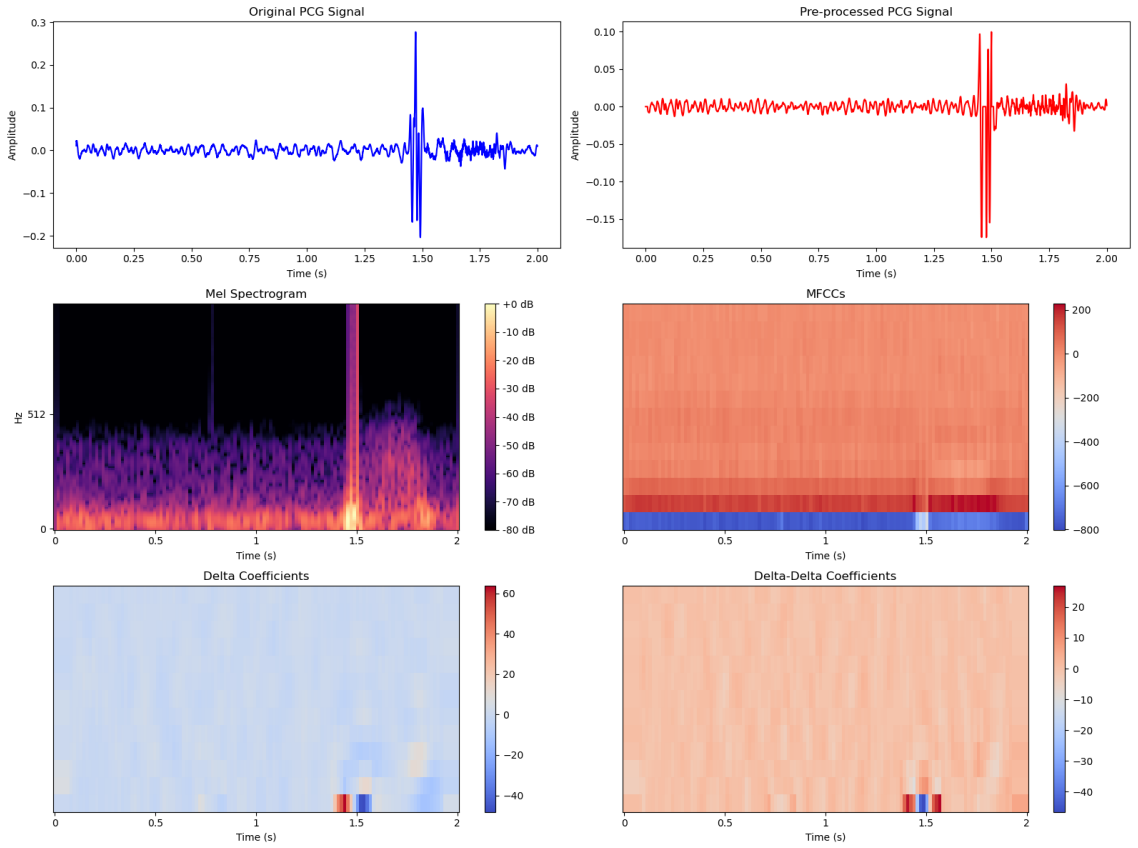


Figure 30 - Representation of the original abnormal PCG signal, together with the pre-processed signal, Mel spectrogram, MFCCs, Delta, and Delta-Delta coefficients.

4.2.2. Electrocardiogram (ECG)

In the ECG preprocessing stage, 15 records with missing data points were identified, requiring linear interpolation to fill the gaps. After experimenting with various filter settings, a 2-35 Hz bandpass filter proved optimal compared to alternatives, such as the 10-35 Hz range. Normalization was also assessed, demonstrating improved performance compared to non-normalized segments. Following these steps, a preprocessing pipeline was established. A fourth-order Butterworth bandpass filter was applied with a 35 Hz low-pass and 2 Hz high-pass filter to address baseline drift [87]. The signal was then down sampled to 500 Hz and normalized to a range between 0 and 1, before being divided into 2-second non-overlapping segments.

To convert the 1D ECG signal into a 2D representation, both Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT) were applied following the initial 1D preprocessing. The CWT was performed using the Morlet wavelet, known for its effectiveness in detecting transient features in non-stationary signals [88]. This

process generated a time-scale matrix representing the convolution of the Morlet wavelet at multiple scales, providing a detailed view of frequency content. The CWT of a signal $x(t)$ is defined as:

$$CWT(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-b}{a} \right) dt \tag{4.5}$$

where $CWT(a,b)$ is the wavelet coefficient at scale a and translation b, and $\psi(t)$ is the Morlet mother wavelet [89], given by:

$$\psi(t) = \pi^{-\frac{1}{4}} e^{-\frac{t^2}{2}} e^{-i\omega_0 t} \tag{4.6}$$

where ω_0 is the central frequency of the wavelet [90].

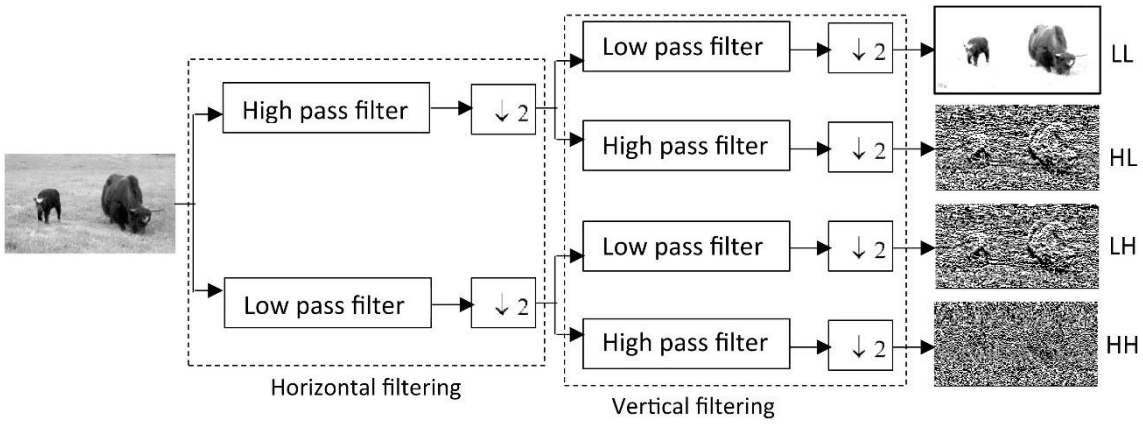


Figure 31 - Image decomposition using 2D Discrete Wavelet Transform (DWT), where high-pass and low-pass filtering followed by down sampling generates four sub-bands: LL, LH, HL, and HH. Adapted from [91].

The DWT was then applied to the CWT output using 2D Haar wavelets. DWT passes the signal through high-pass and low-pass filters, followed by down sampling, to decompose it into four sub-bands: LL, LH, HL, and HH, representing approximation and detail coefficients, as shown in Figure 31. The LL (approximation) sub-band captures the low-frequency components, while LH, HL, and HH represent horizontal, vertical, and diagonal high-frequency details, respectively [92]. These coefficients are calculated as:

$$\mathrm{LL} = \frac{1}{4} \sum_{i,j} I(i,j) \cdot \phi(2x-i) \cdot \phi(2y-j) \tag{4.7}$$

$$\mathrm{LH} = \frac{1}{4} \sum_{i,j} I(i,j) \cdot \phi(2x-i) \cdot \psi(2y-j) \tag{4.8}$$

$$\mathrm{HL} = \frac{1}{4} \sum_{i,j} I(i,j) \cdot \psi(2x-i) \cdot \phi(2y-j) \tag{4.9}$$

$$\mathrm{HH} = \frac{1}{4}\sum_{i,j} I(i,j) \cdot \psi(2x - i) \cdot \psi(2y - j) \tag{4.10}$$

where $I(i,j)$ is the input image (CWT matrix), ϕ is the scaling function (also known as the father wavelet), and ψ is the wavelet function (also known as the mother wavelet).The scaling function captures low-frequency content, while the wavelet function extracts high-frequency details [92], and they are represented as:

$$\phi(t) = \begin{cases} 1 & \text{if } 0 \leq t < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.11}$$

$$\psi(t) = \begin{cases} 1 & \text{if } 0 \leq t < \frac{1}{2}, \\ -1 & \text{if } \frac{1}{2} \leq t < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.12}$$

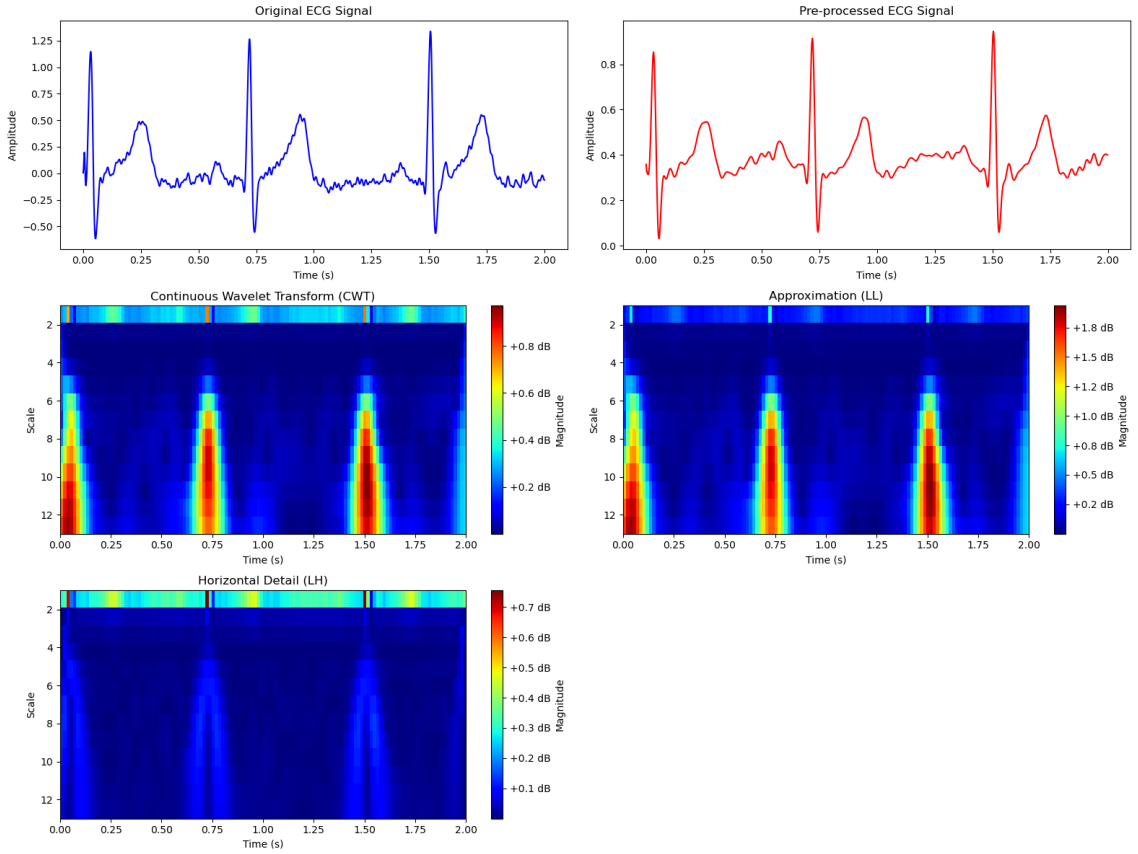


Figure 32 - Representation of the original normal ECG signal, together with the pre-processed signal, Continuous Wavelet Transform and the Approximation and Horizontal Detail coefficients from Discrete Wavelet Transform.

Only the LL and LH sub-bands were retained, as they provided the most significant signal information. Both sub-bands, together with the CWT, were resized to a

uniform (13, 126) shape using interpolation to match the requirements for later modelling methodology. This approach enables multi-resolution analysis, essential for accurate signal representation in subsequent classification tasks. Figures 32 and 33 displays the preprocessing stages for normal and abnormal ECG signals, highlighting prominent R peaks in both cases.

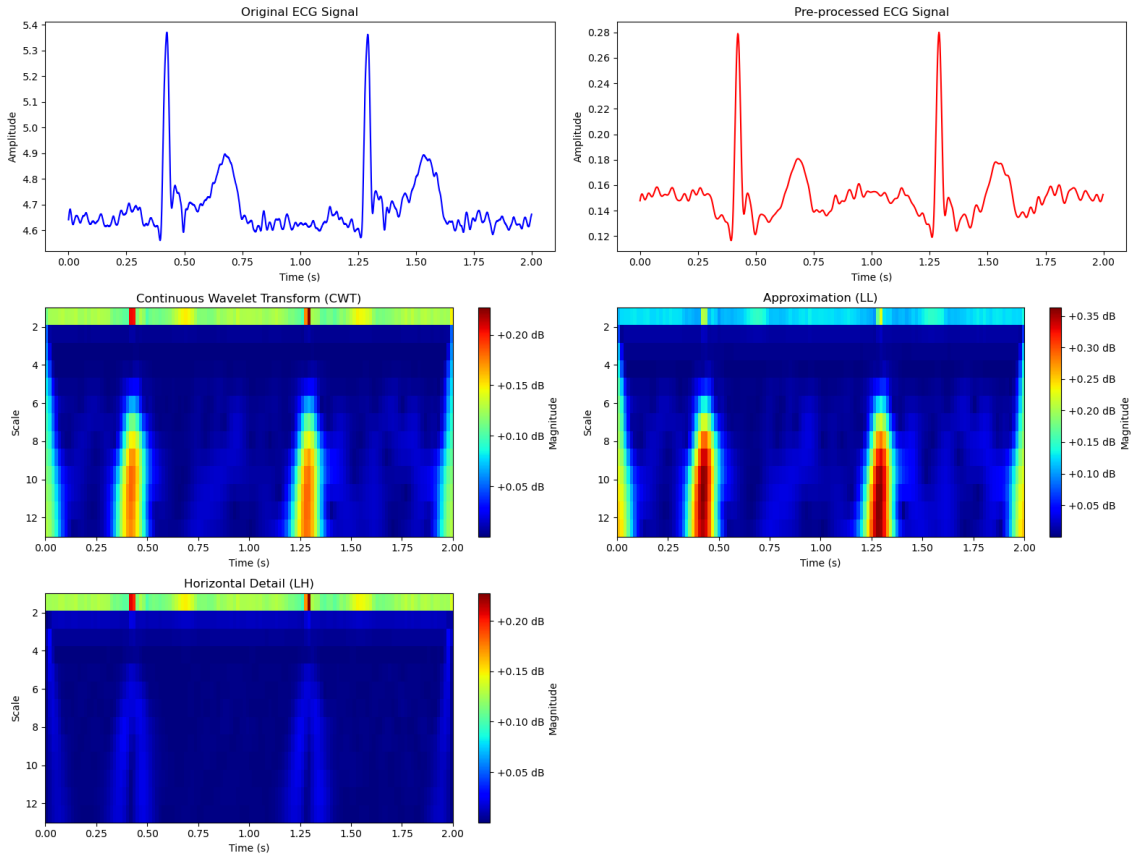


Figure 33 - Representation of the original abnormal ECG signal, together with the pre-processed signal, Continuous Wavelet Transform and the Approximation and Horizontal Detail coefficients from Discrete Wavelet Transform.

4.3. Modelling

In this study, five neural networks were evaluated: a 2D-CNN for PCG signals, a 1D CNN and a 2D CNN for ECG signals, as well as two hybrid models that integrate both modalities. The hybrid models consist of an early fusion (EF) network, combining the 2D CNN architectures for PCG and ECG signals, and a late fusion (LF) network, which merges a 1D CNN for ECG with a 2D CNN for PCG.

The 2D PCG model consists of two 2D convolutional layers, each followed by max-pooling and dropout layers (with a 20% dropout rate) to reduce overfitting. To further

prevent overfitting, a regularization term (L2 penalty) of 0.02 was applied to the loss function in the convolutional layers. These two regularization factors helped to control the model's complexity by limiting the weight' growth, improving its generalization. Batch normalization was applied after the first convolutional layer to improve training stability. The architecture concludes with a dense layer, another dropout layer, and an output layer. Detailed parameters for this network are provided in Table VII.

Table VII - 2D network parameters

Layers	Characteristics
Convolution kernel 1	5×5, stride 1×1
Convolution kernel 2	3×3, stride 1×1
Max pooling kernel	2×2, stride 1×1
Optimizer	Adam
Learning rate	1×10^{-3}
Batch size	32
Max number of epochs	300
Loss function	Binary Cross Entropy
Activation function	ReLU

Table VIII - 1D network parameters

Layers	Characteristics
Convolution kernel	6
Max pooling kernel 1	3, stride 2
Max pooling kernel 2	2, stride 2
Optimizer	Adam
Learning rate	1×10^{-3}
Batch size	32
Max number of epochs	300
Loss function	Binary Cross Entropy
Activation function	ReLU

The 1D ECG model includes three 1D convolutional layers, each followed by batch normalization, max-pooling layers, and a regularization term of 0.02 applied to each convolutional layer. The network ends with two dense layers and an output layer (as shown in Table VIII). The 2D ECG model mirrors the architecture of the 2D PCG model.

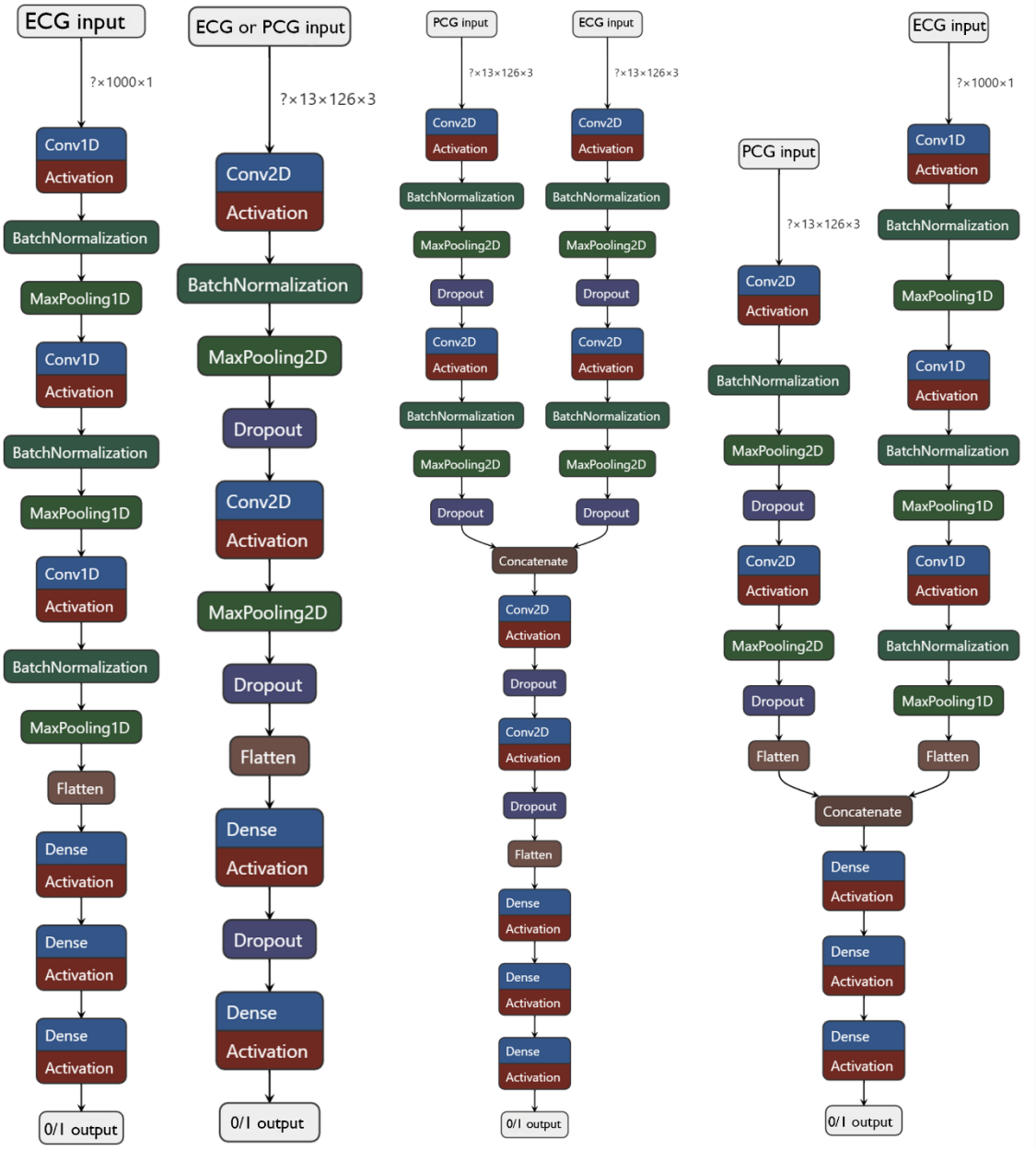


Figure 34 - Model's general scheme. From left to right: ECG 1D-CNN model; PCG/ECG 2D-CNN model; Early fusion model (using the backbones of ECG 2D-CNN and PCG 2D-CNN) and Late fusion model (using the backbones of ECG 1D-CNN and PCG 2D-CNN). Figures created using Roeder, L., Netron app, version 7.3.9. Retrieved from https://github.com/lutzroeder/netron.

For the hybrid networks, both ECG and PCG segments are processed as inputs. In the EF model, the PCG and ECG signals are handled by their respective 2D CNN architectures. After processing, the branches are concatenated, followed by two additional 2D convolutional blocks (kernel size: 3x3, stride: 1x1, regularization: 0.02), two dropout layers (20%), and finally, two dense layers and an output layer. The LF model processes each signal with the 2D CNN for PCG and the 1D CNN for ECG. The outputs from each network's convolutional layers are concatenated into a unified feature vector,

followed by two dense layers and the output layer. All models employ ReLU activation in all convolutional and dense layers, except for the output layer, which uses a Sigmoid activation function.

For all models, two versions of early stopping were tested: monitoring validation loss with a patience of 50 and 20 epochs. However, fixed training for 300 epochs yielded better performance. The architectures of all models are shown in Figure 34.

4.4. Training and Evaluation

The models' performance was assessed using a 5-fold cross-validation approach. To ensure consistency, data within each fold remained fixed across all tasks. Each fold was assigned as follows: one-fold for testing, three folds for training, and one-fold for validation. A critical step in data partitioning was taken to prevent signals from the same patient from appearing in multiple folds, and stratification was applied to maintain balanced class distributions. To address class imbalance, the loss function was modified by incorporating a weighting factor, enhancing the model's attention to the underrepresented class (normal). This weight adjustment was calculated as follows:

$$Weight_{Positives/Negatives} = \left(1 \div \sum Positives/Negatives\right) \times \left(\sum Total\ instances \div 2\right) \quad (4.13)$$

During training, all 2-second segments from each patient were labelled according to the patient's overall classification. During inference, the model's output probabilities for these 2-second segments were averaged to yield a final prediction for each patient.

Several evaluation metrics were employed, including Accuracy, Sensitivity, Specificity, Precision, Receiver Operating Characteristic - Area Under the Curve (ROC-AUC), and F1-score. These metrics were calculated for each fold and then averaged across all five folds using a macro-average approach. ROC curves were generated by calculating the True Positive Rate (TPR) and False Positive Rate (FPR) based on the probabilities and true labels for each of the 405 patients, producing a micro-average.

Each metric provided insight into various performance aspects, particularly in imbalanced datasets like PHY16, where class distribution is skewed. Accuracy, the simplest metric, represents the proportion of total correctly classified instances and is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.14)$$

where TP (True Positives) and TN (True Negatives) refer to correctly predicted instances, while FP (False Positives) and FN (False Negatives) represent incorrect predictions. Although accuracy is easy to interpret, it can be misleading in imbalanced datasets, as it may overemphasize the performance of the majority class.

Recall, also known as Sensitivity or True Positive Rate (TPR), measures the proportion of actual positives that are correctly identified, calculated as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.15}$$

Recall is crucial in applications such as medical diagnostics, where missing a positive instance has severe consequences. However, Recall can be misleading if not considered alongside Precision, as it does not account for the number of false positives.

Specificity, or True Negative Rate (TNR), measures the proportion of correctly identified negatives, and is computed as:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4.16}$$

Specificity is important in minimizing false positives, particularly in applications where accurate identification of the negative class is critical, such as reducing false alarms in medical diagnostics. However, in imbalanced datasets, it may not be sufficient without considering performance on the minority class.

Precision measures the proportion of positive predictions that are actually correct, calculated as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.17}$$

This metric is particularly relevant in scenarios where false positives carry a high cost, such as heart abnormality detection, where unnecessary follow up tests could be costly or invasive, though it should be considered alongside Recall for a balanced perspective, as it does not account false negatives.

The ROC-AUC measures the model's ability to distinguish between positive and negative classes at various thresholds. The ROC curve plots TPR against FPR for different thresholds values between 0 and 1. In a binary classification problem, a sample is classified as belonging to the positive class if the predicted probability exceeds a certain threshold; otherwise, it is classified as negative. Therefore, the ROC curve illustrates the trade-off between TPR and FPR across these thresholds. The AUC

quantifies the model's overall discriminative ability, with values ranging from 0 to 1. The FPR is calculated as:

$$\text{FPR} = \frac{FP}{FP + TN} \tag{4.18}$$

A higher AUC indicates better model performance, but it may not provide granular insights into minority class performance in imbalanced datasets.

The F1-score is the harmonic mean of Precision and Recall, calculated as:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.19}$$

This metric is valuable in cases where balancing Precision and Recall is crucial, such as in detecting heart abnormalities. However, it may be less intuitive than individual metrics like Precision or Recall.

In summary, evaluating models on imbalanced datasets requires considering multiple metrics to gain a complete understanding of performance. Sensitivity, Precision, F1-score, and ROC-AUC offer more nuanced insights compared to simple Accuracy, ensuring that both minority and majority classes are properly accounted for in the evaluation.

4.5. Explainable Artificial Intelligence (XAI) methods

4.5.1 Gradient Class Activation Map (Grad-CAM)

Grad-CAM (Gradient-weighted Class Activation Mapping) is a technique that generates visual explanations for the decisions made by CNNs. By producing a heatmap, Grad-CAM highlights the important regions of an input image or signal that contribute to the model's prediction. This makes it particularly useful for interpreting deep learning models in tasks such as image or signal classification. The main goal of Grad-CAM is to enhance transparency by visualizing which areas of the input are significant for a specific decision. This is accomplished by combining gradient information flow from the final convolutional layer of the network with the feature maps from that layer.

To begin, when an input image or signal is processed by a CNN, it passes through multiple convolutional and pooling layers, eventually generating feature maps in the last convolutional layer. Let A^k represent the k-th feature map in this layer. The class score y^c for a particular class c is calculated based on these feature maps, typically obtained

from the fully connected layers following the final convolutional layer. To assess the importance of each feature map A^k for the class score y^c, the gradient of y^c with respect to A^k is computed using backpropagation:

$$\frac{\partial y^c}{\partial A^k} \tag{4.20}$$

These gradients are then averaged over all spatial locations to obtain the neuron importance weights α_k^c:

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{4.21}$$

where Z is the number of pixels in the feature map A^k, and A_{ij}^k represents the value at the (i,j) −th position of the k-th feature map.

The importance weights α_k^c are then used to compute a weighted combination of forward feature maps, resulting in a coarse localization map $L_{\text{Grad-CAM}}^c$:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{4.22}$$

The ReLU function ensures that only the features with a positive contribution on the class score are considered, setting negative values to zero. The resulting localization map $L_{\text{Grad-CAM}}^c$ is then up sampled to match the size of the input image, producing a heatmap that highlights the most relevant regions for class c in the input image [93].

Grad-CAM has several advantages, including improving interpretability by offering intuitive visual explanations for CNN decisions, which is particularly important in tasks requiring human comprehension, such as medical signal and image analysis. Additionally, Grad-CAM is model-agnostic, meaning it can be applied to any CNN-based architecture without needing modifications. By identifying class-discriminative regions, it helps pinpoint the areas contributing to a particular prediction, making it a valuable tool for debugging and improving model performance.

In this study, Grad-CAM was applied to both 1D and 2D CNN models to generate heatmaps for model interpretation. For the 1D CNN models, a new instance of the model was created for each case, extracting the output of the last convolutional layer and the final model predictions. A gradient tape was initiated to compute gradients, identifying the predicted class index and its associated probability. The gradients of the predicted class score with respect to the designated layer's output were then calculated. The global

average of these gradients was computed for each feature map, forming the basis for the Grad-CAM heatmap. The raw Grad-CAM heatmap was generated by multiplying the designated layer's output by the pooled gradients, and an additional dimension was added for visualization. A ReLU-like operation was applied to set negative values to zero. Two versions of the heatmap were tested: one with normalization by segment and one without it. The heatmap with normalization by segment adjusts each segment independently to its maximum value, improving contrast within segments, while the non-normalized version preserves the original relative intensities across the entire signal.

For the 2D-CNN models, the process was largely the same, except for the dimensionality of the feature maps. After computing the heatmap, it was collapsed along the first axis by summation, aggregating information to highlight significant areas in both PCG and ECG signals. As with the 1D models, a ReLU-like operation was applied to retain positive values. Two versions of the heatmap were again tested, one with normalization by dividing each value by the maximum value in the collapsed heatmap, and one without normalization by segment. The normalization step was introduced to explore whether standardizing values would improve model interpretability and facilitate inter-model comparisons.

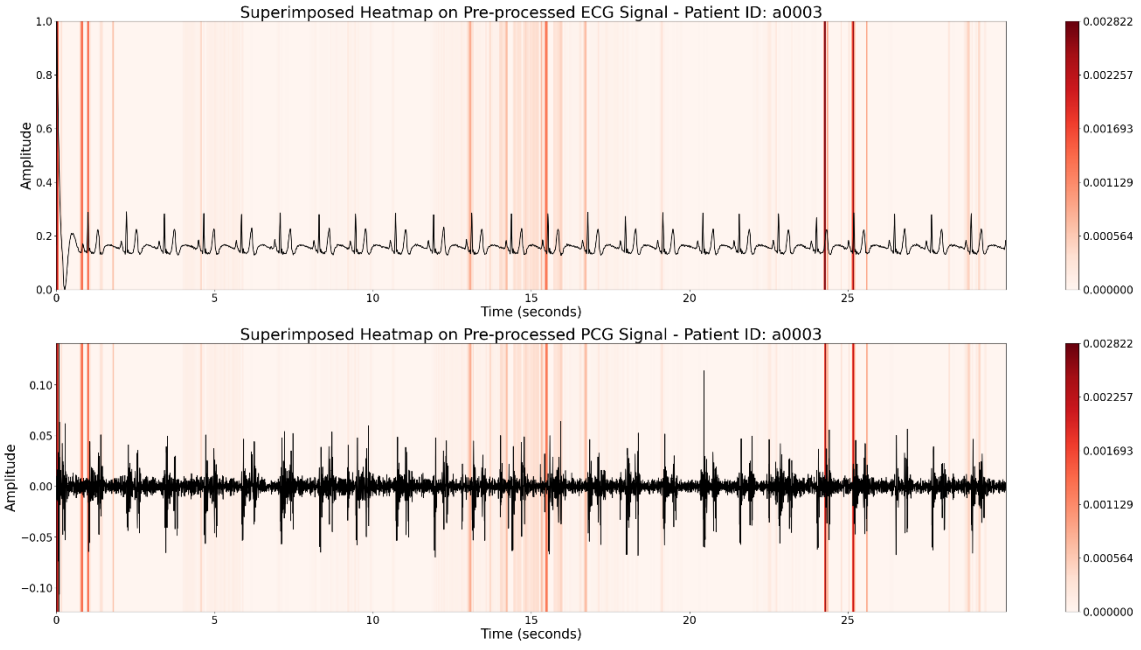


Figure 35 - Grad-CAM output for the Multimodal EF model for patient a0003.

Finally, the Grad-CAM heatmap was superimposed onto the input signals, highlighting the regions that contributed most to the model's decision. These post-processing steps improve the interpretability of the heatmap, offering insights into the

areas of the input signal that most influenced the model's decision-making. Figure 35 illustrates the final outcome for a specific patient's signal, showing the expected Grad-CAM visualization.

4.5.2. SHapley Additive exPlanations (SHAP) Gradient Explainer

Shapley values, originally developed in cooperative game theory, provide a method for fairly distributing the total gain or cost among participants based on their individual contributions [94]. SHAP (SHapley Additive exPlanations), proposed by Lundberg and Lee, adapts this concept to explain machine learning models by unifying several interpretability methods under a common framework, which can be viewed as approximations of Shapley values [95]. They extended this framework to deep learning models, leading to the development of techniques like Gradient SHAP.

Gradient SHAP combines ideas from SHAP and integrated gradients, which were introduced by Sundararajan et al. in 2017 [96]. Integrated gradients calculate feature attributions by integrating the gradients of the model's output with respect to the input features along a path from a baseline input to the actual input. In other words, they generate feature attributions by explaining the difference between a model's output for a sample and its output for a reference sample. The integrated gradient for a feature i is given by:

$$\mathrm{IntegratedGrad}_i(x, x') = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F\left(x' + \alpha(x - x')\right)}{\partial x_i} d\alpha \qquad (4.23)$$

where x is the input, x' is the baseline input, α ranges from 0 to 1 (representing the interpolation between the baseline and the input) and F is the model's output function. The integral computes the average gradient of the model's output with respect to the input, integrated along the path from x' to x.

Choosing the reference sample x' is crucial in this method. For instance, in image tasks, using an all-zero baseline may result in black pixels being ignored. One solution is to integrate gradients over multiple reference inputs, but this can be computationally expensive in terms of both time and memory. Gradient SHAP addresses this by incorporating an expectation over a distribution of reference inputs, sampled from a background dataset, rather than relying on a single reference [97]. This results in a combined expectation of gradients that sum to the difference between the expected model output and the current output. To further reduce computational complexity, a

method called Expected Gradients was introduced. This method approximates attributions by collapsing multiple integrals using sampling, allowing quick calculation across multiple references. Instead of directly integrating over the training distribution (which is intractable), the integral is reformulated as an expectation. For a model f and a feature i, with x as the sample, x' as a reference and D representing the distribution of the background dataset from which reference values x' are sampled, Expected Gradients is defined as:

$$\text{ExpectedGradients}_i(x) = E_{x' \sim D_{train}, \alpha \sim U(0,1)} \left[(x_i - x_i') \frac{\partial f\left(x' + \alpha(x - x')\right)}{\partial x_i} \right] \quad (4.24)$$

Therefore, Gradient SHAP works by selecting a background dataset, sampling x' from the training dataset and sampling α (a scalar that scales the difference between x and x') from uniform distribution $U(0,1)$. This approach represents the model's expected behaviour over a distribution of inputs. To finish, it computes the value inside the expectation for each sample and averages them, and thus approximates to the SHAP values.

In this study, Gradient SHAP was applied to all five neural network models: 2D CNN for PCG, 1D CNN and 2D CNN for ECG, and both the Multimodal Early Fusion (EF) and Multimodal Late Fusion (LF) models. The explainer was initialized with the respective model and the training dataset as the background dataset, which was appropriate because the training data represents the distribution the models were trained on. Gradient SHAP was then applied to all instances in the dataset, providing a thorough understanding of feature importance across different model architectures. In this context, each data point in the signal was treated as a feature.

For 2D models, the process also involved collapsing the SHAP heatmap along its first axis (the time axis) through summation, as done with Grad-CAM. Normalization of the SHAP values by segment was tested to explore whether this improved interpretability or not. Only the top 50 positive and 50 negative SHAP values were considered, focusing on the most and least influential points for model classification. This helped enhance the visual representation and provided a clearer explanation of how each data point contributed to the model's prediction. Figure 36 illustrates the expected result for a specific patient's signal, demonstrating how SHAP values assign importance (either positive, red points, or negative, blue points) to individual data points for each prediction, offering a detailed and interpretable view of model behaviour.

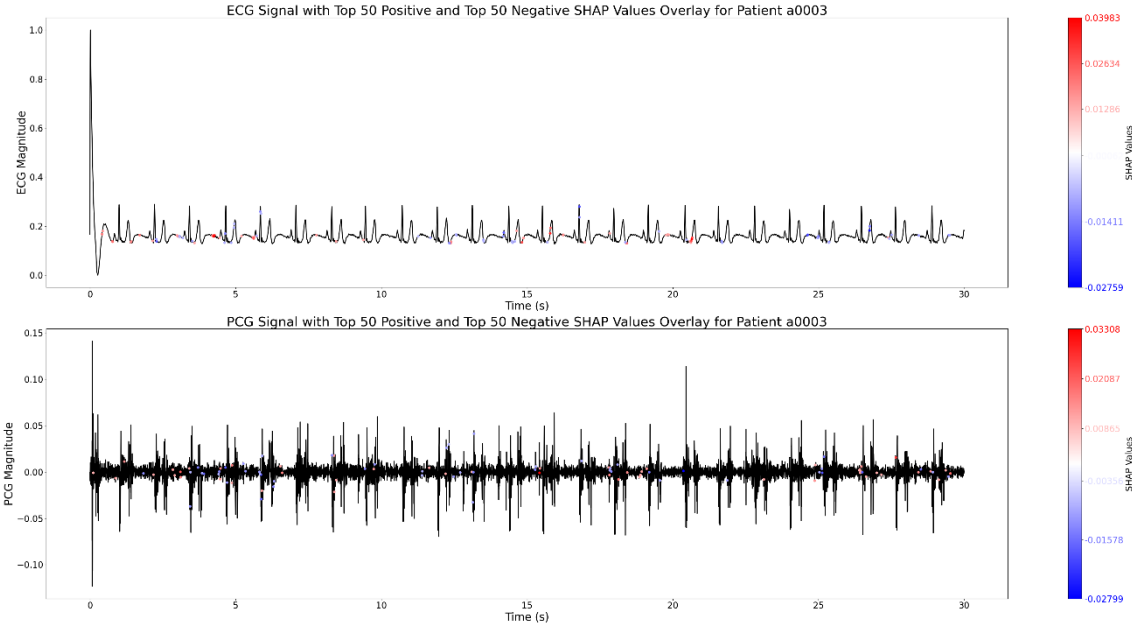


Figure 36 - SHAP output for the Multimodal EF model for patient a0003.

5. Results and Discussion

This chapter presents the results of the five model architectures (Section 4.1) and the visualizations generated using Grad-CAM and SHAP (Section 4.2). Additionally, a comparison of the models' performance and insights from the two XAI techniques is discussed.

5.1. Model´s evaluation and statistical test

The results presented in Figure 37, demonstrate that both the 2D ECG and multimodal EF models outperform the other architectures. The 2D ECG model achieved a ROC-AUC score of 0.829 ± 0.046 and an F1-score of 0.850 ± 0.046, while the EF model slightly surpassed it with a ROC-AUC score of 0.840 ± 0.029 and an identical F1-score of 0.850 ± 0.025. The close ROC-AUC scores of the 2D ECG and EF models highlight their similar ability to distinguish between positive and negative cases. In comparison, the LF model showed a lower ROC-AUC of 0.797 ± 0.027. The ROC-AUC metric is critical as it provides a single score that summarizes the model's discriminatory ability across different thresholds, offering a comprehensive evaluation of performance.

Interestingly, the 1D ECG and LF models achieved slightly higher F1-scores (0.850 ± 0.029 and 0.852 ± 0.032, respectively), indicating a better balance between precision and recall. This is especially important in medical applications where both high recall (to avoid missed cases) and high precision (to avoid unnecessary treatments) are essential. The F1-score, as a harmonic mean of precision and recall, becomes particularly useful in situations with imbalanced class distributions, as is often the case in medical diagnostics.

The 1D ECG model also stands out for its exceptional sensitivity/recall (0.920 ± 0.043), making it highly effective at identifying individuals with abnormalities, such as cardiac issues. High recall is crucial in healthcare, as it reduces the risk of missing critical cases, thereby improving patient outcomes by ensuring that fewer anomalies go undetected. Conversely, the 2D ECG and EF models achieved the highest specificity (0.650 ± 0.158 and 0.669 ± 0.133, respectively), indicating their effectiveness in correctly identifying individuals without medical conditions. High specificity helps reduce the number of false positives, which is vital for preventing unnecessary treatments or testing.

The 2D ECG and EF models also showed higher precision (0.861 ± 0.053 and 0.866 ± 0.043, respectively), meaning they were better at correctly identifying abnormal

signals when a positive prediction was made. This minimizes false alarms, ensuring that most of the abnormal signals flagged by these models are indeed abnormal, thereby reducing the psychological and financial burden on patients and the healthcare system. Additionally, both models demonstrated superior accuracy (0.790 ± 0.053 and 0.790 ± 0.026, respectively), although, given the class imbalance in the dataset, accuracy alone is not a reliable metric.

In terms of variability, the 2D ECG and 2D PCG models exhibited greater fluctuations in performance compared to the other models. This is evident from the larger standard deviations in specificity and sensitivity/recall metrics for the 2D ECG model (0.650 ± 0.158 and 0.847 ± 0.092 respectively) and in specificity and precision metrics for the 2D PCG model (0.583 ± 0.113 and 0.827 ± 0.042 respectively). This variability suggests that these models may have less stable performance, indicating challenges in handling specific cases. Also, regarding the multimodal models, LF appears to have more variability, and thus is less stable. In contrast, the EF model appears to be more stable when taking all metrics in consideration.

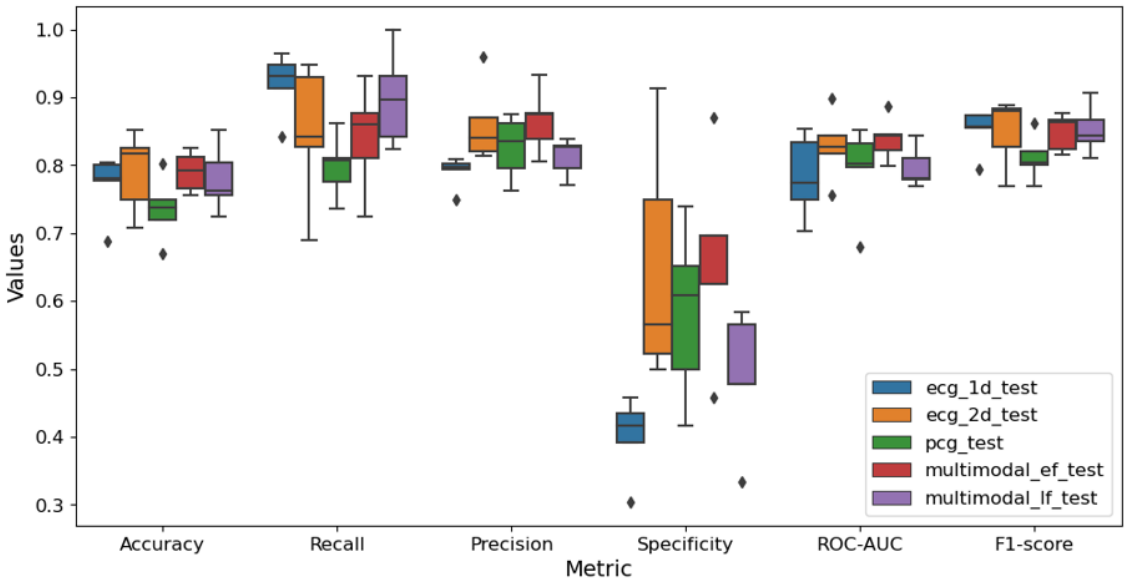


Figure 37 - Boxplots illustrating the metrics calculated for each of the five folds during cross-validation, across the five models.

These findings align with the ROC curves in Figure 38, where the 2D ECG and EF models demonstrate better overall performance by covering a larger area under the curve. The confusion matrices in Figure 39 further illustrates these differences. The EF model excels at predicting true negatives, accurately identifying normal cases, while the 1D ECG model is most effective at detecting true positives, highlighting its strength in

identifying abnormal cases. Additionally, the EF model produced the fewest false positives, reducing the risk of incorrectly classifying normal signals as abnormal, while the 1D ECG model had the fewest false negatives, minimizing missed abnormal cases, which is crucial for patient safety.

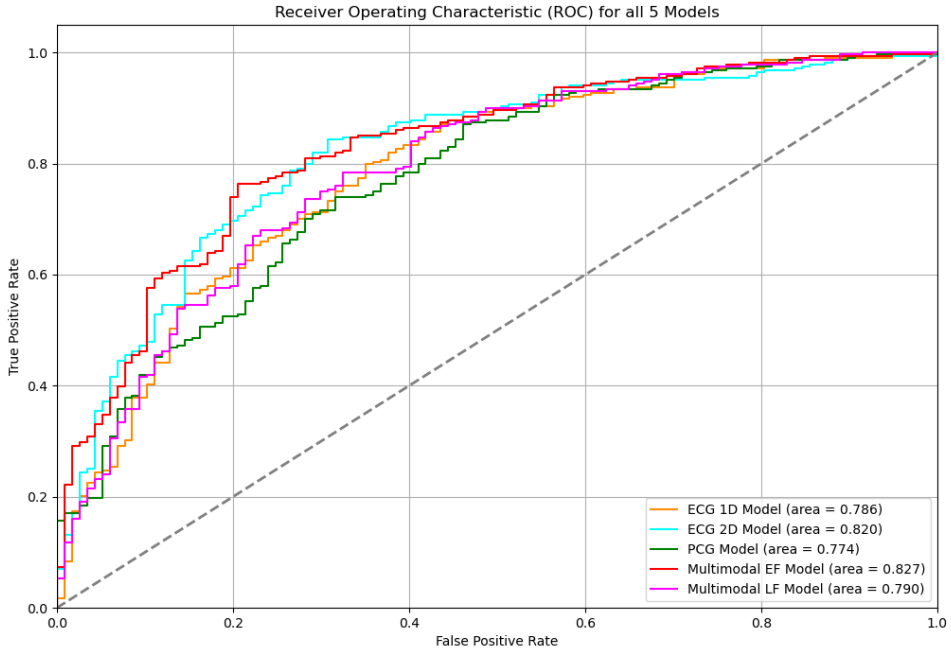


Figure 38 - ROC curves for all 5 models, calculated using the prediction probabilities of all patients during the 5-fold cross validation.

In contrast, the PCG model struggled the most with identifying true positives, resulting in the highest number of false negatives. Similarly, the 1D ECG model had the highest number of false positives, misclassifying more normal cases as abnormal. Interestingly, this result is somewhat unexpected, given that the PHY2016 dataset was originally designed for PCG analysis. However, the ECG analysis in this dataset surprisingly demonstrated better overall performance. These findings underscore the trade-offs and unique strengths of each model in the context of ECG and PCG signal classification.

To rigorously evaluate the differences between the five models, several statistical tests were conducted across various metrics. A paired t-test was employed to compare the performance of different model pairs across metrics like accuracy, recall, precision, specificity, ROC-AUC, and F1-score. It is a parametric test that determines whether there is a statistically significant difference between the means of two related samples. A p-

value less than 0.05 indicated statistically significant differences. Additionally, non-parametric tests such as the Wilcoxon Signed-Rank Test and McNemar's Test were performed, although they did not yield significant results.

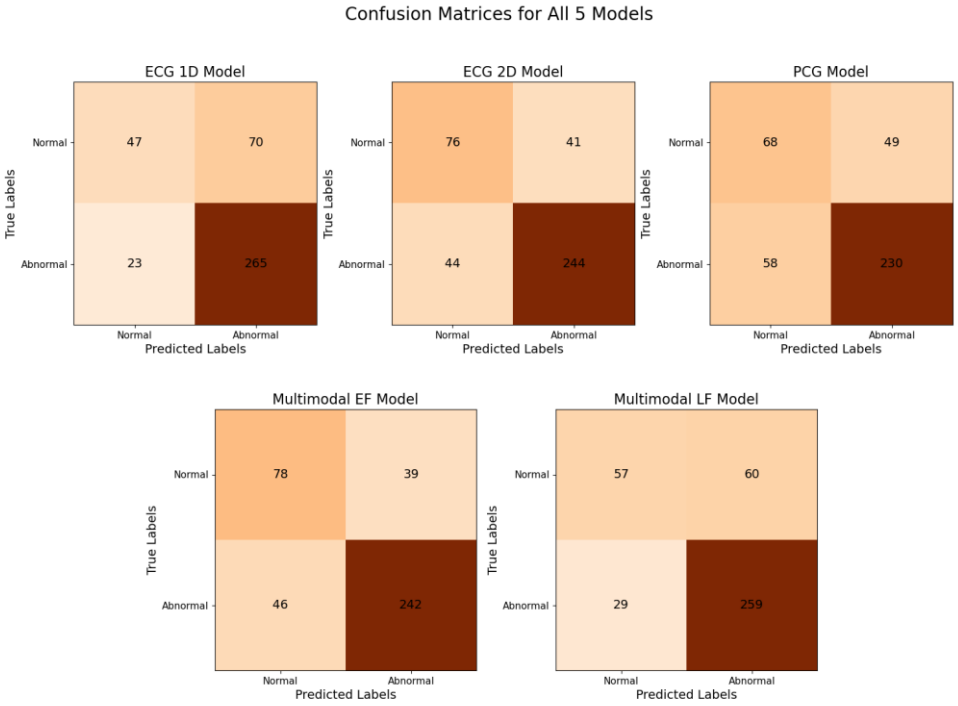


Figure 39 - Confusion matrices for all 5 models, summing the obtained confusion matrices from each fold of the 5-fold cross validation.

Key results emerged from these comparisons (Figure 40). No significant differences were found in accuracy between any pair of models. However, in recall, two significant pairs were identified: between the 1D ECG and PCG models ($p = 0.005$, favouring the 1D ECG model) and between the PCG and multimodal LF models ($p = 0.009$, favouring the LF model). In specificity, three significant pairs were observed: between the 1D ECG and 2D ECG models ($p = 0.028$, favouring the 2D ECG), between the 1D ECG and EF models ($p = 0.026$, favouring the EF model), and between the EF and LF models ($p = 0.038$, favouring the EF model). In terms of precision, four significant pairs were detected, with the multimodal EF model consistently performing better. Lastly, significant differences were found in F1-score ($p = 0.033$) and ROC-AUC ($p = 0.045$), both favouring the LF and EF models against PCG and multimodal LF models respectively.

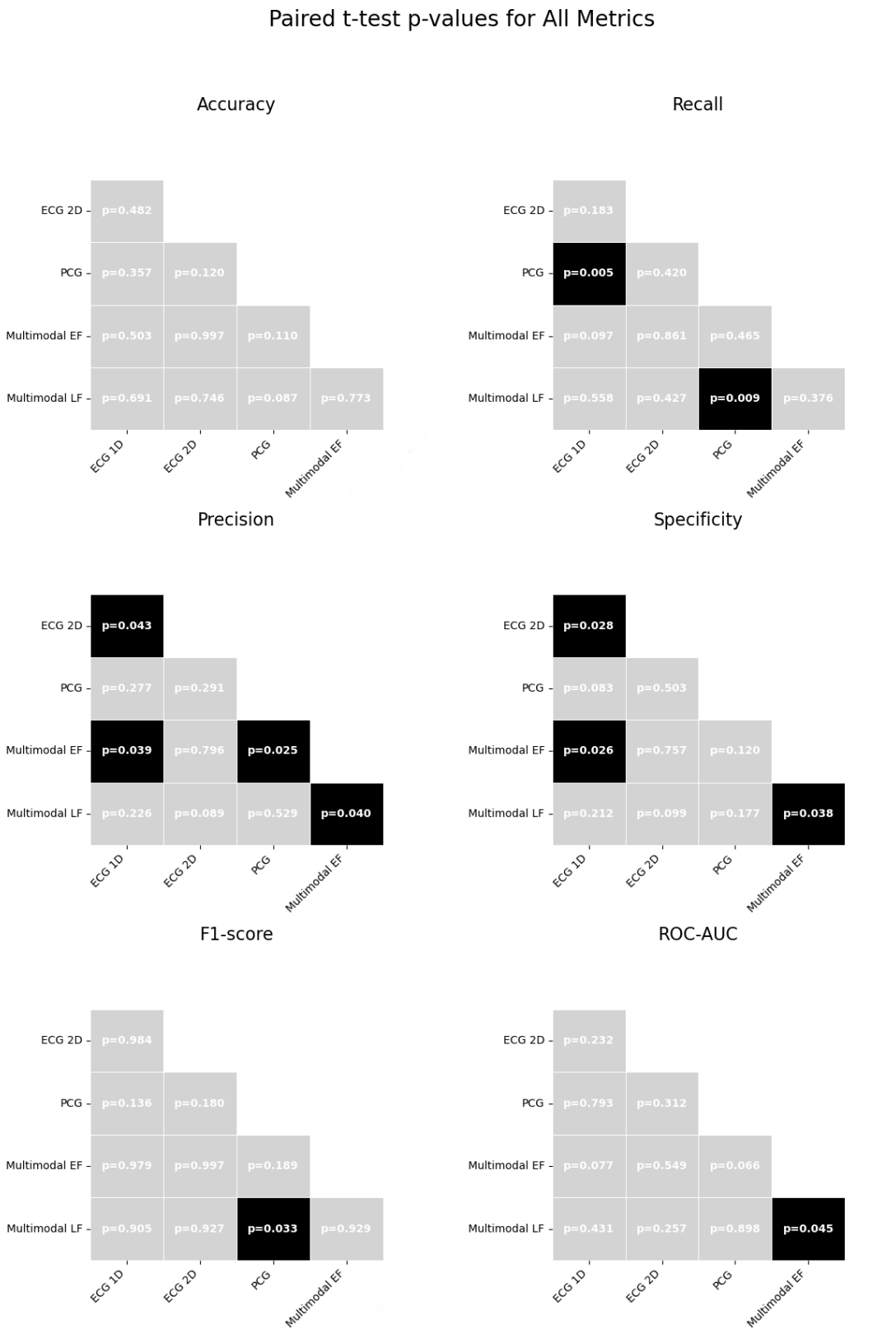


Figure 40 - Paired t-test results across 6 metrics (Accuracy, Recall, Precision, Specificity, F1-score and ROC-AUC) for all possible model pairs.

While the results are not as strong as state-of-the-art models tested on the PHY16 dataset, they include more extensive testing, such as statistical significance analysis. Additionally, evaluating the computational efficiency of the models on a low-power device, such as a multimodal stethoscope, would be necessary to verify their real-world applicability.

5.2. Application of Explainable Artificial Intelligence

In this section, Grad-CAM was applied to all five models using one example per classification scenario, aiming to cover a range of outcomes. These scenarios include correct classifications of both abnormal and normal conditions, as well as incorrect classifications of both, resulting in a total of four examples per model. The effect of segment-wise normalization on Grad-CAM visualizations was also examined, with the same patient segments used for comparison with non-normalized data. Specifically, Grad-CAM was computed for the following layers: the final 1D convolutional layer for the 1D ECG model, the last 2D convolutional layers for the 2D ECG and 2D PCG models, the final 2D convolutional layer after concatenation in the multimodal early fusion model, and the last convolutional layers of each modality (1D for ECG and 2D for PCG) before concatenation in the multimodal late fusion model. To ensure clearer interpretability, the analysis was constrained to signal lengths between 4 to 6 seconds.

Similarly, SHAP values were calculated for each model, using the same classification scenarios and patient examples as in the Grad-CAM analysis. This evaluation considered the contribution of each feature (i.e., each data point) to the model's predictions and compared the results with and without segment-wise normalization. The combination of Grad-CAM and SHAP results for each model allowed for a comprehensive assessment of feature importance and model decision-making processes, highlighting the respective strengths and limitations of both interpretability techniques.

5.2.1. Grad-CAM results

The Grad-CAM analysis of the 1D ECG model reveals a strong focus on the R peaks in both correct and incorrect classifications. This emphasis, shown in Figure 41, often leads to misclassifications, as not all heart abnormalities are linked to the R peaks. When the heatmap results are normalized by segments (Figure 42), this focus on the R peaks

becomes even more pronounced. In fact, some heatmap peaks that were absent in the non-normalized version appear in the normalized version, reinforcing the model's tendency to prioritize R peaks. In misclassified cases, the model not only highlights R peaks but also indiscriminately targets all peaks in the signal, including T and P waves.

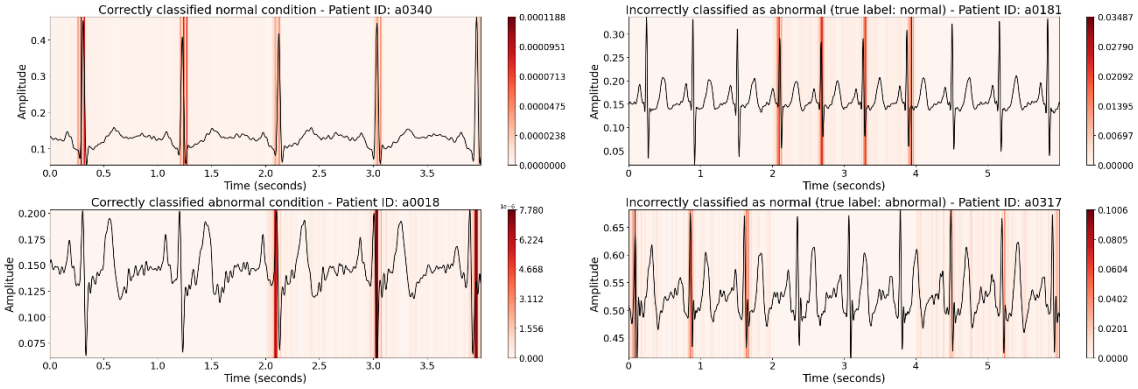


Figure 41 - Grad-CAM heatmaps for both normal and abnormal patients, with correct and incorrect classifications, generated using the 1D ECG model.

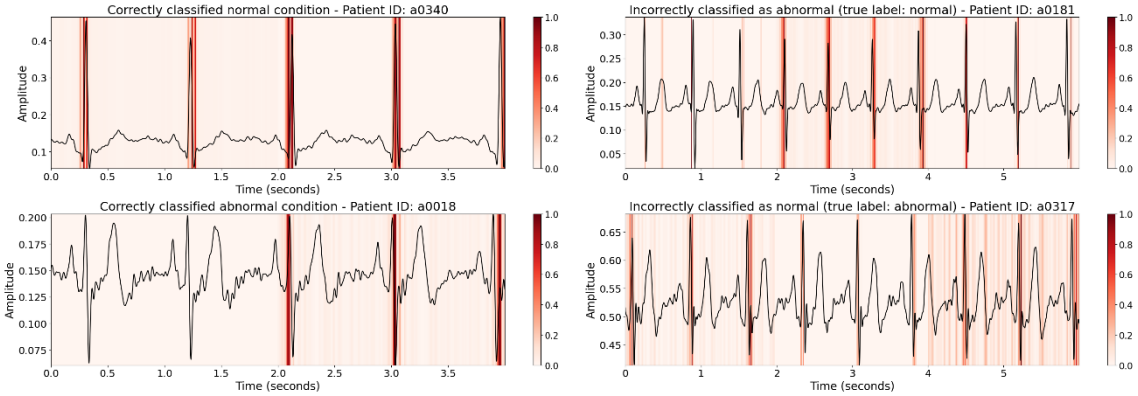


Figure 42 - Grad-CAM heatmaps for both normal and abnormal patients, with correct and incorrect classifications, normalized by segment, generated using the 1D ECG model.

In contrast, the 2D ECG model, as illustrated in Figure 43, shifts its attention from R peaks to the broader QRS complex, also considering other significant peaks like the T and P waves. Interestingly, when the model focuses on the P and T waves, it occasionally results in misclassifications. Normalizing the data by segment (Figure 44) helps clarify these misclassifications, as the model tends to highlight all peaks, including noise, when the signal quality or preprocessing introduces artifacts.

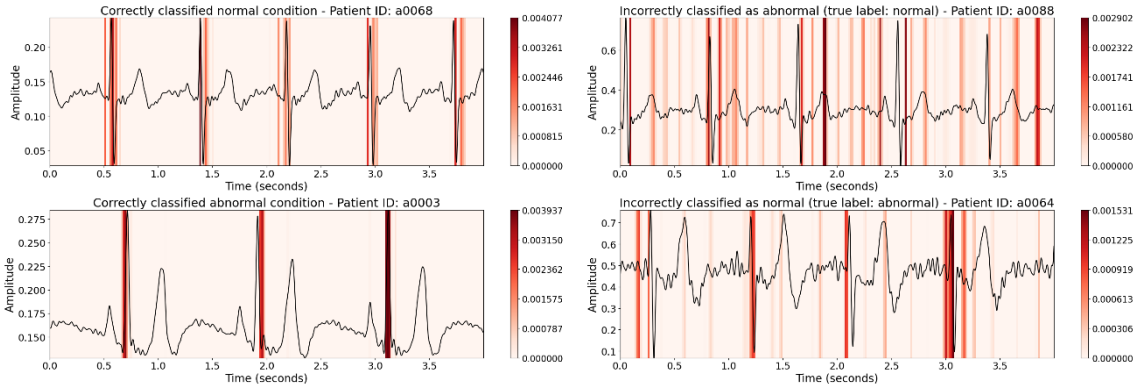


Figure 43 - Grad-CAM heatmaps for both normal and abnormal patients, with correct and incorrect classifications, generated using the 2D ECG model.

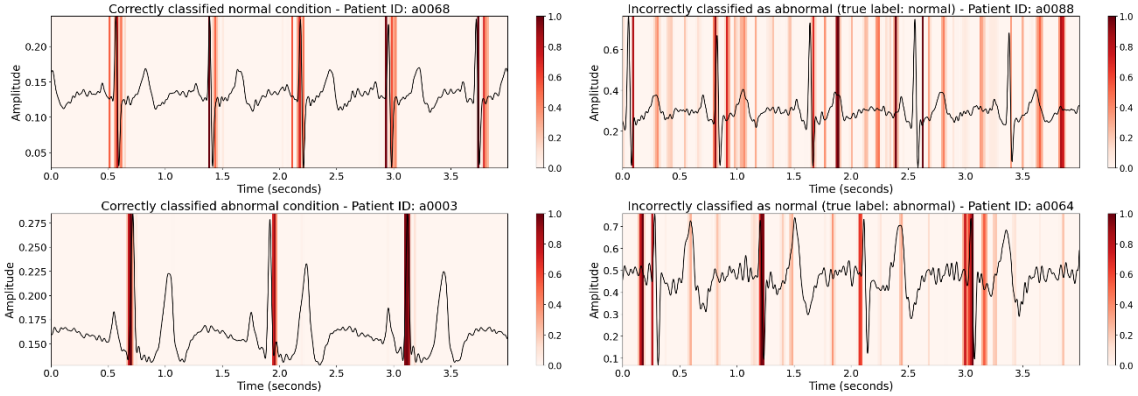


Figure 44 - Grad-CAM heatmaps for both normal and abnormal patients, with correct and incorrect classifications, normalized by segment, generated using the 2D ECG model.

For the 2D PCG model, the focus is primarily on the segments between the S1 and S2 heart sounds, with particular attention to the diastolic intervals (between S2 and S1). Misclassifications in this model often stem from noise in the sample data or low-amplitude peaks, as observed with the misclassified normal patient in Figure 45, where the model struggles with the S2 sound. This issue likely arises from challenges in signal acquisition and preprocessing. Interestingly, for correctly classified abnormal signals, particularly those associated with systolic murmurs, the model focuses more on the diastolic phase (the silent phase between S2 and S1) rather than the systolic phase (S1 to S2), likely because the diastolic interval appears to be longer than normal. Normalization (Figure 46) further reveals the model's tendency to highlight unusual peaks and abnormalities, particularly between heart sounds. In cases of misclassification, the model's difficulty in identifying relevant features between S1 and S2 highlights its challenge in dealing with noisy or ambiguous data.

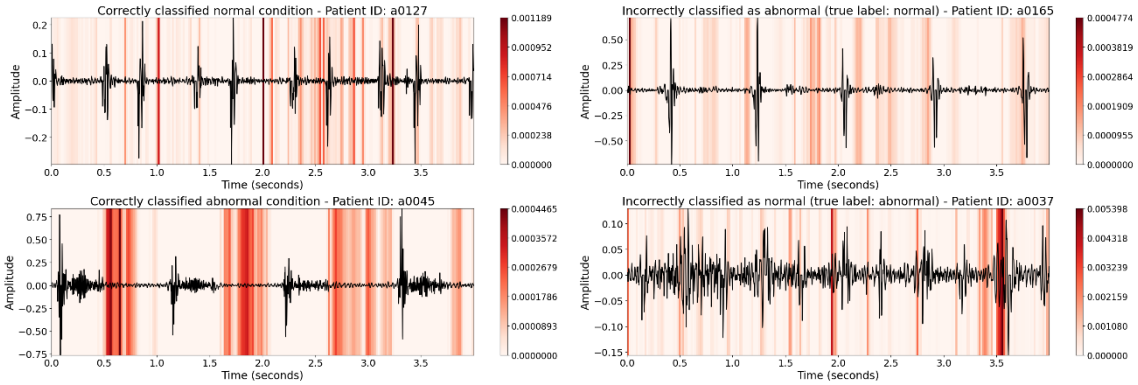


Figure 45 - Grad-CAM heatmaps for both normal and abnormal patients, with correct and incorrect classifications, generated using the 2D PCG model.

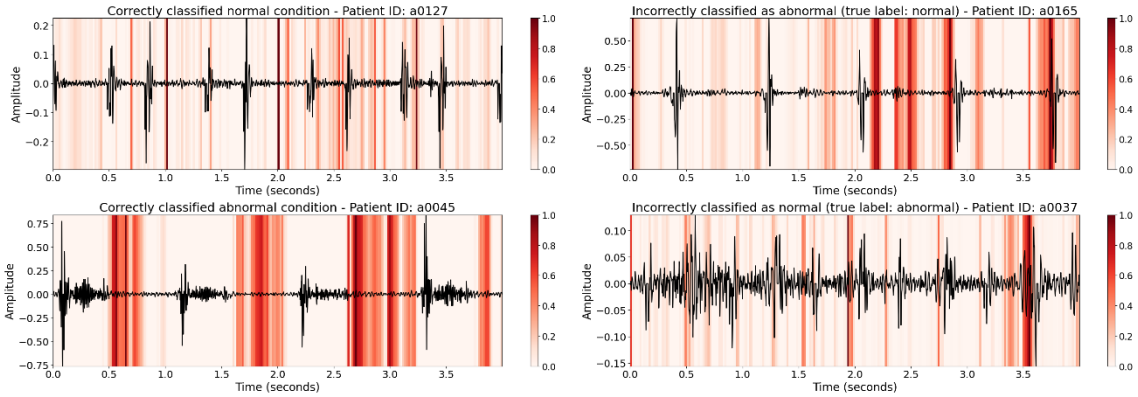


Figure 46 - Grad-CAM heatmaps for both normal and abnormal patients, with correct and incorrect classifications, normalized by segment, generated using the 2D PCG model.

Turning to the Multimodal EF model (Figure 47), it appears to correlate the S1 heart sound with the QRS complex (indicating the start of systole) and the S2 heart sound with the end of the T wave (marking the end of systole and the onset of diastole) to detect abnormalities within these associations. Since the Grad-CAM heatmap is generated from the final 2D convolutional layer, it reflects both ECG and PCG data simultaneously. Misclassifications occur when the model fails to accurately capture these key features. After normalization (Figure 48), reveals that the model also examines additional features, like the P wave, in its search for anomalies. The normalization process uncovers additional heatmap regions not visible in the non-normalized version, providing deeper insights into the model's decision-making process.

For the Multimodal LF model (Figure 49), each branch of the model focuses on its respective modality independently, as the Grad-CAM heatmaps are produced before the layers are concatenated. As a result, there is no clear integration of features from

both modalities. The ECG branch primarily focuses on the QRS complex, while the PCG branch targets the S1 and S2 heart sounds and abnormalities occurring during systole.

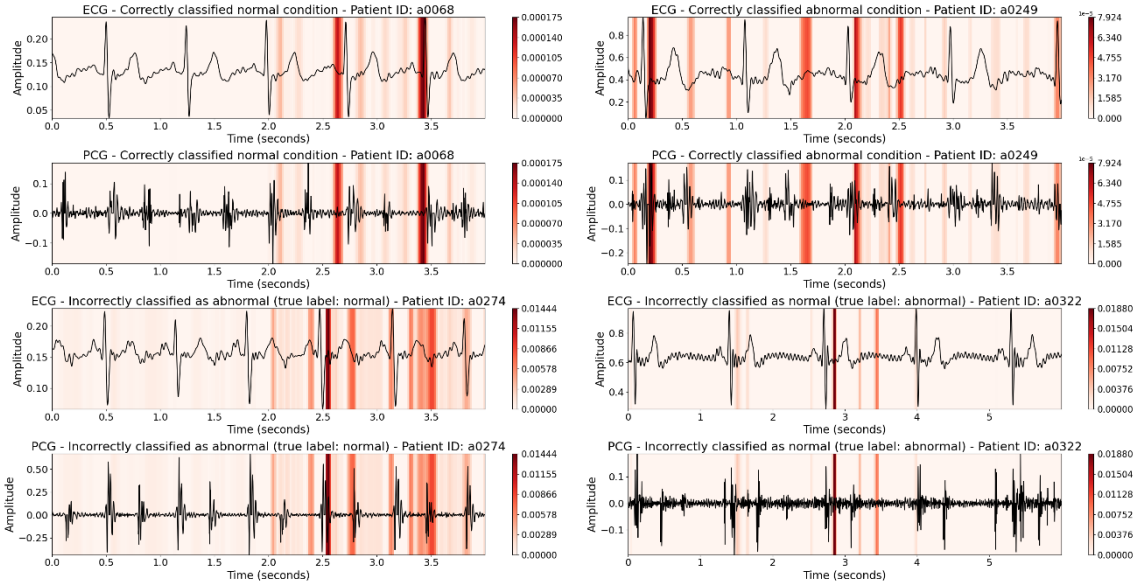


Figure 47 - Grad-CAM heatmaps for both normal and abnormal patients, with correct and incorrect classifications, generated using the Multimodal EF model. First signal represented is ECG, followed by PCG.

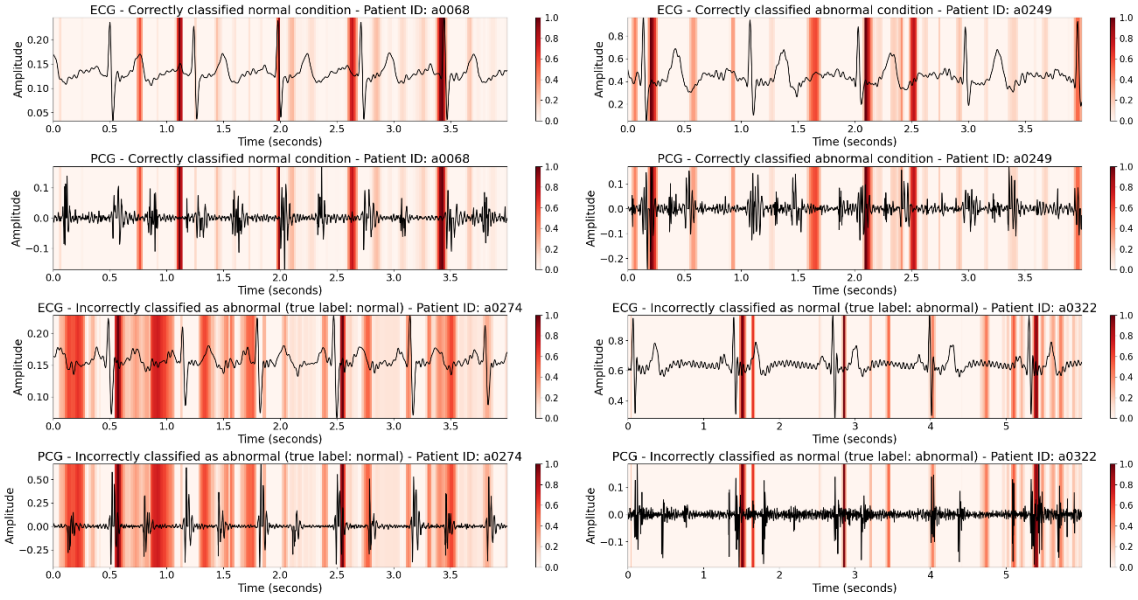


Figure 48 - Grad-CAM heatmaps for both normal and abnormal patients, with correct and incorrect classifications, normalized by segment, generated using the Multimodal EF model. First signal represented is ECG, followed by PCG.

Normalizing the results (Figure 50) reveals additional heatmap patterns that were not evident in the non-normalized version, offering a clearer view of the model's focus areas. Notably, the multimodal approach alters the focus of each modality compared to their standalone models. For example, the ECG branch, which previously concentrated on R

peaks in the 1D CNN model, now emphasizes the QRS complex, while the PCG branch shifts its attention from the intervals between S2 and S1 to the S1 and S2 sounds themselves, along with the intervals between them. This suggests a complementary interaction between the two modalities, with each one compensating for the other's reduced focus on certain features.

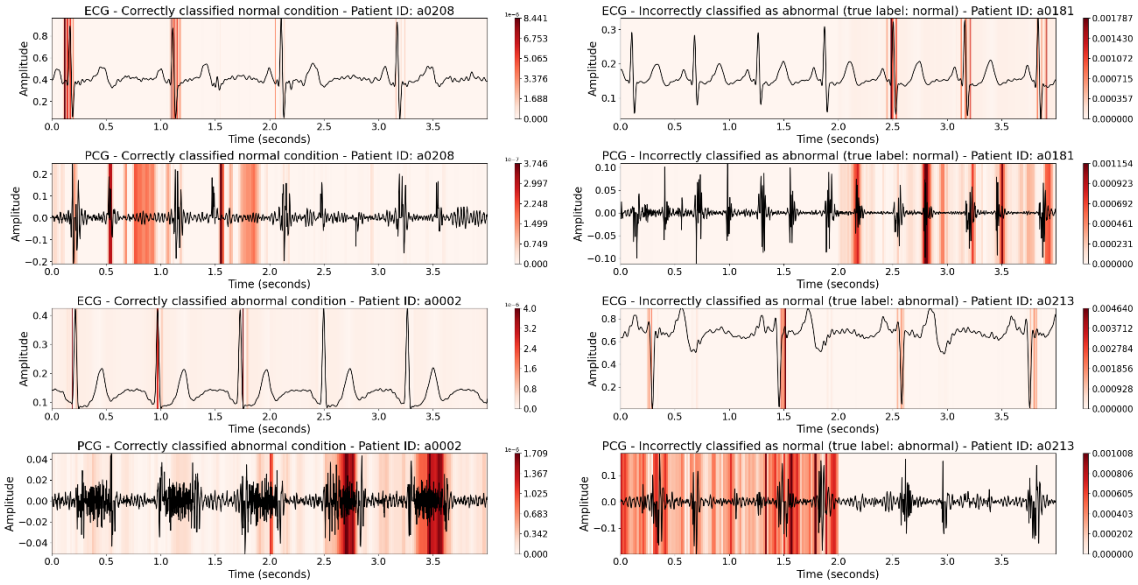


Figure 49 - Grad-CAM heatmaps for both normal and abnormal patients, with correct and incorrect classifications, generated using the Multimodal LF model. First signal represented is ECG, followed by PCG.

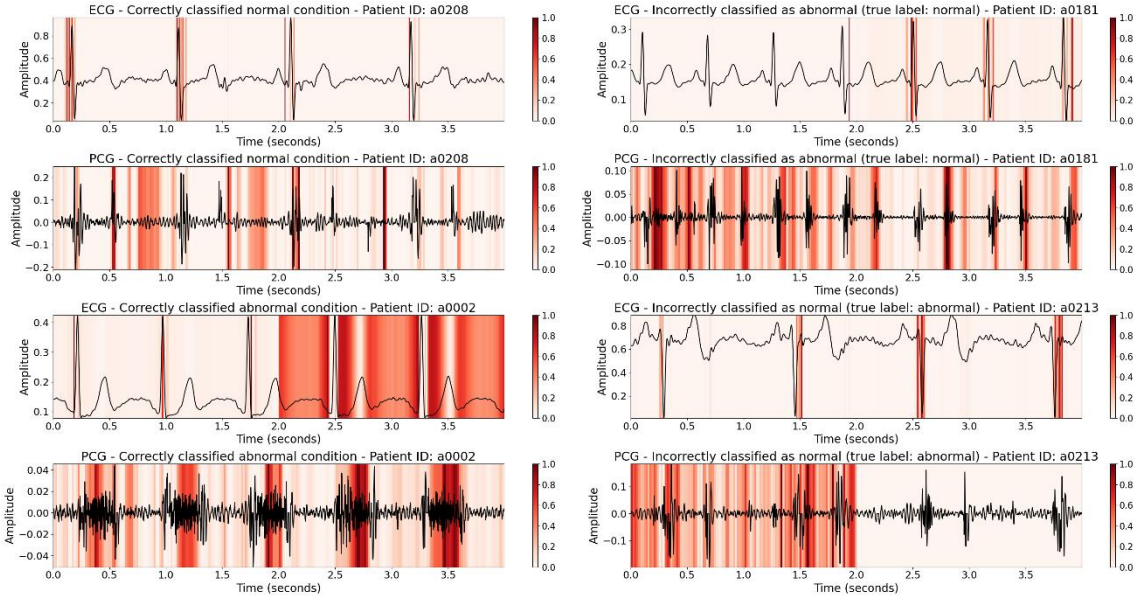


Figure 50 - Grad-CAM heatmaps for both normal and abnormal patients, with correct and incorrect classifications, normalized by segment, generated using the Multimodal LF model. First signal represented is ECG, followed by PCG.

5.2.2. Gradient SHAP results and comparison with Grad-CAM

The SHAP results for the 1D ECG model, as seen in Figure 51, show that in correct classifications, positive SHAP values are mainly concentrated on the Q and S points, with some appearing on the T wave in abnormal cases. Negative values tend to cluster around certain R-peaks, between Q and R points, and the ST segments. In misclassified segments, however, positive SHAP points are more commonly found on the R peaks and ST segments. When SHAP values are normalized by segment (Figure 52), there are notable shifts. Positive SHAP points in correct classifications now focus more on the QRS complex, particularly the Q and S points, and some extend in the T and P waves. Negative points are more concentrated on the ST segment and between T and P waves. In misclassified cases, only minor shifts are observed, with positive and negative points moving slightly within the QRS complex and the ST segment.

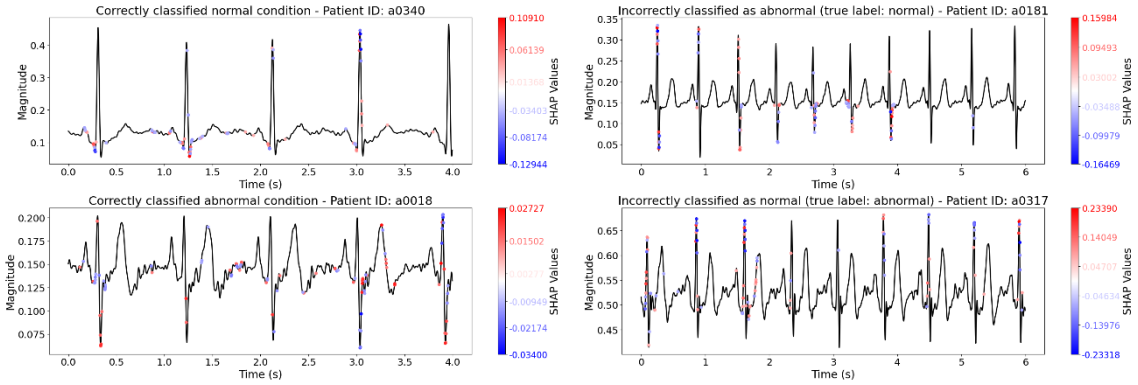


Figure 51 - Gradient SHAP, highlighting the top 50 positive and top 50 negative SHAP points for both normal and abnormal patients, with correct and incorrect classifications, generated using the 1D ECG model.

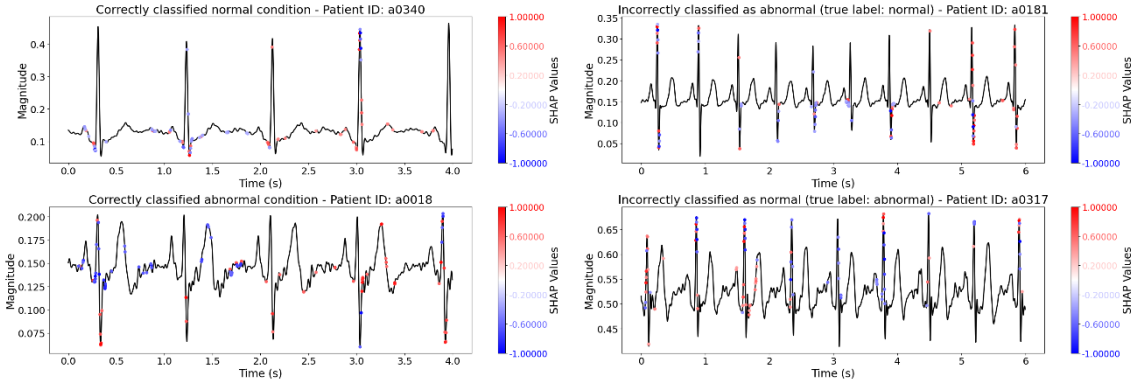


Figure 52 - Gradient SHAP, highlighting the top 50 positive and top 50 negative SHAP points for both normal and abnormal patients, with correct and incorrect classifications, normalized by segment, generated using the 1D ECG model.

When comparing these SHAP findings with the Grad-CAM results (Figures 41 and 42), several differences come out. Grad-CAM shows that the model predominantly

focuses on the R peaks in both correct and incorrect classifications, with normalization amplifying this effect and highlighting additional peaks. In contrast, SHAP provides a broader view, highlighting the QRS complex and emphasizing the T and P waves in correct classifications. Misclassifications in SHAP are more distributed across the QRS complex and ST segments, offering a different perspective on the model's behaviour compared to Grad-CAM's narrow focus on R peaks.

For the 2D ECG model (Figure 53), positive SHAP values in correct classifications are mainly concentrated on the Q and S points, with additional emphasis on the T and P waves in abnormal cases. Negative values tend to appear on the R peaks and ST segments. In misclassified segments, positive points are found more often on the R peaks and T waves, while negative values are seen on the ST segments. Normalizing the SHAP values (Figure 54) alters the distribution slightly.

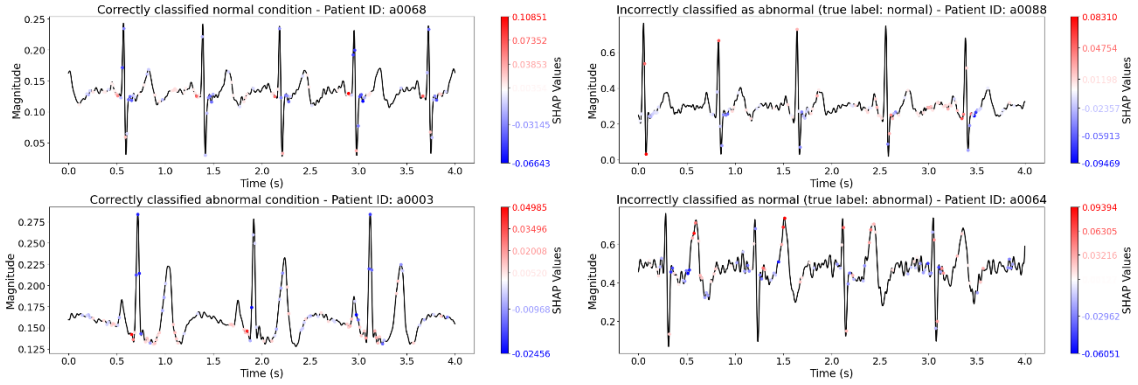


Figure 53 - Gradient SHAP, highlighting the top 50 positive and top 50 negative SHAP points for both normal and abnormal patients, with correct and incorrect classifications, generated using the 2D ECG model.

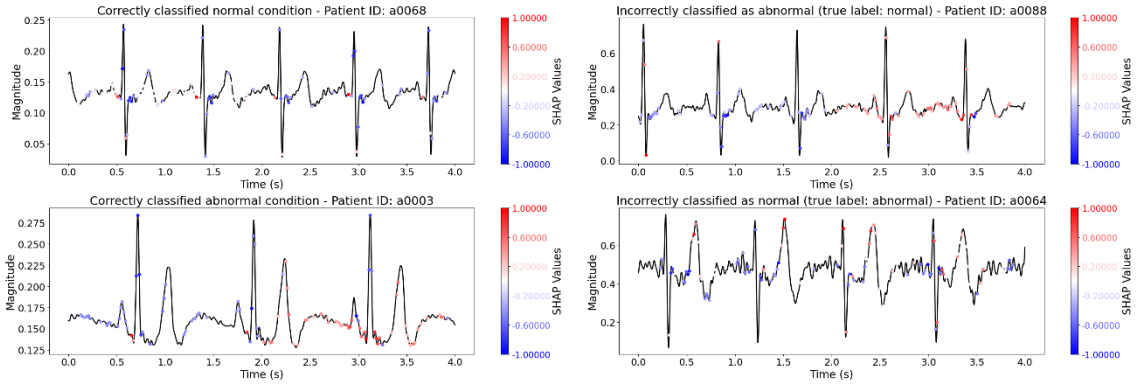


Figure 54 - Gradient SHAP, highlighting the top 50 positive and top 50 negative SHAP points for both normal and abnormal patients, with correct and incorrect classifications, normalized by segment, generated using the 2D ECG model.

In correct classifications, the positive SHAP points shift to focus more on the Q points and at the abnormal signal's T and P waves. Negative points remain in similar locations,

with a broader spread in abnormal cases. In incorrect classifications, there are no significant changes, except for more intense colouring due to normalization.

When comparing SHAP with Grad-CAM (Figures 43 and 44), SHAP offers a more granular perspective, highlighting specific ECG features such as the Q and S points and ST segments, whereas Grad-CAM focuses on broader regions, like the QRS complex and the T and P waves. Normalization in both methods reveals different patterns, with SHAP providing finer detail on feature importance.

Turning to the 2D PCG model (Figure 55), SHAP results show that in correct normal classifications, positive values appear primarily in the diastolic phase (between the S2 and S1 sounds), with some points in the systolic phase (S1 to S2). Negative values are evenly distributed across the signal. In correctly classified abnormal signals, positive SHAP values concentrate on the murmur region (between S1 and S2).

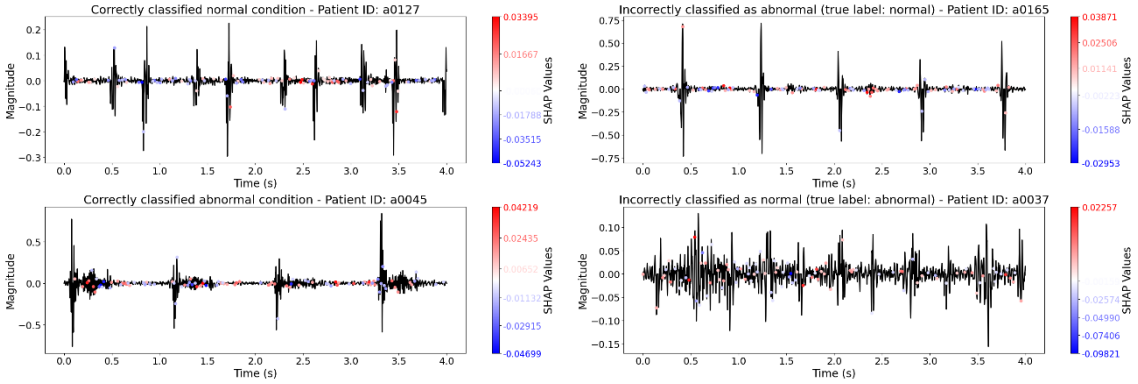


Figure 55 - Gradient SHAP, highlighting the top 50 positive and top 50 negative SHAP points for both normal and abnormal patients, with correct and incorrect classifications, generated using the 2D PCG model.

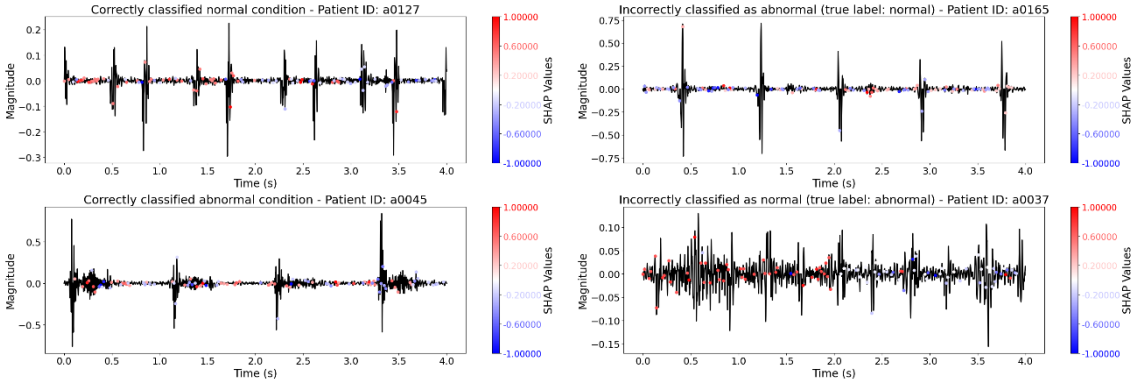


Figure 56 - Gradient SHAP, highlighting the top 50 positive and top 50 negative SHAP points for both normal and abnormal patients, with correct and incorrect classifications, normalized by segment, generated using the 2D PCG model.

In misclassified signals, both positive and negative points are scattered throughout the signal. After normalizing SHAP values (Figure 56), there is little change in point distribution, though positive points become more prominent on the fundamental sounds, particularly in normal cases, and the colours become more perceptible.

Comparing these SHAP results with Grad-CAM (Figures 45 and 46), both techniques highlight the importance of segments between S1 and S2 and S2 and S1. Grad-Cam tends to focus more between S2 and S1 and SHAP prioritizes the region between S1 and S2, which is particularly relevant since systolic murmurs are more common than diastolic murmurs [98]. This might suggest that SHAP is capturing more clinically relevant features in this context. However, SHAP provides a more detailed view of positive and negative contributions, especially in the murmur regions. Grad-CAM focuses more broadly on these intervals, while SHAP pinpoints specific regions of the signal that contribute to the classification' results.

For the Multimodal EF model (Figure 57), in correct normal classifications, SHAP highlights the P wave and ST segment in ECG and the S1 sound and systolic phase in PCG. Positive points also appear in the diastolic phase. In correct abnormal classifications, SHAP values focus on the QRS complex and between the T and P waves in ECG, while PCG shows emphasis on the murmur region. In misclassifications, positive SHAP points are concentrated on the T and P waves in ECG and the systolic phase in PCG. In the misclassified normal patient, negative Shap values focus more in the QRS complex for ECG and in the diastolic phase for PCG, while in the misclassified abnormal patient, positive SHAP values concentrate in the QRS complex of the ECG and in the beginning and end of the systolic phase in PCG. Normalization (Figure 58) causes some shifts, with positive SHAP values more evenly distributed in correct normal cases, and focusing primarily on the QRS complex in correct abnormal cases. Misclassified cases show minor shifts in SHAP points, particularly in the PR and QT intervals in ECG.

Comparing SHAP and Grad-CAM for the Multimodal EF model (Figures 47 and 48), SHAP provides a more detailed breakdown of feature importance across ECG and PCG signals, while Grad-CAM focuses on broader correlations between the S1 sound, QRS complex, and the S2 sound with the T wave. Normalization in Grad-CAM uncovers additional features such as the P wave, while SHAP continues to provide a finer level of detail in both modalities.

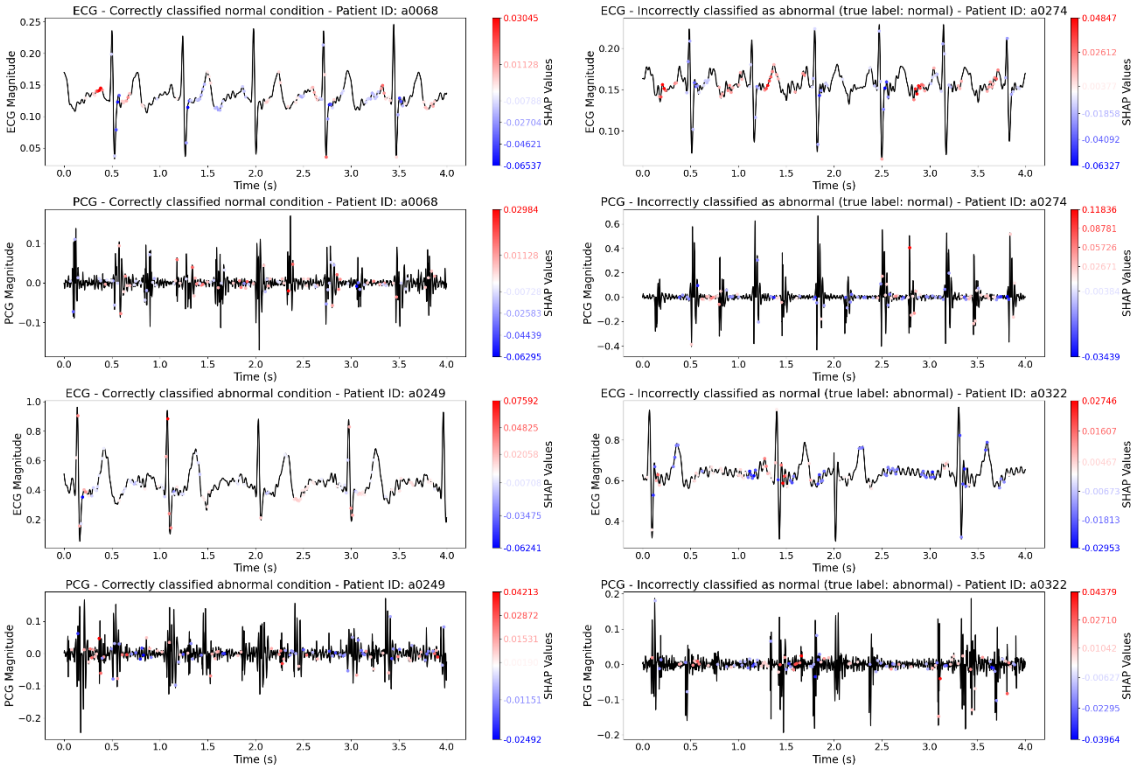


Figure 57 - Gradient SHAP, highlighting the top 50 positive and top 50 negative SHAP points for both normal and abnormal patients, with correct and incorrect classifications, generated using the Multimodal EF model.

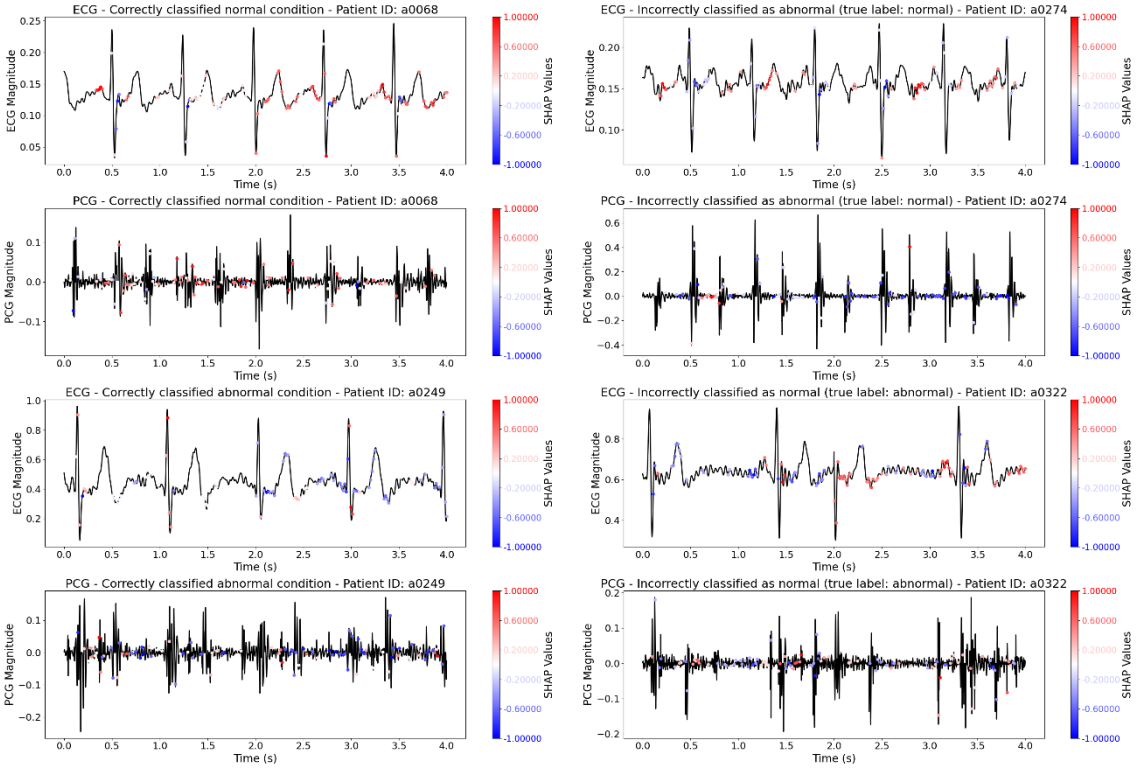


Figure 58 - Gradient SHAP, highlighting the top 50 positive and top 50 negative SHAP points for both normal and abnormal patients, with correct and incorrect classifications, normalized by segment, generated using the Multimodal EF model.

Finally, in the Multimodal LF model (Figure 59), SHAP points in correct classifications alternate within the QRS complex and between the P and T waves in ECG, while in PCG, positive values appear between the S1 and S2 sounds for normal cases and over the murmur for abnormal cases. In incorrect normal classifications, SHAP points shift within the QRS and ST segments, and within the PR and QT intervals in misclassified abnormal classification, scattering across the PCG in both misclassified cases. Normalization (Figure 60) makes the SHAP points more visible without significantly altering their locations.

In comparison with Grad-CAM (Figures 49 and 50), SHAP reveals a more detailed distribution of feature importance, while Grad-CAM shows that each branch of the Multimodal LF model independently focuses on its respective modality. SHAP demonstrates how features are prioritized differently across the ECG and PCG signals, whereas Grad-CAM provides a more unified focus on the QRS complex and fundamental heart sounds. Normalization in both methods helps clarify the model's behaviour, with SHAP offering a finer breakdown of feature importance and Grad-CAM highlighting broader trends.

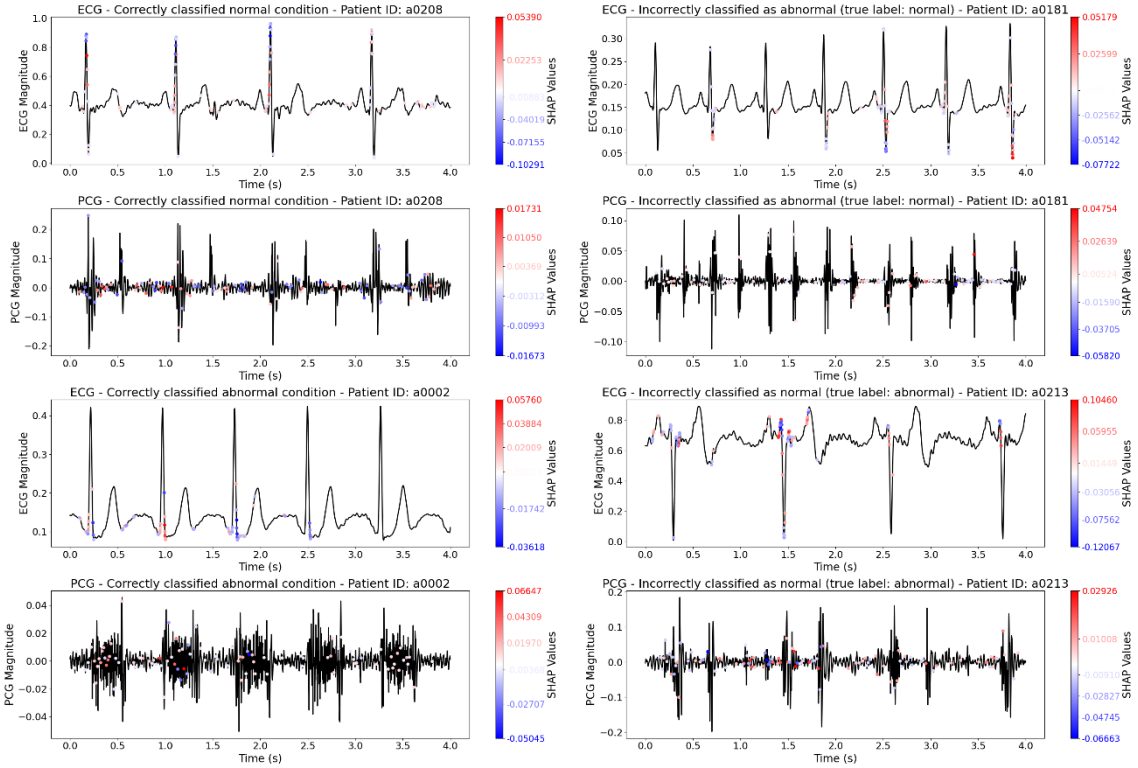


Figure 59 - Gradient SHAP, highlighting the top 50 positive and top 50 negative SHAP points for both normal and abnormal patients, with correct and incorrect classifications, generated using the Multimodal LF model.

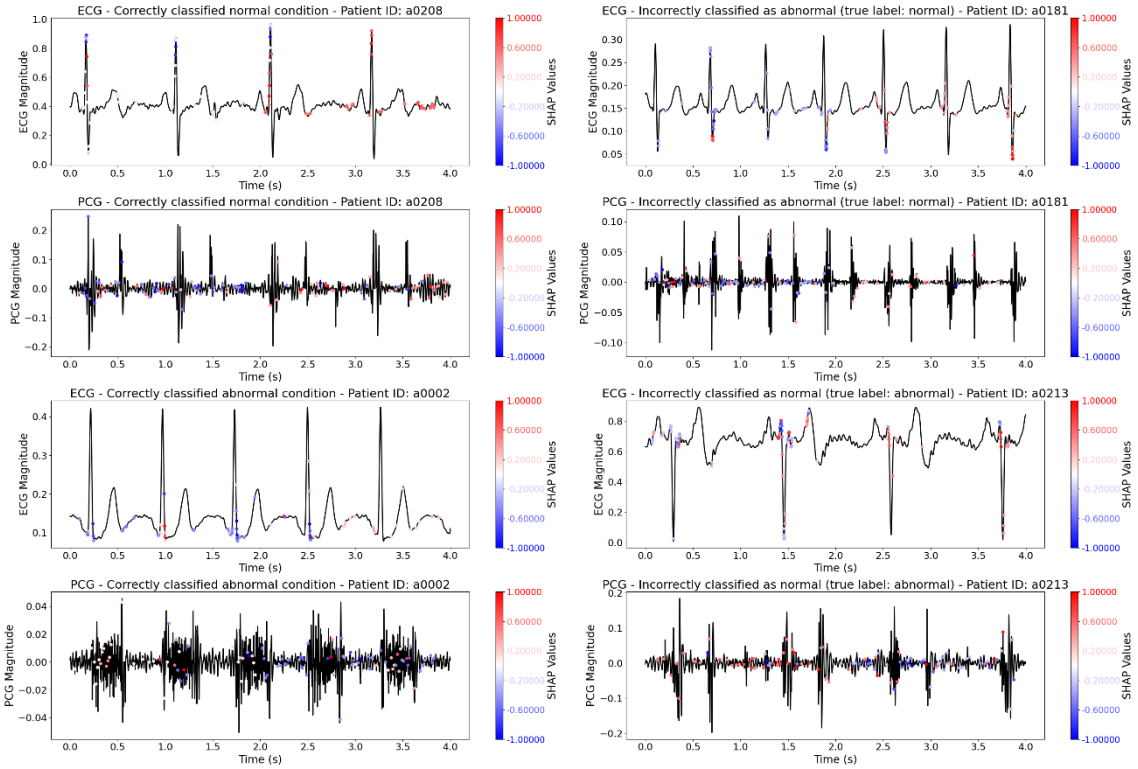


Figure 60 - Gradient SHAP, highlighting the top 50 positive and top 50 negative SHAP points for both normal and abnormal patients, with correct and incorrect classifications, normalized by segment, generated using the Multimodal LF model.

In conclusion, both SHAP and Grad-CAM reveal key features in ECG and PCG signals, such as the QRS complex, P and T waves, and the S1 and S2 heart sounds. SHAP provides a more granular view of feature importance, while Grad-CAM offers a broader focus. Together, these techniques offer complementary insights into the models' decision-making processes, revealing both their strengths and limitations in handling noisy or ambiguous data. Normalizing the data enhances both methods, providing clearer interpretations and deeper understanding of the regions deemed important by the models for specific classifications.

6. Conclusion

This thesis involved the design and evaluation of five model architectures, each incorporating an XAI component. In this concluding chapter, the key findings of the study are summarized, and potential future research directions are outlined.

6.1. Main outcomes

This study demonstrates that multimodal models combining ECG and PCG signals exhibit greater stability and achieve a more favourable balance between precision and recall compared to single-modality cardiovascular disease classifiers. Among the two multimodal architectures explored, early fusion outperforms late fusion, showing statistically significant improvements in specificity, precision, and ROC-AUC metrics. Additionally, the 2D ECG model outperforms the 1D ECG model in terms of discriminative power, indicating that preprocessing ECG signals into higher dimensions enhances model performance, with significant gains in precision and specificity.

Regarding the two explainability methods, SHAP and Grad-CAM, the analysis reveals that the models effectively focus on critical features of both ECG (QRS complex, P, and T waves) and PCG (S1, S2, and their intervals), essential for distinguishing normal from abnormal signals. While SHAP provides a more detailed breakdown of feature importance across different signal components, Grad-CAM offers a broader visualization of the models' focus areas. Together, these techniques provide a comprehensive understanding of each model's strengths and limitations in identifying key features amid noise or ambiguous data. Normalization of the data further enhances the clarity of visual interpretations from both methods, helping to highlight the most relevant regions influencing the model's predictions.

Overall, these findings highlight the potential of integrating ECG and PCG signals within multimodal deep learning frameworks to improve diagnostic accuracy and reliability. The superior performance of early fusion models, along with the complementary insights offered by SHAP and Grad-CAM, underlines the value of combining multiple interpretability methods in the development of more robust cardiovascular diagnostic tools.

6.2. Future Work

Future research should prioritize several key areas to further improve the performance and applicability of multimodal ECG and PCG models, as well as enhance the interpretability provided by XAI methods. A primary objective, already in progress, is to expand the diversity and volume of synchronized ECG and PCG datasets by incorporating data from a broader range of patient demographics and medical conditions. In addition to this, further validation on an external dataset is required to assess the robustness of the models and improve their generalizability. Additionally, employing data augmentation techniques, such as those using generative deep learning models [99], could simulate a wider variety of physiological variations, thereby further boosting model robustness.

Another promising direction is the development of real-time diagnostic tools that leverage multimodal models for clinical use. Optimizing these models for speed and efficiency will be crucial for their application in real-time patient monitoring systems and low-power devices like multimodal stethoscopes. These advancements could also support personalized medicine, where future models could be adapted to individual patient profiles by incorporating personalized risk factors and medical histories, leading to more accurate and tailored diagnostics.

In terms of explainability, future work should explore integrating additional XAI techniques, such as LIME (which may require adaptation for biological time series data), alongside SHAP and Grad-CAM. This hybrid approach could provide more comprehensive explanations of model decisions by combining insights from multiple techniques into a unified interpretability framework. As models evolve to classify multiple cardiovascular conditions, XAI methods should also be adapted to explain how the model distinguishes between these conditions, especially as more extensive datasets become available.

Validating the outputs of XAI methods is essential to ensure that the explanations generated are reliable and clinically meaningful. Approaches such as the Jaccard index with shapelets, as well as visual validation by specialized physicians [100], are promising for validating these outputs. Additionally, incorporating Uncertainty Quantification (UQ) will help assess the confidence of the model's explanations, which is particularly valuable in clinical settings. Techniques like the Spatial Uncertainty Estimator [101] have demonstrated the importance of UQ in enhancing the reliability of XAI. The combined

use of UQ and explanation quality assessment can improve model insights and contribute to more informed decision-making and better patient outcomes.

Finally, creating a feedback loop system where clinicians and physicians can review and provide feedback on the model's explanations would foster continuous learning and model refinement. This interaction would enhance the model's accuracy and reliability, making it more effective in real-world clinical applications.

By addressing these areas, future research can build upon this study to develop more accurate, robust, and interpretable multimodal diagnostic tools, making them suitable for widespread clinical use.

References

[1] 'Cardiovascular diseases (CVDs)'. Accessed: Aug. 12, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[2] Q. Counseller and Y. Aboelkassem, 'Recent technologies in cardiac imaging', *Front. Med. Technol.*, vol. 4, p. 984492, Jan. 2023, doi: 10.3389/fmedt.2022.984492.
[3] M. C. Kusko and K. Maselli, 'Introduction to Cardiac Auscultation', in *Learning Cardiac Auscultation: From Essentials to Expert Clinical Interpretation*, A. J. Taylor, Ed., London: Springer, 2015, pp. 3–14. doi: 10.1007/978-1-4471-6738-9_1.
[4] A. Dupre, S. Vieau, and P. A. Iaizzo, 'Basic ECG Theory, 12-Lead Recordings and Their Interpretation', in *Handbook of Cardiac Anatomy, Physiology, and Devices*, P. A. Iaizzo, Ed., Totowa, NJ: Humana Press, 2009, pp. 257–269. doi: 10.1007/978-1-60327-372-5_17.
[5] X. Bao, Y. Deng, N. Gall, and E. Kamavuako, 'Analysis of ECG and PCG Time Delay around Auscultation Sites', Jan. 2020, pp. 206–213. doi: 10.5220/0008942602060213.
[6] W. Chen, Q. Sun, X. Chen, G. Xie, H. Wu, and C. Xu, 'Deep Learning Methods for Heart Sounds Classification: A Systematic Review', *Entropy*, vol. 23, no. 6, p. 667, May 2021, doi: 10.3390/e23060667.
[7] N. Musa *et al.*, 'A systematic review and Meta-data analysis on the applications of Deep Learning in Electrocardiogram', *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 7, pp. 9677–9750, 2023, doi: 10.1007/s12652-022-03868-z.
[8] R. Hettiarachchi *et al.*, '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)*, May 2021, pp. 1–5. doi: 10.1109/ISCAS51556.2021.9401093.
[9] H. Li *et al.*, 'Dual-Input Neural Network Integrating Feature Extraction and Deep Learning for Coronary Artery Disease Detection Using Electrocardiogram and Phonocardiogram', *IEEE Access*, vol. 7, pp. 146457–146469, 2019, doi: 10.1109/ACCESS.2019.2943197.
[10] S. Kundu, 'AI in medicine must be explainable', *Nat. Med.*, vol. 27, no. 8, pp. 1328–1328, Aug. 2021, doi: 10.1038/s41591-021-01461-z.
[11] M. Mersha, K. Lam, J. Wood, A. K. AlShami, and J. Kalita, 'Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction', *Neurocomputing*, vol. 599, p. 128111, Sep. 2024, doi: 10.1016/j.neucom.2024.128111.
[12] D. P. Zipes, J. Jalife, and W. G. Stevenson, Eds., *Cardiac electrophysiology: from cell to bedside*, Seventh edition. Philadelphia, PA: Elsevier, 2018.
[13] P. Vijayaraman *et al.*, 'Cardiac Conduction System Pacing: A Comprehensive Update', *JACC Clin. Electrophysiol.*, vol. 9, no. 11, pp. 2358–2387, Nov. 2023, doi: 10.1016/j.jacep.2023.06.005.
[14] I. Pinto, A. Fred, C. Rodrigues, and H. Plácido da Silva, *Electrophysiology of the Heart and the Electrocardiogram: Visual Depictions*. 2020.
[15] K. Wang, *Atlas of Electrocardiography*. Jaypee Brothers Medical Publishers (P) Ltd., 2013. doi: 10.5005/jp/books/11969.
[16] M. Gavaghan, 'Cardiac Anatomy and Physiology: A Review', *AORN J.*, vol. 67, no. 4, pp. 800–822, Apr. 1998, doi: 10.1016/S0001-2092(06)62644-6.
[17] S. Douedi and H. Douedi, 'P wave', in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2024. Accessed: Sep. 01, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK551635/
[18] M. Al-Akchar and M. S. Siddique, 'Long QT Syndrome', in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2024. Accessed: Sep. 01, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK441860/

[19] A. H. Kashou, H. Basit, and A. Malik, 'ST Segment', in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2024. Accessed: Sep. 01, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK459364/
[20] Q. Yao, R. Wang, X. Fan, J. Liu, and Y. Li, 'Multi-class Arrhythmia detection from 12-lead varied-length ECG using Attention-based Time-Incremental Convolutional Neural Network', *Inf. Fusion*, vol. 53, pp. 174–182, Jan. 2020, doi: 10.1016/j.inffus.2019.06.024.
[21] P. P. MD, R. M. MD, R. K. MD, and P. A. N. MD, *Podrid's Real-World ECGs: Volume 1, The Basics: A Master's Approach to the Art and Practice of Clinical ECG Interpretation*. Cardiotext Publishing, 2012.
[22] S. Turnbull *et al.*, 'Accuracy of a Single-Lead ECG Device for Diagnosis of Cardiac Arrhythmias Compared Against Cardiac Electrophysiology Study', *Heart Lung Circ.*, Jul. 2024, doi: 10.1016/j.hlc.2024.05.008.
[23] L. Fabricius Ekenberg *et al.*, 'Wireless Single-Lead versus Standard 12-Lead ECG, for ST-Segment Deviation during Adenosine Cardiac Stress Scintigraphy', *Sensors*, vol. 23, no. 6, p. 2962, Mar. 2023, doi: 10.3390/s23062962.
[24] N. Rafie, A. H. Kashou, and P. A. Noseworthy, 'ECG Interpretation: Clinical Relevance, Challenges, and Advances', *Hearts*, vol. 2, no. 4, Art. no. 4, Dec. 2021, doi: 10.3390/hearts2040039.
[25] R. M. Rangayyan and R. J. Lehner, 'Phonocardiogram signal analysis: a review', *Crit. Rev. Biomed. Eng.*, vol. 15, no. 3, pp. 211–236, 1987.
[26] V. N. Varghees and K. I. Ramachandran, 'A novel heart sound activity detection framework for automated heart sound analysis', *Biomed. Signal Process. Control*, vol. 13, pp. 174–188, Sep. 2014, doi: 10.1016/j.bspc.2014.05.002.
[27] L. Resnekov, 'Understanding Heart Sounds and Murmurs, With an Introduction to Lung Sounds', *JAMA*, vol. 254, no. 1, pp. 124–125, Jul. 1985, doi: 10.1001/jama.1985.03360010134047.
[28] J. Nameche, J. Amaro, M. Silva, F. Ferreira, and F. Lopes, 'A Signal Acquisition and Processing Device to Assist Human Heart Sound-based Diagnosis', presented at the Portuguese Conf. on Pattern Recognition - RecPad, Oct. 2012.
[29] R. M. Rangayyan, Ed., *Biomedical Signal Analysis*, 1st ed. Wiley, 2015. doi: 10.1002/9781119068129.
[30] S. R. McGee, *Evidence-based Physical Diagnosis*. Elsevier Health Sciences, 2012.
[31] Z. Jiang *et al.*, 'Automated valvular heart disease detection using heart sound with a deep learning algorithm', *IJC Heart Vasc.*, vol. 51, p. 101368, Apr. 2024, doi: 10.1016/j.ijcha.2024.101368.
[32] W. Phanphaisarn, A. Roeksabutr, P. Wardkein, J. Koseeyaporn, and P. Yupapin, 'Heart detection and diagnosis based on ECG and EPCG relationships', *Med. Devices Auckl. NZ*, vol. 4, pp. 133–144, Aug. 2011, doi: 10.2147/MDER.S23324.
[33] F. Chakir, A. Jilbab, C. Nacir, and A. Hammouch, 'Recognition of cardiac abnormalities from synchronized ECG and PCG signals', *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 673–677, Jun. 2020, doi: 10.1007/s13246-020-00875-2.
[34] T. Shiraga *et al.*, 'Improving Valvular Pathologies and Ventricular Dysfunction Diagnostic Efficiency Using Combined Auscultation and Electrocardiography Data: A Multimodal AI Approach', *Sensors*, vol. 23, no. 24, p. 9834, Dec. 2023, doi: 10.3390/s23249834.
[35] P. Colli Franzone, L. F. Pavarino, and S. Scacchi, *Mathematical Cardiac Electrophysiology*, vol. 13. in MS&A, vol. 13. Cham: Springer International Publishing, 2014. doi: 10.1007/978-3-319-04801-7.
[36] F. Safara, S. Doraisamy, A. Azman, A. Jantan, and A. R. Abdullah Ramaiah, 'Multi-level basis selection of wavelet packet decomposition tree for heart sound classification', *Comput. Biol. Med.*, vol. 43, no. 10, pp. 1407–1414, Oct. 2013, doi: 10.1016/j.compbiomed.2013.06.016.

[37] R. M. F. L. da Silva and A. de Souza Maciel, 'Conduction Disorders: The Value of Surface ECG', *Curr. Cardiol. Rev.*, vol. 17, no. 2, pp. 173–181, Mar. 2021, doi: 10.2174/1573403X16666200511090151.
[38] D. R. Sarvamangala and R. V. Kulkarni, 'Convolutional neural networks in medical image understanding: a survey', *Evol. Intell.*, vol. 15, no. 1, pp. 1–22, 2022, doi: 10.1007/s12065-020-00540-3.
[39] Z. Ebrahimi, M. Loni, M. Daneshtalab, and A. Gharehbaghi, 'A review on deep learning methods for ECG arrhythmia classification', *Expert Syst. Appl. X*, vol. 7, p. 100033, Sep. 2020, doi: 10.1016/j.eswax.2020.100033.
[40] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
[41] A. Ghosh, A. Sufian, F. Sultana, A. Chakrabarti, and D. De, 'Fundamental Concepts of Convolutional Neural Network', in *Recent Trends and Advances in Artificial Intelligence and Internet of Things*, V. E. Balas, R. Kumar, and R. Srivastava, Eds., Cham: Springer International Publishing, 2020, pp. 519–567. doi: 10.1007/978-3-030-32644-9_36.
[42] S. Ioffe and C. Szegedy, 'Batch normalization: accelerating deep network training by reducing internal covariate shift', in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, in ICML'15. Lille, France: JMLR.org, Jul. 2015, pp. 448–456.
[43] D. Kingma and J. Ba, 'Adam: A Method for Stochastic Optimization', *Int. Conf. Learn. Represent.*, Dec. 2014.
[44] J. Duchi, E. Hazan, and Y. Singer, 'Adaptive Subgradient Methods for Online Learning and Stochastic Optimization', *J. Mach. Learn. Res.*, vol. 12, no. 61, pp. 2121–2159, 2011.
[45] G. Hinton, N. Srivastava, and K. Swersky, 'Lecture 6a: Overview of Mini-Batch Gradient Descent', 2020. [Online]. Available: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf
[46] F. Chakir, A. Jilbab, C. Nacir, and A. Hammouch, 'Recognition of cardiac abnormalities from synchronized ECG and PCG signals', *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 673–677, Jun. 2020, doi: 10.1007/s13246-020-00875-2.
[47] S. A. Singh, S. A. Singh, N. D. Devi, and S. Majumder, 'Heart Abnormality Classification Using PCG and ECG Recordings', *Comput. Sist.*, vol. 25, no. 2, Art. no. 2, May 2021, doi: 10.13053/cys-25-2-3447.
[48] J. Li, L. Ke, Q. Du, X. Chen, and X. Ding, 'Multi-modal cardiac function signals classification algorithm based on improved D-S evidence theory', *Biomed. Signal Process. Control*, vol. 71, p. 103078, Jan. 2022, doi: 10.1016/j.bspc.2021.103078.
[49] J. Wang, J. Zang, Q. An, H. Wang, and Z. Zhang, 'A pooling convolution model for multi-classification of ECG and PCG signals', *Comput. Methods Biomech. Biomed. Engin.*, pp. 1–14, Jan. 2024, doi: 10.1080/10255842.2023.2299697.
[50] C. Liu *et al.*, 'An open access database for the evaluation of heart sound algorithms', *Physiol. Meas.*, vol. 37, no. 12, pp. 2181–2213, Dec. 2016, doi: 10.1088/0967-3334/37/12/2181.
[51] P. Li, Y. Hu, and Z.-P. Liu, 'Prediction of cardiovascular diseases by integrating multi-modal features with machine learning methods', *Biomed. Signal Process. Control*, vol. 66, p. 102474, Apr. 2021, doi: 10.1016/j.bspc.2021.102474.
[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', *Comput. Biol. Med.*, vol. 138, p. 104914, Nov. 2021, doi: 10.1016/j.compbiomed.2021.104914.
[53] Z. Wang and T. Oates, 'Imaging time-series to improve classification and imputation', presented at the IJCAI International Joint Conference on Artificial Intelligence, 2015, pp. 3939–3945.

[54] M. Morshed and S. A. Fattah, 'A Deep Neural Network for Heart Valve Defect Classification From Synchronously Recorded ECG and PCG', *IEEE Sens. Lett.*, vol. 7, no. 9, pp. 1–4, Sep. 2023, doi: 10.1109/LSENS.2023.3307053.
[55] H. Han, M. Xiang, C. Lian, D. Liu, and Z. Zeng, 'A Multimodal Deep Neural Network for ECG and PCG Classification With Multimodal Fusion', in *2023 13th International Conference on Information Science and Technology (ICIST)*, Dec. 2023, pp. 124–128. doi: 10.1109/ICIST59754.2023.10367180.
[56] A. Kazemnejad, S. Karimi, P. Gordany, G. D. Clifford, and R. Sameni, 'An open-access simultaneous electrocardiogram and phonocardiogram database', *Physiol. Meas.*, vol. 45, no. 5, May 2024, doi: 10.1088/1361-6579/ad43af.
[57] Z. Wang, W. Yan, and T. Oates, 'Time series classification from scratch with deep neural networks: A strong baseline', in *2017 International Joint Conference on Neural Networks (IJCNN)*, May 2017, pp. 1578–1585. doi: 10.1109/IJCNN.2017.7966039.
[58] Y. Chang, M. Xiang, Z. Ling, Q. Liu, D. Liu, and C. Lian, 'A CNN-Transformer-Based Multimodal Model for Cardiac Abnormal Signal Detection', in *2023 International Conference on Neuromorphic Computing (ICNC)*, Dec. 2023, pp. 466–471. doi: 10.1109/ICNC59488.2023.10462741.
[59] H. Zhang *et al.*, 'Co-learning-assisted progressive dense fusion network for cardiovascular disease detection using ECG and PCG signals', *Expert Syst. Appl.*, vol. 238, p. 122144, Mar. 2024, doi: 10.1016/j.eswa.2023.122144.
[60] J. Y. Zhu, H. Liu, and X. W. Liu, 'Cardiovascular Disease Detection Based on Multi-Modal Data Fusion and Multi-Branch Residual Network', in *Proceedings of the 2023 International Conference on Frontiers of Artificial Intelligence and Machine Learning*, in FAIML '23. New York, NY, USA: Association for Computing Machinery, Mar. 2024, pp. 25–28. doi: 10.1145/3616901.3616907.
[61] M. F. Rifet Ibrahim, T. Alkanat, M. Meijer, A. Schlaefer, and P. Stelldinger, 'End-to-End Multi-Modal Tiny-CNN for Cardiovascular Monitoring on Sensor Patches', in *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, Mar. 2024, pp. 18–24. doi: 10.1109/PerCom59722.2024.10494450.
[62] G. M., V. Ravi, S. V, G. E.A, and S. K.P, 'Explainable Deep Learning-Based Approach for Multilabel Classification of Electrocardiogram', *IEEE Trans. Eng. Manag.*, vol. 70, no. 8, pp. 2787–2799, Aug. 2023, doi: 10.1109/TEM.2021.3104751.
[63] F. Liu *et al.*, 'An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection', *J. Med. Imaging Health Inform.*, vol. 8, no. 7, pp. 1368–1373, Sep. 2018, doi: 10.1166/jmihi.2018.2442.
[64] A. Anand, T. Kadian, M. K. Shetty, and A. Gupta, 'Explainable AI decision model for ECG data of cardiac disorders', *Biomed. Signal Process. Control*, vol. 75, p. 103584, May 2022, doi: 10.1016/j.bspc.2022.103584.
[65] P. Wagner *et al.*, 'PTB-XL, a large publicly available electrocardiography dataset', *Sci. Data*, vol. 7, no. 1, p. 154, May 2020, doi: 10.1038/s41597-020-0495-6.
[66] K. H. Le, H. H. Pham, T. B. T. Nguyen, T. A. Nguyen, T. N. Thanh, and C. D. Do, 'LightX3ECG: A Lightweight and eXplainable Deep Learning System for 3-lead Electrocardiogram Classification', *Biomed. Signal Process. Control*, vol. 85, p. 104963, Aug. 2023, doi: 10.1016/j.bspc.2023.104963.
[67] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, 'A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients', *Sci. Data*, vol. 7, no. 1, p. 48, Feb. 2020, doi: 10.1038/s41597-020-0386-x.
[68] N. Alamatsaz, L. Tabatabaei, M. Yazdchi, H. Payan, N. Alamatsaz, and F. Nasimi, 'A lightweight hybrid CNN-LSTM explainable model for ECG-based arrhythmia detection', *Biomed. Signal Process. Control*, vol. 90, p. 105884, Apr. 2024, doi: 10.1016/j.bspc.2023.105884.

[69] G. B. Moody and R. G. Mark, 'The impact of the MIT-BIH Arrhythmia Database', *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, May 2001, doi: 10.1109/51.932724.
[70] S. Petrutiu, A. V. Sahakian, and S. Swiryn, 'Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans', *EP Eur.*, vol. 9, no. 7, pp. 466–470, Jul. 2007, doi: 10.1093/europace/eum096.
[71] J. Ojha, H. Haugerud, A. Yazidi, and P. G. Lind, 'Exploring Interpretable AI Methods for ECG Data Classification', in *Proceedings of the 5th ACM Workshop on Intelligent Cross-Data Analysis and Retrieval*, in ICDAR '24. New York, NY, USA: Association for Computing Machinery, Jun. 2024, pp. 11–18. doi: 10.1145/3643488.3660294.
[72] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, H. Ghaemmaghami, and C. Fookes, 'A Robust Interpretable Deep Learning Classifier for Heart Anomaly Detection Without Segmentation', *IEEE J. Biomed. Health Inform.*, vol. 25, no. 6, pp. 2162–2171, Jun. 2021, doi: 10.1109/JBHI.2020.3027910.
[73] Yaseen, G.-Y. Son, and S. Kwon, 'Classification of Heart Sound Signal Using Multiple Features', *Appl. Sci.*, vol. 8, no. 12, Art. no. 12, Dec. 2018, doi: 10.3390/app8122344.
[74] Z. Wang, K. Qian, H. Liu, B. Hu, B. W. Schuller, and Y. Yamamoto, 'Exploring interpretable representations for heart sound abnormality detection', *Biomed. Signal Process. Control*, vol. 82, p. 104569, Apr. 2023, doi: 10.1016/j.bspc.2023.104569.
[75] A. Bhardwaj, S. Singh, and D. Joshi, 'Explainable Deep Convolutional Neural Network for Valvular Heart Diseases Classification Using PCG Signals', *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–15, 2023, doi: 10.1109/TIM.2023.3274174.
[76] A. Mordvintsev and M. Tyka, 'Inceptionism: Going Deeper into Neural Networks', 2015, [Online]. Available: https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html
[77] C. Divakar, R. Harsha, K. Radha, D. V. Rao, N. Madhavi, and T. Bharadwaj, 'Explainable AI for CNN-LSTM Network in PCG-Based Valvular Heart Disease Diagnosis', in *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Jan. 2024, pp. 92–97. doi: 10.1109/Confluence60223.2024.10463207.
[78] S. Li, J. Sun, H. Yang, J. Pan, T. Guo, and W. Wang, 'Interpretable End-to-End heart sound classification', *Measurement*, vol. 237, p. 115113, Sep. 2024, doi: 10.1016/j.measurement.2024.115113.
[79] H. Weytjens and J. De Weerdt, 'Process Outcome Prediction: CNN vs. LSTM (with Attention)', in *Business Process Management Workshops*, A. Del Río Ortega, H. Leopold, and F. M. Santoro, Eds., Cham: Springer International Publishing, 2020, pp. 321–333. doi: 10.1007/978-3-030-66498-5_24.
[80] S. E. Schmidt, C. Holst-Hansen, C. Graff, E. Toft, and J. J. Struijk, 'Segmentation of heart sound recordings by a duration-dependent hidden Markov model', *Physiol. Meas.*, vol. 31, no. 4, p. 513, Mar. 2010, doi: 10.1088/0967-3334/31/4/004.
[81] G. Saxena, N. Farooqui, and S. Ali, 'Extricate Features Utilizing Mel Frequency Cepstral Coefficient in Automatic Speech Recognition System', *Int. J. Eng. Manuf.*, vol. 12, pp. 14–21, Dec. 2022, doi: 10.5815/ijem.2022.06.02.
[82] 'Speech Communications: Human and Machine, 2nd Edition | Wiley', Wiley.com. Accessed: Jul. 27, 2024. [Online]. Available: https://www.wiley.com/en-us/Speech+Communications%3A+Human+and+Machine%2C+2nd+Edition-p-9780780334496
[83] S. Braun, 'WINDOWS', in *Encyclopedia of Vibration*, S. Braun, Ed., Oxford: Elsevier, 2001, pp. 1587–1595. doi: 10.1006/rwvb.2001.0052.
[84] S. A. Fulop, 'The Fourier Power Spectrum and Spectrogram', in *Speech Spectrum Analysis*, S. A. Fulop, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 69–106. doi: 10.1007/978-3-642-17478-0_4.

[85] P. Singh, 'An Approach to Extract Feature using MFCC', *IOSR J. Eng.*, vol. 4, pp. 21–25, Aug. 2014, doi: 10.9790/3021-04812125.
[86] Md. A. Hossan, S. Memon, and M. A. Gregory, 'A novel approach for MFCC feature extraction', in *2010 4th International Conference on Signal Processing and Communication Systems*, Dec. 2010, pp. 1–5. doi: 10.1109/ICSPCS.2010.5709752.
[87] N. Giordano, S. Rosati, G. Balestra, and M. Knaflitz, 'Can Multi-source Phonocardiography Enable Inexperienced Users to Record Heart Sounds for Telemonitoring Applications? A Comparative Analysis', presented at the 2023 Computing in Cardiology Conference, Nov. 2023. doi: 10.22489/CinC.2023.126.
[88] M. Aqil and A. Jbari, 'Continuous Wavelet Analysis and Extraction of ECG Features', in *Emerging Technologies in Biomedical Engineering and Sustainable TeleMedicine*, J. Alja'am, S. Al-Maadeed, and O. Halabi, Eds., Cham: Springer International Publishing, 2021, pp. 51–68. doi: 10.1007/978-3-030-14647-4_5.
[89] B. Russell and J. Han, 'Jean Morlet and the Continuous Wavelet Transform', vol. 28, 2016.
[90] P. Wallisch, M. Lusignan, M. Benayoun, T. I. Baker, A. S. Dickey, and N. G. Hatsopoulos, 'Chapter 9 - Wavelets', in *Matlab for Neuroscientists*, P. Wallisch, M. Lusignan, M. Benayoun, T. I. Baker, A. S. Dickey, and N. G. Hatsopoulos, Eds., London: Academic Press, 2009, pp. 133–140. doi: 10.1016/B978-0-12-374551-4.00009-9.
[91] P. Parida and N. Bhoi, 'Wavelet based transition region extraction for image segmentation', *Future Comput. Inform. J.*, vol. 2, no. 2, pp. 65–78, Dec. 2017, doi: 10.1016/j.fcij.2017.10.005.
[92] D. F. Walnut, *An Introduction to Wavelet Analysis*. in Applied and Numerical Harmonic Analysis. Boston, MA: Birkhäuser Boston, 2004. doi: 10.1007/978-1-4612-0001-7.
[93] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization', in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 618–626. doi: 10.1109/ICCV.2017.74.
[94] L. S. Shapley, '17. A Value for n-Person Games', in *Contributions to the Theory of Games (AM-28), Volume II*, H. W. Kuhn and A. W. Tucker, Eds., Princeton University Press, 1953, pp. 307–318. doi: 10.1515/9781400881970-018.
[95] S. M. Lundberg and S.-I. Lee, 'A unified approach to interpreting model predictions', in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 4768–4777.
[96] M. Sundararajan, A. Taly, and Q. Yan, 'Axiomatic attribution for deep networks', in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, in ICML'17. Sydney, NSW, Australia: JMLR.org, Aug. 2017, pp. 3319–3328.
[97] G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, and S.-I. Lee, 'Improving performance of deep learning models with axiomatic attribution priors and expected gradients', *Nat. Mach. Intell.*, vol. 3, no. 7, pp. 620–631, Jul. 2021, doi: 10.1038/s42256-021-00343-w.
[98] R. B. Ford and E. M. Mazzaferro, 'Section 2 - Patient Evaluation and Organ System Examination', in *Kirk & Bistner's Handbook of Veterinary Procedures and Emergency Treatment (Ninth Edition)*, R. B. Ford and E. Mazzaferro, Eds., Saint Louis: W.B. Saunders, 2012, pp. 295–380. doi: 10.1016/B978-1-4377-0798-4.00002-5.
[99] Y. Xia, W. Wang, and K. Wang, 'ECG signal generation based on conditional generative models', *Biomed. Signal Process. Control*, vol. 82, p. 104587, Apr. 2023, doi: 10.1016/j.bspc.2023.104587.

[100] I. Neves *et al.*, 'Interpretable heartbeat classification using local model-agnostic explanations on ECGs', *Comput. Biol. Med.*, vol. 133, p. 104393, Jun. 2021, doi: 10.1016/j.compbiomed.2021.104393.
[101] S. Seoni *et al.*, 'Application of spatial uncertainty predictor in CNN-BiLSTM model using coronary artery disease ECG signals', *Inf. Sci.*, vol. 665, p. 120383, Apr. 2024, doi: 10.1016/j.ins.2024.120383.