POLITECNICO DI TORINO

Master Degree
in Biomedical Engineering

Master Thesis

Deep learning model optimization for the classification of 27 cardiac arrhythmias:
a study on performance improvement and clinical feedback integration

Supervisor
Prof.ssa Gabriella OLMO
Dott.ssa Federica AMATO
Dott. Marco BOLOGNA

Candidate
Ivan CALDERONE

Academic Year 2023-2024

Abstract

Cardiovascular disease poses a significant health risk and remains one of the leading causes of death worldwide. The 12-lead electrocardiogram is a comprehensive and widely accessible diagnostic tool for identifying cardiac abnormalities. Early and accurate diagnosis allows for timely treatment and intervention, helping to prevent severe complications. It is therefore crucial to have automated classification tools capable of recognizing these cardiac alterations, streamlining clinical practice, which is currently very time-consuming. This thesis contributes to this context by building on a solution developed by a team ranked third in the PhysioNet/Computing in Cardiology Challenge 2020, which aimed to classify 27 cardiac arrhythmias using a new scoring metric. This metric awards partial credit for misdiagnoses that result in similar outcomes or treatments as the correct diagnosis, according to cardiologists. Starting from the original model, a squeeze-and-excite ResNet ensemble using the same signal from eight truncated leads of varying lengths, several preprocessing steps were applied to improve performance. Furthermore, clinical expertise was incorporated through collaboration with a cardiologist from a hospital in Turin. One of the most promising modifications, integrated into the final solution, was the addition of features that describe the patient's condition, such as gender and age, indicating that more patient-specific information leads to better outcomes and more tailored treatment approaches. After multiple refinements, the final model consists of two phases: the first is a binary model that discriminates between healthy and altered classes, achieving an accuracy of 88% and a macro F1 score of 93%. The second phase distinguishes 19 classes by grouping certain arrhythmias and integrating clinical knowledge. The results are promising, with overall performance improvements on both the validation set and hidden test set. The integration of clinical data has proven to be crucial, highlighting the importance of close collaboration with field experts to refine the model and ensure its clinical applicability in real-world settings.

Contents

List of Tables

List of Figures

List of Abbreviations

ECG Electrocardiogram

SA Sinoatrial

AV Atrioventricular

PVC Premature ventricular contractions

PAC Premature atrial contractions

AI Artificial intelligence

CNN Convolutional neural network

RNN Recurrent neural network

LSTM Long Short-Term Memory

BiLSTM Bidirectional Long Short-Term Memory

ATT Attention

MIT-BIH Massachusetts Institute of Technology - Beth Israel Hospital

CPSC China Physiological Signal Challenge

PTB Physikalisch-Technische Bundesanstalt

WFDB WaveForm DataBase

SNOMED Systematized Nomenclature of Medicine

GPU Graphical processing unit

CPU Central processing unit

SSH Secure Shell

HIPAA Health Insurance Portability and Accountability Act

ResNet Residual Network

SE Squeeze-and-Excitation

BCE Binary Cross Entropy

ReLU Rectified Linear Unit

HRV Heart Rate Variability

CSI Cardiac Sympathetic Index

CVI Cardiac Vagal Index

AUC Area under the curve

MRI Magnetic Resonance Imaging

Chapter 1

Background

1.1 Thesis objective

The primary objective of this thesis is to develop and refine a model capable of accurately classifying 27 different cardiac arrhythmias based on electrocardiogram signals, building on a solution originally ranked third in the PhysioNet/Computing in Cardiology Challenge 2020. This model aims to improve the efficiency of arrhythmia classification through deep learning techniques while ensuring clinical relevance by incorporating expert knowledge from cardiologists. The entire project was developed through collaboration with the company SynbrAIn, which provided specific and essential equipment for carrying out this work, contributing to the development and validation of the model. The ultimate goal is to create a system that not only performs well in terms of accuracy and robustness but also integrates patient-specific data, such as age and gender, to tailor predictions more effectively. By collaborating closely with cardiologists and integrating clinical insights, the thesis strives to develop a tool that can be directly applicable in clinical practice, ultimately aiding in the timely and accurate diagnosis of cardiac arrhythmias.

1.2 Thesis organisation

The thesis is organized as follows. In Chapter 2, an introduction is provided to the anatomy and physiology of the heart, the tools used to record its electrical activity, and the solutions implemented so far, with a particular focus on the PhysioNet/Computing in Cardiology Challenge 2020. Chapter 3 describes the datasets used for this Challenge (Section 3.2), the instrumentation used thanks to the collaboration with the company (Section 3.1), and the workflow implemented, with particular attention to the pre-processing steps (Section 3.3) and the model used

for the classification task (Section 3.4). Chapter 4 presents the results obtained, while Chapter 5 provides an analysis of the results, with particular attention to the limitations encountered during this project.

Chapter 2

Introduction

2.1 Heart and Cardiac Arrhythmias: an Overview

2.1.1 Anatomy and Physiology of the Heart

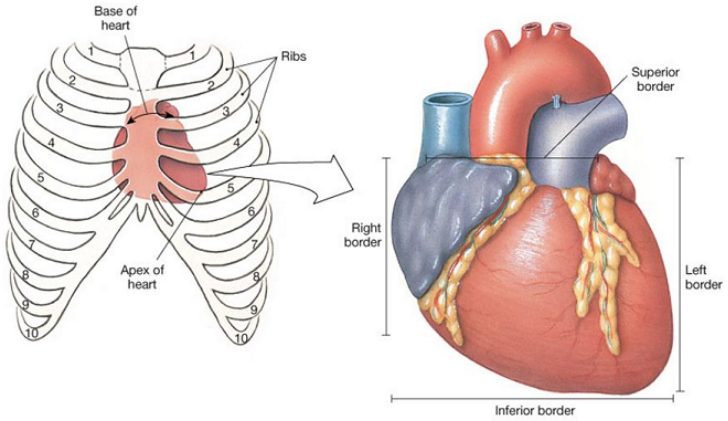

Figure 2.1. Position of the heart.

The heart, a hollow muscular organ located in the mediastinum, shown in Fig. 2.1, has a conical shape and dimensions similar to those of a fist, with the apex pointing downwards and to the left. This vital organ is divided into two distinct sections: the left section, through which oxygenated blood circulates, and the right section, where venous blood rich in carbon dioxide flows. Each section comprises two cavities: an upper one, called the atrium, and a lower one, known as the ventricle. The atria communicate with the underlying ventricles through the atrioventricular orifice, equipped with a cardiac valve that regulates blood flow; the

tricuspid valve is located on the right, while the bicuspid or mitral valve is situated on the left. The cardiac valves play a crucial role in preventing the backflow of blood into the atria during the systolic phase, which is the ventricular contraction. Tha anatomy of the heart is shown in Fig. 2.2.

The heart is structurally divided into two halves: the interatrial septum divides the atrial portion, while the interventricular septum separates the ventricles. The entire organ is enveloped by a protective membrane called the pericardium, and its wall is composed of three distinct layers: the epicardium, the outer layer; the myocardium, the intermediate and thickest layer, responsible for cardiac contraction; and the endocardium, the innermost layer that lines the heart cavities. [33]

In the context of blood flow, the right atrium receives deoxygenated blood from the superior and inferior vena cava, which transport it from peripheral tissues. From here, the blood passes to the right ventricle, from which the pulmonary artery originates, responsible for carrying it to the lungs for reoxygenation. Conversely, the left atrium receives oxygenated blood through four pulmonary veins, and from the left ventricle, the aorta originates, distributing oxygenated blood throughout the body. Being a constantly active muscle, the heart requires an abundant supply

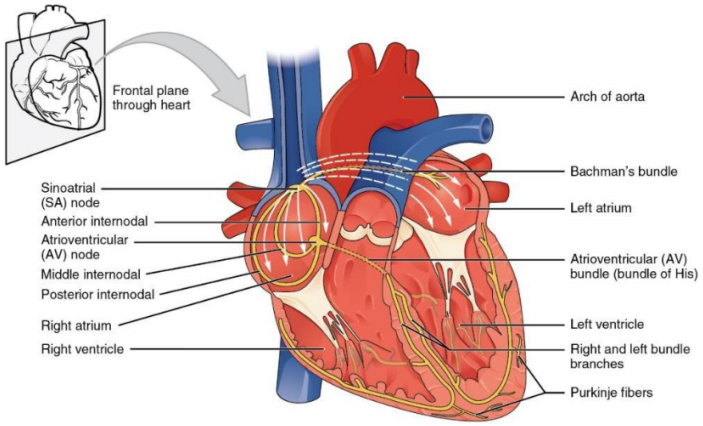

Figure 2.2. Anatomy of the heart and description of the blood flow.

of oxygen and nutrients, which is guaranteed by the coronary system. This system consists of a network of vessels that branch off from the aorta and cover the heart, ensuring the necessary supply to all the cells of the cardiac tissue. The left and right coronary arteries branch into an extensive network of increasingly smaller arteries, reaching the capillaries, which ensure the capillary nutrition of the myocardium. The functioning of the heart is marked by a cardiac cycle, a complex process that repeats rhythmically, ensuring the proper circulation of blood throughout the body. The cardiac cycle, shown in Fig. 2.3 is divided into two main phases: diastole and

systole, which respectively represent the periods of relaxation and contraction of the heart chambers. [26]

During diastole, the ventricles relax, allowing blood to flow from the atria to the ventricles through the atrioventricular orifices, whose valves (tricuspid on the right and mitral on the left) are open. In this phase, the pressure in the ventricles is lower than that in the atria, favoring ventricular filling. At the same time, the semilunar valves (aortic and pulmonary) remain closed, preventing the backflow of blood from the aorta and pulmonary artery into the ventricles.

With the completion of ventricular filling, the cycle enters the systolic phase. During ventricular systole, the ventricles contract in response to electrical impulses generated by the sinoatrial node and propagated through the heart's conduction system. Ventricular contraction causes an increase in pressure within the ventricles, which leads to the closure of the atrioventricular valves, preventing the backflow of blood into the atria. Simultaneously, the rising ventricular pressure exceeds that in the arteries, causing the semilunar valves to open and allowing the expulsion of blood into the main arterial vessels: the aorta and the pulmonary artery. Once the blood has been ejected, the ventricles begin to relax, and the

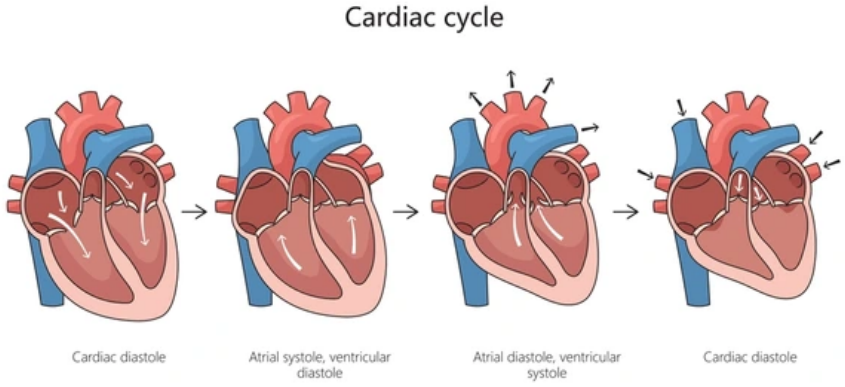

Figure 2.3. Graphical description of the cardiac cycle.

pressure within them rapidly decreases. This leads to the closure of the semilunar valves, preventing the backflow of blood into the ventricles, and marks the beginning of a new diastolic phase, during which the heart prepares to receive venous and oxygenated blood from the atria once again. The cardiac cycle, therefore, is a continuous and coordinated process, whose efficiency is ensured by the precise interaction between myocardial contractions, pressure variations in the heart chambers, and the synchronous functioning of the valves. This cycle ensures that

blood is constantly pumped through the pulmonary and systemic circuits, guaranteeing tissue oxygenation and the maintenance of homeostasis. The heart, through its muscular structure and the complex organization of its cavities, cyclically performs a series of phases that constitute the cardiac cycle, ensuring the continuous circulation of blood. This process is tightly regulated by an intrinsic electrical conduction system that coordinates the contractions of the heart chambers, thus generating the heartbeat. The heartbeat begins in the sinoatrial (SA) node, located in the right atrium, which acts as the natural pacemaker of the heart. The SA node generates spontaneous electrical impulses that propagate through the atrial walls, causing the atria to contract and blood to pass into the ventricles. After reaching the atrioventricular (AV) node, the impulse undergoes a brief delay, allowing the ventricles to complete their filling before contracting. Subsequently, the electrical impulse rapidly spreads through the bundle of His and the Purkinje fibers, causing the synchronized contraction of the ventricles and the expulsion of blood into the arteries. This delicate balance of contractions is recorded by the electrocardiogram, which captures the electrical activity of the heart, providing a visual representation of the waves corresponding to the different phases of the cardiac cycle. The electrocardiogram (ECG), therefore, becomes an essential tool for evaluating cardiac function and diagnosing any abnormalities in the rhythm or conduction of the heartbeat.

2.1.2 Electrocardiogram

The electrical activity of heart cells generates a flow of currents within the heart, which manifests as potential variations on the skin's surface, as shown in Fig. 2.4. The variation of the potential difference can be measured using specific devices, and their variation thruogh time constitutes the electrocardiogram. Practically, the ECG consists of 12 temporal traces representing as many potential differences, detected between various points on the body surface using electrodes placed on the body. Essentially, the ECG is the projection in 12 directions in three-dimensional space of the cardiac vector, resulting from the moments of the electric dipoles generated in the heart during the wavefront progression. Each of the 12 ECG leads represents the magnitude of the cardiac vector, and thus the electrical activity, in the corresponding direction at each moment in time. The 12 directions are selected to divide the space to express the activity in the right-left, superior-inferior, and anterior-posterior orientations of the body.

The ECG is a pseudo-periodic signal, whose amplitude generally ranges in the order of millivolts. Its morphology typically consists of five peaks, positive or negative in the presence of anomalies, which are called waves. These waves represent deviations of the signal from the baseline and are denoted by the letters P, Q, R, S, and T. The P wave corresponds to the depolarization of the atria, while the

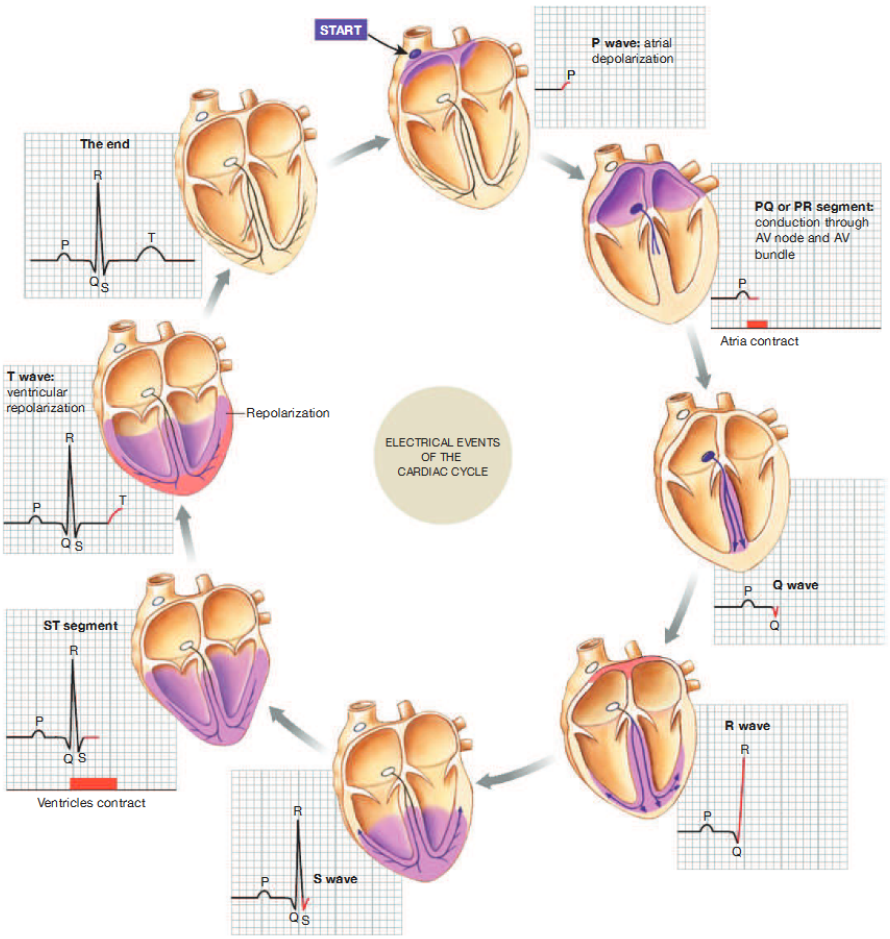

Figure 2.4. Correlation between an ECG and the electrical events in the heart. In each phase of the cycle, the regions of depolarization (in purple) and repolarization (in orange) are highlighted.

QRS complex represents the depolarization of the ventricles, occurring at the level of the septum, apex, and base. The T wave, finally, reflects the repolarization of the ventricles. In the ECG, a specific wave for atrial repolarization is not visible, as it is masked by the simultaneous ventricular depolarization. Additionally, the ECG contains segments representing periods where no potential differences are recorded, while the intervals between the waves indicate the times taken for conduction, depolarization, or periods with no electrical activity. For example, the PR interval represents the time required for the potential to propagate from the atria to the ventricles.

The basic arrangement of electrodes for recording the ECG involves creating

a triangle, whose vertices are ideally located near the roots of the upper limbs and the pubic region. However, for convenience, the electrodes are usually placed on the wrists and the left leg. This triangle is commonly known as Einthoven's Triangle and it is shown in Fig. 2.5.

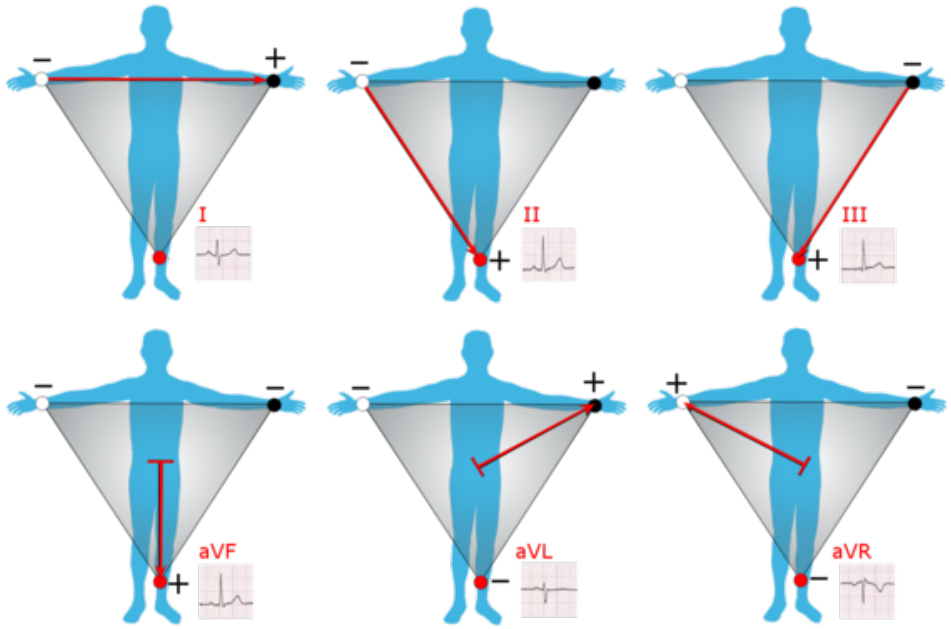

Figure 2.5. Bipolar leads and unipolar leads. Einthoven's triangle.

The standard ECG leads are twelve and are divided into three main categories: bipolar leads, augmented unipolar leads of Goldberger, and precordial unipolar leads of Wilson. Regarding the bipolar leads, or main leads, some conventions must be adopted: lead I represents the potential difference between the left wrist and the right wrist; lead II reflects the potential difference between the left ankle and the right wrist; finally, lead III is given by the difference between the potential of the left ankle and that of the left wrist. These leads are measured on the frontal plane and are linearly dependent on each other, following Kirchhoff's law, where I + III = II. However, these leads are not sufficient to adequately record all possible variations of the cardiac vector, making it necessary to adopt additional leads, such as the augmented leads of Goldberger, to improve the evaluation of cardiac events.

The unipolar leads of Goldberger, measured on the frontal plane, are obtained along the bisectors of Einthoven's triangle and reflect the potential difference between the explored limb and the average potentials of the other two unexplored

limbs. The term augmented refers to the need to amplify the signal to make it comparable with the other leads. The resulting leads are lead aVL, aVR, and aVF.

$$aVF = LL - \frac{1}{2}(RA + LA) \tag{2.1}$$

$$aVR = RA - \frac{1}{2}(LA + LL) \tag{2.2}$$

$$aVL = LA - \frac{1}{2}(RA + LL) \tag{2.3}$$

where RA is the potential in the right arm, LA is the potential in the left arm, and LL is the potential in the left leg.

These six leads can be derived using four electrodes, three exploring and one reference, placed at a certain distance from the heart. However, since these electrodes are placed far from the heart, they do not allow precise analysis of anomalies localized in specific areas of the heart muscle. For this reason, it is necessary to have electrodes closer to the heart, placed on the chest. The leads thus obtained, called precordial leads, are obtained by considering the potential difference between a reference electrode, placed on the sternum in the fourth intercostal space, and six other points appropriately located in the intercostal space. The position is shown in Fig. 2.6. These leads are linearly independent of each other and are identified as leads V1, V2, V3, V4, V5, and V6. [19]

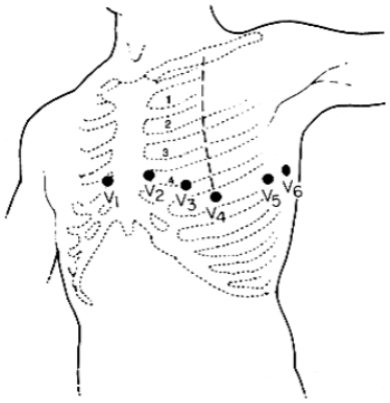

Figure 2.6. Position of the electrodes for precordial leads.

The ECG presents normal values that indicate the regular functioning of the heart. Among the most characteristic parameters, we find the heart rate, which under normal conditions ranges between 60 and 100 beats per minute, the PR interval, which should be between 120 and 200 milliseconds, and the QRS complex,

whose duration varies between 80 and 100 milliseconds. If these values deviate from the normal ranges, they could indicate the presence of heart diseases or rhythm abnormalities, making the ECG an essential tool for the diagnosis and monitoring of heart conditions.

2.1.3 Cardiac Arrhythmias Types and ECGs Implications

Figure 2.7. ECG showing a normal sinus rhythm: regular P waves, narrow QRS complexes, well-defined T waves, and consistent wave shapes.

The term arrhythmia refers to any alteration in the frequency or rhythm of the heartbeat. Cardiac arrhythmias represent a varied set of disorders resulting from changes in the generation or conduction of electrical impulses in the heart. These impulses can travel at speeds higher or lower than normal, or cause an irregular heartbeat. Under normal conditions, the heart follows a regular rhythm, controlled by an internal electrical system that coordinates the contractions of the atria and ventricles. This mechanism is regulated by the sinoatrial node, which acts as the heart's natural pacemaker by generating electrical impulses that propagate through specific conduction pathways. When the heart functions correctly, the electrocardiogram, as shown in Fig. 2.7, displays a normal sinus rhythm, characterized by regular P waves, followed by QRS complexes and a stable ST segment. This tracing reflects an orderly and synchronized electrical conduction through the heart.

In the presence of arrhythmias, this balance can be compromised. Arrhythmias can manifest as tachycardias, which is an acceleration of the heartbeat exceeding 100 beats per minute, or as bradycardias, where the heartbeat slows down below 60 beats per minute due to a decrease in the discharge rate of the sinus node. Bradycardias can result from dysfunction of the sinoatrial node or conduction blocks. In the ECG, bradycardia presents with a heart rate lower than normal but with a regular rhythm and well-defined P waves. [34]

A particularly significant rhythm anomaly is atrial fibrillation, one of the most common and potentially serious arrhythmias, with significant implications for cardiac health and patients' quality of life. In this condition, the atria activate in a

chaotic and disorganized manner, resulting in the absence of distinct P waves in the ECG. These waves are replaced by irregular fibrillatory waves, accompanied by an equally irregular and variable ventricular rhythm, as evidenced by non-uniform RR intervals. These signs are crucial for distinguishing atrial fibrillation from other forms of arrhythmia, such as atrial flutter, which manifests with rapid and regular P waves, creating a characteristic sawtooth pattern in the ECG. [30]

Other cardiac anomalies include ectopic beats such as premature ventricular contractions (PVC) and premature atrial contractions (PAC). PVC are premature contractions originating from the ventricles and appear in the ECG as wide and deformed QRS complexes, followed by a compensatory pause. PAC, on the other hand, are premature contractions originating from the atria and appear in the ECG as premature P waves, often followed by a compensatory interval.

Among the anomalies detectable on the ECG are also deviations of the heart's electrical axis from normal values. Right axis deviation occurs when the axis is shifted to the right, while left axis deviation occurs when the axis is shifted to the left. These deviations can be associated with various clinical conditions, such as bundle branch blocks. A bundle branch block occurs when there is an interruption in the conduction of the electrical impulse through one of the branches of the ventricular conduction system, which can be a right or left bundle branch block. In right bundle branch block, the ECG shows widened QRS complexes and a specific pattern called rabbit ears in the precordial leads. In left bundle branch block, the QRS complexes are also widened, with a typical notched pattern in the lateral leads. These blocks can indicate underlying heart diseases and affect overall cardiac function. Some example of cardiac anomalies are shown in Fig. 2.8. [13]

The shape of the waves and the duration of the intervals in the ECG can also provide valuable indications of cardiac anomalies. For example, the T wave represents ventricular repolarization and, if inverted, can indicate myocardial ischemia or other conditions. A prolonged PR interval, which reflects the time required for the impulse to conduct from the atria to the ventricles, can suggest an atrioventricular block. This slowdown in conduction can result from various factors affecting the passage of the electrical potential from one heart chamber to another.

In conclusion, alterations in the electrocardiographic signal are fundamental for the diagnosis of arrhythmias, which are closely interconnected. The ECG is a crucial diagnostic tool for identifying and managing cardiac arrhythmias, allowing not only the detection of rhythm alterations but also obtaining essential details on anomalies in the heart's electrical conduction. The ability to monitor and evaluate these anomalies is essential for ensuring effective management of arrhythmias and improving patients' quality of life.

Figure 2.8. Examples of arrhythmias and their shapes in the ECG. **A.** Normal sinus rhythm. **B.** Atrial fibrillation. **C.** Left bundle branch block. **D.** Right bundle branch block. **E.** Premature atrial contraction. **F.** Premature ventricular contraction. **G.** Ectopic beats. **H.** Myocardial Infarction. **I.** Sinus Bradycardia. **J.** Supraventricular Tachycardia. **K.** Atrial flutter. **L.** Ventricular fibrillation.

2.2 Importance of diagnosing cardiac arrhythmias and development of automatic processing algorithms

Manual interpretation of ECGs is a time-consuming process that requires highly qualified personnel to achieve an accurate diagnosis. The problem becomes more pronounced in places where there is a shortage of medical experts and clinical equipment, especially in developing countries. This motivates the need for a reliable, automatic, and low-cost system for monitoring and diagnosis. In fact, automation in the detection and classification of cardiac anomalies can support doctors in diagnosing an increasing number of recorded ECGs. However, successes in this field have been limited. Over the past decade, the rapid development of machine learning techniques has included an increasing number of 12-lead ECG classifiers in the form of both time sequences and two-dimensional images. Many of these algorithms are capable of correctly identifying cardiac anomalies, but most have been trained, tested, or developed on single, small, or relatively homogeneous datasets. Additionally, many methods focus on identifying a limited number of cardiac arrhythmias, which do not represent the complexity and difficulty of interpreting ECGs.

2.2.1 State of the art

Arrhythmia classification can be divided into two main categories. The first concerns morphological arrhythmias, which form due to an irregularity in a single heartbeat, while the second includes rhythmic arrhythmias, generated by a series of irregular heartbeats. To deal with both macro-categories of arrythmias this problem, various databases have been made available online, some characterized by labels for each individual beat, such as the MIT-BIH Arrhythmia Database, and others containing ECG traces labeled in their entirety. The main stages involved in arrhythmia classification include data pre-processing, feature extraction, feature dimension reduction or optimization (if applicable), classification, and, if necessary, a post-processing phase. [17]

ECG signal pre-processing includes various steps such as filtering to remove baseline and powerline interference, noise reduction using band-pass filters, artifact removal, amplitude normalization to ensure uniformity, R-peak detection to identify ventricular depolarization, segmentation to isolate individual beats, quality control to manage artifacts, resampling to obtain a uniform sampling frequency, interpolation to handle missing data, and heart rate normalization for comparative analysis. These steps ensure that ECG data are clean and standardized, thus facilitating the diagnosis and monitoring of cardiac issues. [7]

Regarding feature extraction, numerous solutions have been proposed that integrate clinical knowledge with automatic classification tools to improve performance. The extracted features often belong to three different domains: temporal, with measures of heart rate variability; frequency domain, with statistical measures such as skewness and entropy; and finally, nonlinear features. These characteristics are subsequently input into classification models or sometimes concatenated with the results of convolution operations to add information not extracted by neural networks. It can thus be stated that these machine learning-related features play a crucial role in the analysis and classification of ECG data for arrhythmia detection and disease diagnosis, supporting signal interpretation. [11]

In the context of automatic ECG trace classification, early approaches were based on machine learning techniques, such as rule-based classification methods and clustering algorithms. The advent of artificial intelligence (AI) and advanced computational techniques has significantly revolutionized diagnostics in healthcare. With technological progress, artificial neural networks and deep learning algorithms have gained increasing importance. AI-based methods, which integrate neural networks, Bayesian networks, fuzzy logic systems, and machine learning models such as linear or logistic regression, decision trees, k-nearest neighbors, random forests, and support vector machines, have proven capable of accurately predicting cardiovascular outcomes in patients. However, these traditional methods often require extensive feature engineering and may struggle with highly complex or unstructured data, limiting their ability to generalize. In contrast, deep learning approaches, which automatically learn significant features from ECG data, continuously improve classification capabilities, reducing the need for human intervention and addressing some of the limitations of previous models.

Among the most relevant deep learning models are convolutional neural networks (CNN) and recurrent neural networks (RNN). Some examples of CNN-based models include a model consisting of 16 convolution blocks with residual connections proposed to detect 12 kinds of heart arrhythmias from the original single lead II ECG input, which achieves an F1 average score of 0.837, better than the average of cardiologists (0.780) [20], or a deep learning model of 10 layers proposed for the diagnosis of 10 types of myocardial infarction based on 12 lead ECG signals with 2-s of signals as input, and attain a diagnostic accuracy of 98.97%. [2] More recently, a proposed model combines a transformer with a convolutional neural network and a denoising autoencoder for inter-patient ECG arrhythmia classification. The transformer is used to capture long-range dependencies in ECG data, improving the model's ability to understand complex temporal sequences. The CNN is integrated to extract local features from ECG signals, enhancing the model's ability to identify specific arrhythmia patterns. The denoising autoencoder is employed to reduce noise in ECG data, improving signal quality and, consequently, classification accuracy. The model is designed to address the challenges of arrhythmia

classification in inter-patient contexts, where individual variations can make accurate identification of cardiac anomalies difficult. [48] Another recent solution that has shown particular interest is the one proposed by [28]. In this work, the authors present an automated multi-label cardiac arrhythmia classification network using CNN, which can detect and classify 45 cardiac arrhythmia classes, surpassing many previous models in terms of the number of classes handled and the quality of predictions. The model incorporates both the residual structure and channel attention mechanism. Thus, two key schemes have been developed to improve classification performance: the Global Channel Attention Block and the Short Residual Block. The Global Channel Attention Block incorporates dilated convolutions to preserve overall features. It focuses on the important characteristics of each arrhythmia class from the original electrocardiogram data during the training process. The Short Residual Block employs a residual structure to enhance classification accuracy. The network's performance is evaluated using a large-scale 12-lead electrocardiogram database for arrhythmia study on PhysioNet, specifically Shaoxing People's Hospital and Ningbo First Hospital, and the 2018 China Physiological Signal Challenge (CPSC) dataset. The proposed classification network performs well in all average evaluation metrics on the Ningbo and Shaoxing dataset, but did not perform well in all evaluation metrics when tested on the CPSC 2018 dataset. This again highlights how the proposed solutions are highly specific to a single dataset and similarly struggle to generalize to new ones.

Regarding models based on recurrent neural networks, they have achieved promising results in classifying heart rhythms. A recent example of this type of classifier is a model based on self-attention Long Short-Term Memory Fully Convolutional Network (LSTM-FCN) for arrhythmia classification and uncertainty assessment. ArrhyMon is designed to detect and classify six different types of arrhythmias, in addition to normal ECG patterns. This model integrates fully convolutional network layers and a self-attention-based LSTM architecture to capture and exploit both global and local features embedded in ECG sequences. Additionally, ArrhyMon incorporates an uncertainty model based on a deep ensemble that generates a confidence measure for each classification result. The model is evaluated using three publicly available datasets (MIT-BIH, PhysioNet Cardiology Challenge 2017 and 2020/2021), demonstrating state-of-the-art classification performance with an average accuracy of 99.63%. The confidence measures shown by the model closely correlated with subjective diagnoses made by physicians. [39]

Another relevant aspect is that artificial intelligence can also be applied to ECG signals from wearable devices, enabling real-time analysis and interpretation of cardiovascular health. Thanks to these intelligent wearable devices, equipped with sophisticated algorithms, users can proactively manage their heart health by constantly monitoring and interpreting ECG data.

Among the currently available solutions of particular interest is the one proposed in [23], which uses a smartphone application to analyze a photographic image of a 12-lead ECG, offering an automated interpretation that can subsequently be validated by cardiologists. The application is based on an AI-powered system composed of six deep neural networks, trained on standard 12-lead ECGs to detect 20 essential diagnostic patterns, divided into six categories: rhythm, acute coronary syndrome, conduction abnormalities, ectopia, cardiac chamber enlargement, and cardiac axis. The system, trained on over 900000 ECGs from publicly available datasets (such as PTB-XL or CPSC), has demonstrated superior diagnostic performance compared to traditional approaches in 13 of the 20 evaluated patterns, and was not inferior for the remaining ones. However, among the issues encountered is the quality of the captured image, particularly the difficulty in focusing on the ECG trace from which the classification is performed, an issue that will require future studies and modifications. [22]

Despite this, these solutions are particularly interesting as they allow, with high accuracy, albeit lower than the one of the doctors, the recognition of numerous cardiac arrhythmias. They also promote a more proactive approach to health management, marking the beginning of a new era of preventive and patient-centered clinical care. However, some critical issues remain, such as managing limited computational resources and the need to comply with strict data security and privacy standards in the medical field, which require a delicate balance. This complex and constantly evolving landscape highlights both the challenges and opportunities offered by automatic arrhythmia classification, with the aim of making healthcare increasingly personalized and centered on patient needs.

The integration of advanced AI technologies and deep learning algorithms has significantly improved the ability to classify ECG signals, surpassing the limitations of traditional methods and opening new perspectives for the early diagnosis of arrhythmias and cardiac diseases. However, several open challenges remain and require further research. One of the main difficulties lies in the quality of the available data, often affected by artifacts or residual noise, despite advanced preprocessing techniques. The presence of these disturbances can negatively impact classification accuracy, particularly when it comes to detecting rare arrhythmias, which are frequently underrepresented in the databases used for training. The scarcity of data related to these conditions leads to unbalanced models that are difficult to generalize. Moreover, even though deep learning models have achieved impressive results, their often black-box nature makes it challenging to understand the decisions made. This is a particularly relevant issue in the clinical field, where trust in AI models depends on their interpretability.

2.2.2 PhysioNet Computing in Cardiology challenge 2020

The PhysioNet/Computing in Cardiology Challenge 2020 offer the opportunity to address issues related to automation in the detection and classification of cardiac anomalies by providing data from a wide range of sources with a broad set of cardiac anomalies. [3] The PhysioNet Challenge is an initiative that invited participants from academia, industry, and other sectors to solve clinically important issues that have significant clinical implications. As in previous years, the Challenge has an unofficial phase and an official phase that took place over several months. PhysioNet co-organizes the Challenge annually in collaboration with the Computing in Cardiology conference. The goal of the PhysioNet Challenge 2020 is to identify clinical diagnoses from 12-lead ECG recordings.

Participants are asked to design and implement a working open-source algorithm that, based solely on the provided clinical data, can automatically identify any of the 27 cardiac anomalies present in a 12-lead ECG recording. The winners of the Challenge are the teams whose algorithm achieved the highest score for the recordings in the hidden test set. A new scoring function is developed to evaluate the participants' algorithms, assigning partial credit to incorrect diagnoses that lead to treatments or outcomes similar to the true diagnosis, as some incorrect diagnoses are more harmful than others and should be evaluated accordingly.

Database	Total patients	Recordings in Training Set	Recordings in Validation Set	Recordings in Test Set	Total Recordings
CPSC	9458	10330	1463	1463	13256
INCART	32	74	0	0	74
PTB	19175	22353	0	0	22353
Georgia	15742	10344	5167	5167	20678
Undiscolsed	Unknown	0	0	10000	10000

Table 2.1. Numbers of patients and recordings in the training, validation, and test databases for the Challenge.

Data from five different sources are used, totaling 66,361 recordings, and it is possible for each patient to have one or more ECGs. Two sources are split to form training, validation, and test sets; two sources are included only as training data; and one source is included only as test data. The split is shown in the Tab 2.1. The training data and clinical diagnoses of the ECGs are made public, while the validation and test data are kept hidden. The training, validation, and test data are matched as closely as possible for age, sex, and diagnosis. The completely hidden dataset is never published, allowing for the evaluation of common machine learning issues such as overfitting. The test set as a whole includes data from the same sources as some training sets, as well as an entirely new set recorded from an institution geographically distinct from the training. Therefore, while there may be a small number of ECGs from patients present in both the training and test data, there is at least one test database where the likelihood of patients from the

training database being represented in the test data is extremely low.

Each 12-lead ECG recording is acquired in a hospital or clinical setting. The data acquisition specifications depend on the source of the databases, which are assembled worldwide and therefore vary. The quality of the labels depends on clinical or research practices and includes automatically generated labels, reviewed by a single cardiologist, and judged by multiple cardiologists.

Diagnosis	Code	Abbreviation
1st degree AV block	270492004	IAVB
Atrial fibrillation	164889003	AF
Atrial flutter	164890007	AFL
Bradycardia	426627000	Brady
Complete right bundle branch block	713427006	CRBBB
Incomplete right bundle branch block	713426002	IRBBB
Left anterior fascicular block	445118002	LAnFB
Left axis deviation	39732003	LAD
Left bundle branch block	164909002	LBBB
Low QRS voltages	251146004	LQRSV
Nonspecific intraventricular conduction disorder	698252002	NSIVCB
Pacing rhythm	10370003	PR
Premature atrial contraction	284470004	PAC
Premature ventricular contractions	427172004	PVC
Prolonged PR interval	164947007	LPR
Prolonged QT interval	111975006	LQT
Q wave abnormal	164917005	QAb
Right axis deviation	47665007	RAD
Right bundle branch block	59118001	RBBB
Sinus arrhythmia	427393009	SA
Sinus bradycardia	426177001	SB
Sinus rhythm	426783006	NSR
Sinus tachycardia	427084000	STach
Supraventricular premature beats	63593006	SVPB
T wave abnormal	164934002	TAb
T wave inversion	59931005	TInv
Ventricular premature beats	17338001	VPB

Figure 2.9. Diagnoses, Systematized Nomenclature of Medicine (SNOMED) codes and abbreviations in the posted training databases for the 27 diagnoses that were scored for the Challenge.

The training data contain 111 diagnoses or classes. Of these 111 diagnoses,

27 are used, shown in the Fig. 2.9, to evaluate the participants' algorithms because they are relatively common, of clinical interest, and more easily recognizable from ECG recordings. All data are provided in MATLAB and WFDB (Waveform Database) compatible formats. Each ECG recording has a MATLAB v4 binary file for the ECG signal data and an associated WFDB header text file describing the recording and patient attributes, including the diagnosis or diagnoses, i.e., the labels for the recording, making it a multi-label problem. For each 12-lead ECG recording, and not for each single lead, the algorithm must identify a set of one or more classes and a probability score of membership or confidence for each class.

For evaluation, each team's training code is run on the training data and then the trained code of each team is run on the hidden validation and test sets by the organizers, executed sequentially on the recordings to use them as realistically as possible. It is possible to submit code implementations in MATLAB or Python, while other languages, including Julia and R, were supported but received little interest from participants during the unofficial phase. Participants containeriz their code in Docker and submitted it to GitHub or Gitlab repositories. Virtual machines on Google Cloud are used for the runs, imposing a maximum time limit of 72 hours for training on the training set, and 24 hours for running the trained classifiers on the test set.

In total, the Challenge has 1395 algorithm submissions from 217 teams from academia and industry, while the total number of successful entries was 707, with 397 successful entries during the unofficial phase of the Challenge and 310 successful entries during the official phase. During the official phase, each entry is evaluated on the validation set. The final score and ranking are based on the test set. A total of 70 teams successfully run their code on the test data.

The highest scores are observed in the hidden CPSC datasets shown in Fig. 2.10, which contain a larger number of recordings in the training set compared to the other three hidden datasets. A drop in scores of about 50% is observed for the undisclosed hidden set, for which no recordings are included in the training or validation.

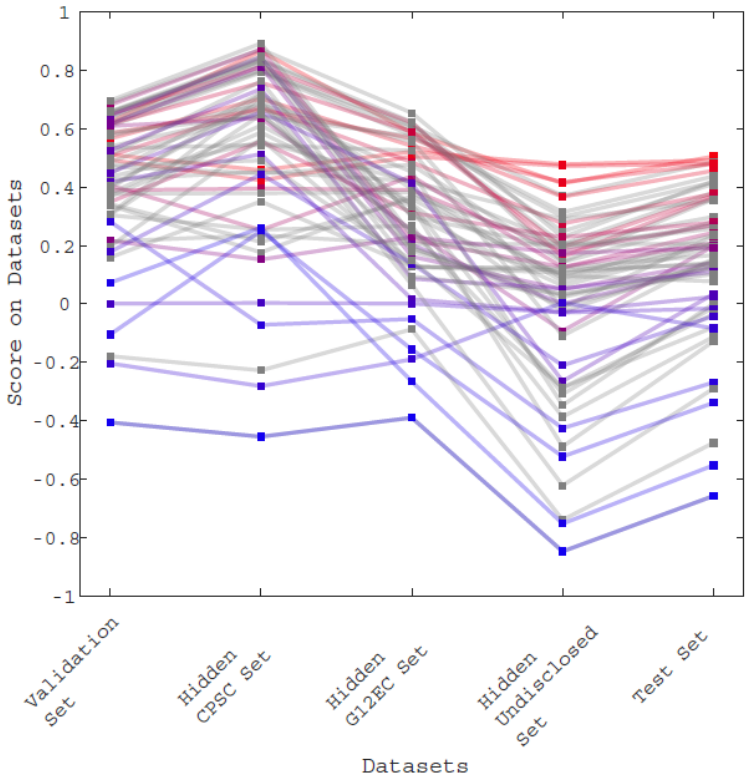

Figure 2.10. Scores of the final 70 algorithms that were able to completely evaluated on the validation set, the hidden CPSC set, the hidden Georgia (G12EC) set, the hidden undisclosed set, and the test set. The points indicate the score of each individual algorithm on each dataset, with the higher points showing algorithms with the highest scores on each dataset. The ranks on the test set are further indicated by color, with red indicating the best ranked algorithms and blue indicating the worst ranked algorithm on the test set. Source: [3]

Chapter 3

Materials and Methods

3.1 Instrumentation and working environment

During the course of the project, several software and hardware tools provided by company SynbrAIn were employed. In particular, on the software side, Docker was used, a container-based application virtualization technology, accessible via SSH connection once created. As for the hardware, graphics processing units (GPUs) hosted on a server provided by SynbrAIn were used.

Docker is an open-source platform designed to simplify the development, distribution, and management cycle of applications through containerization. The main functionality of Docker is to allow developers to package applications along with their dependencies in environments isolated from the operating system, known as containers. Containers are less resource-intensive compared to virtual machines, which are the other commonly adopted solution for isolating software components. This approach ensures consistency in the execution of applications across different development, testing, and production environments, thus simplifying the processes of creating, managing, and scaling microservices. [8]

In the context of this thesis, a Docker container was used to access Jupyter Notebook, providing a graphical interface of the server accessible via an Secure Shell(SSH) tunnel connected to the local computer. Jupyter Notebook is an open-source tool that allows the creation and sharing of documents containing executable code, equations, visualizations, and narrative text. Thanks to the isolation of containers, there was no need to worry about conflicting dependencies or libraries. To create a container, it is first necessary to define an image, which represents the base virtual environment of which the containers are specific instances. The instructions for defining an image are contained within a Dockerfile, which specifies the base image, the application source code, the dependencies listed in the *requirements.txt* file, and the configurations necessary for running the service.

In particular, the Dockerfile used had the following structure:

Listing 3.1. Dockerfile.

```
FROM python:3.9.12
WORKDIR /data

# Copy requirements file and install dependencies
COPY requirements.txt .
RUN pip install -r requirements.txt
RUN pip install jupyter

# Set non-interactive mode for Debian
ENV DEBIAN_FRONTEND=noninteractive

# Expose the Jupyter Notebook port
EXPOSE 8888


# Set the entry point for running Jupyter Notebook
ENTRYPOINT ["sh", "-c", "jupyter notebook --allow-root
--no-browser --ip=0.0.0.0"]
```

The Dockerfile used in this thesis is designed to configure a Docker image with Python, PyTorch, Jupyter Notebook, and some additional libraries, creating an optimized environment for running Python applications, with a particular focus on deep learning applications using PyTorch. The container configuration includes exposing port 8888 to allow remote access to Jupyter Notebook.

After uploading the Dockerfile to the server in the desired directory, simply access the server, navigate to the directory containing the Dockerfile, and execute the following commands to create the image:

Listing 3.2. Creation of container and connection creation.

```
# DOCKER IMAGE CREATION
docker build -t NAME .  # (NAME: Image name selected)

# CONTAINER CREATION
docker run -ti --gpus all -p xxxx:8888 --shm-size=16g
--ulimit memlock=-1 -d -v /yyyy:/data -e
JUPYTER_TOKEN="docker" --ulimit stack=6710886 zzzz
# (xxxx: port selected from docker ps command; /yyyy:
path to the project; zzzz: name of the container)

## CONNECTION CREATION
```

```
Open a new terminal:
-ssh -N -f -L localhost:xxxx:localhost:xxxx -p423 username@IP
(xxxx: port selected from docker ps command;
423: port used; username: user id;)

Open browser:
- localhost:xxxx (xxxx: port selected from docker ps command)
```

Once these commands have been executed and the container has been created and the connection through the SSH tunnel has been established, you can open a browser and access http://localhost:xxxx/ to view Jupyter Notebook, thus obtaining a graphical interface of the server. It is important to note that through these instructions, it is possible to generate the Dockerfile image by assigning a predefined token, activating any GPUs present on the server, and creating the necessary volumes to access the server's data. The server used for the project is equipped with two NVIDIA graphics processing units: a GeForce RTX 3090 and a GeForce RTX 2080 Ti. The GeForce RTX 3090 has 24 GB (24576 MiB) of memory, while the GeForce RTX 2080 Ti has 11 GB (11264 MiB) of memory. The greater memory of the GeForce RTX 3090 offers several significant advantages. Firstly, it allows for the handling of larger datasets directly on the GPU, reducing the need for frequent transfers between the central processing unit (CPU) and GPU, which can slow down the processing. Additionally, it enables the training of more complex machine learning models and deep neural networks, which require more memory to store weights and activations during training. Finally, a GPU with more memory can handle more processes simultaneously, improving the overall efficiency and productivity of the system.

To monitor the memory usage of the GPUs, the command `nvidia-smi` is used. This command provides detailed information about the NVIDIA GPUs installed in the system, including the total available memory, the currently used memory, and the processes that are using the GPU memory. Thanks to `nvidia-smi`, it is possible to obtain a complete overview of the GPU resources in use and optimize the system's performance according to the specific needs of the project.

3.2 Dataset

3.2.1 Training set

The dataset used is provided by the organizers of the Cardiology Challenge 2020 [3] and consists of various sources from around the world. In particular, the available databases are:

- **CPSC Database and CPSC-Extra Database**: Derived from the China Physiological Signal Challenge 2018. These data, from 9458 different patients, include ECG recordings of varying durations between 6 and 60 seconds, sampled at 500 Hz. The CPSC consists of 6877 recordings, with an average duration of 15.9 seconds; while the CPSC-Extra consists of 3453 recordings, with an average duration of 15.9 seconds, for a total of 10330 recordings. [32]

- **St Petersburg INCART Database**: This database consists of 75 annotated recordings extracted from 32 Holter records. Each record is 30 minutes long, each sampled at 257 Hz.

- **Physikalisch Technische Bundesanstalt (PTB)**: Includes two public databases: the PTB Diagnostic ECG Database and PTB-XL. The former contains 516 recordings sampled at 1000 Hz with an average duration of 110.8 seconds, while the PTB-XL includes 21,837 clinical ECGs of 10 seconds, sampled at 500 Hz with an average duration of 10 seconds. [46]

- **Georgia 12-lead ECG Challenge (G12EC) Database**: Represents a unique demographic of the southeastern United States and contains 10,344 ECGs of 10 seconds, sampled at 500 Hz.

The combined datasets account for a total of 43101 recordings in the training data, the specifications of which are shown in Tab 3.1. All data are provided in WFDB

Dataset	Number of recordings	Mean Duration (seconds)	Mean Age (years)	Sex (Male/Female)
CPSC	6877	15.9	60.2	54%/46%
CPSC-Extra	3453	15.9	63.7	53%/46%
INCART	72	1800.0	56.0	54%/46%
PTB	516	110.8	56.3	73%/27%
PTB-XL	21837	10.0	59.8	52%/48%
Georgia	10344	10.0	60.5	54%/46%

Table 3.1. Number of recordings, mean duration of recordings, mean age of patients in recordings, sex of patients in recordings, and sample frequency of recordings for each dataset.

format, with MATLAB v4 binary files for ECG signals and WFDB header text files that describe the signal metadata, including essential information such as the number of samples, sampling frequency, channel names, units of measurement, and calibration parameters, including the diagnosis. An example of a header file is provided in Fig. 3.1. The WFDB format is a set of standards for reading and storing physiological signal data and their annotations. [1] These files can store signals from one or more channels, interlaced in a single file for multichannel recordings. In total, the labels associated with the entire dataset amount to 111.

```
A0001 12 500 7500
A0001.mat 16x1+24 1000.0(0)/mV 16 0 28 -1716 0 I
A0001.mat 16x1+24 1000.0(0)/mV 16 0 7 2029 0 II
A0001.mat 16x1+24 1000.0(0)/mV 16 0 -21 3745 0 III
A0001.mat 16x1+24 1000.0(0)/mV 16 0 -17 3680 0 aVR
A0001.mat 16x1+24 1000.0(0)/mV 16 0 24 -2664 0 aVL
A0001.mat 16x1+24 1000.0(0)/mV 16 0 -7 -1499 0 aVF
A0001.mat 16x1+24 1000.0(0)/mV 16 0 -290 390 0 V1
A0001.mat 16x1+24 1000.0(0)/mV 16 0 -204 157 0 V2
A0001.mat 16x1+24 1000.0(0)/mV 16 0 -96 -2555 0 V3
A0001.mat 16x1+24 1000.0(0)/mV 16 0 -112 49 0 V4
A0001.mat 16x1+24 1000.0(0)/mV 16 0 -596 -321 0 V5
A0001.mat 16x1+24 1000.0(0)/mV 16 0 -16 -3112 0 V6
# Age: 74
# Sex: Male
# Dx: 59118001
# Rx: Unknown
# Hx: Unknown
# Sx: Unknown
```

Figure 3.1. Example of a WFDB header file for a 12-lead ECG recording from CPSC database. The first line specifies the record name, the number of leads, the sampling frequency, and the number of samples. The subsequent lines describe the details of each signal, including the file name, sample format, sampling frequency, and gain. The comment lines provide additional information about the patient and the diagnosis.

For this reason, all signals whose class was not among the 27 expected described in Fig. 2.9 were excluded. Furthermore, the INCART data was excluded since it has only 74 30-minute records with a sampling frequency of 257 Hz and is significantly different from other datasets. From these data, the training and validation datasets were created. This division was maintained as proposed by the group to identify the model to work on in order to maintain a final comparison on the metrics obtained as described in Section 3.5. After excluding the recordings belonging to the INCART database and those whose labels were not among the 27 expected, the total number of recordings is 37716, which are subsequently randomly divided into 80 % as the training set, amounting to 30172 recordings, and 20 % as the validation set, amounting to 7544 recordings. [52]

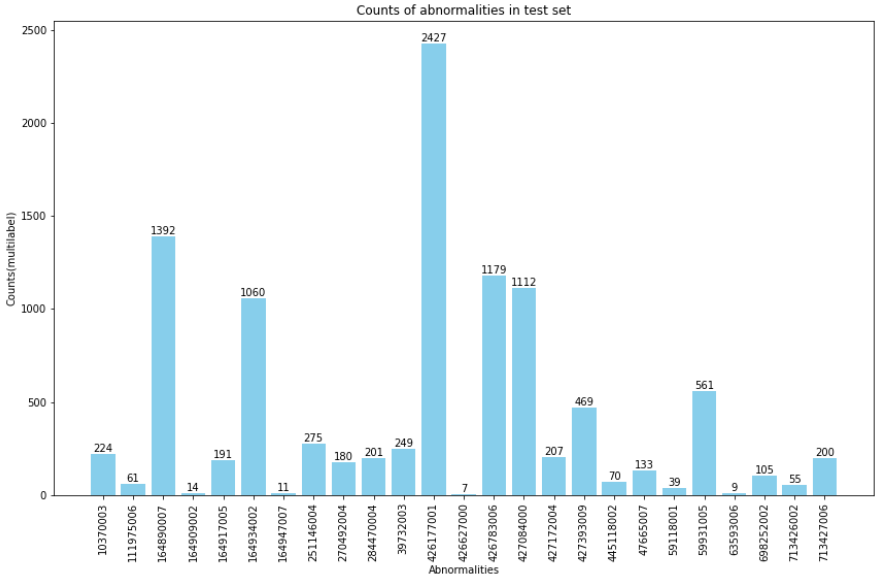

Figure 3.2. Abnormalities distribution in the test set.

3.2.2 Additional test set

To further validate the model's generalizability, an external dataset from the PhysioNet/Computing in Cardiology Challenge 2021 was applied as external validation: the Ningbo dataset. [41] This dataset is a large database of 12-lead electrocardiograms created for studies on arrhythmia and other cardiovascular conditions. This database was developed under the auspices of Chapman University, Shaoxing People's Hospital (Shaoxing Hospital Zhejiang University School of Medicine), and Ningbo First Hospital. The dataset contains 45,152 ECGs from 34,905 patients, with a sampling frequency of 500 Hz and a length of 10 seconds. All signals are provided in .mat format, and each is associated with a .header file containing metadata and technical characteristics. The ECGs include various common rhythms and other cardiovascular conditions, all labeled by expert professionals. The initial labels contain almost all 27 labels expected by the challenge, except for atrial fibrillation and ventricular premature beats. Specifically, there are 7,615 elements labeled with atrial flutter; however, after visualization, several of these do not present the characteristic 'sawtooth' pattern, indicating a possible phase of inaccurate labeling. The dataset presents a diverse demographic distribution: the average age of patients is 57.7 years, with 43 % women, 56 % men, and 1 % of patients without specified gender. This demographic diversity ensures that the dataset represents a wide range of patients, thus improving the robustness and generalizability of models developed using these data. Among all records, 6500

with labels in 25 types of ECG anomalies were randomly selected to form an external validation set, ensuring that the number of elements for each class was at least equal to the minimum value among all expected classes. In this case, the initial distribution counted only 7 elements labeled with bradycardia; for this reason, there are at least 7 elements for each of the 25 classes, while the others were randomly extracted until reaching the 6500 selected elements. The extraction of these elements led to a new average age value of 58.0 years, while the gender distribution was 56.4 % male and 43.6 % female, with a distribution shown in Fig. 3.2.

3.3 Data analysis

The preliminary data analysis examined the two available subsets (training and validation) with the aim of analyzing their distributions in terms of available covariates, signal lengths, and associated anomalies. In particular, the first analysis focused on the distribution of the gender variable in terms of percentage and the distribution of age divided into groups encompassing different age values. The results are shown in Fig. 3.3 and Fig. 3.4.

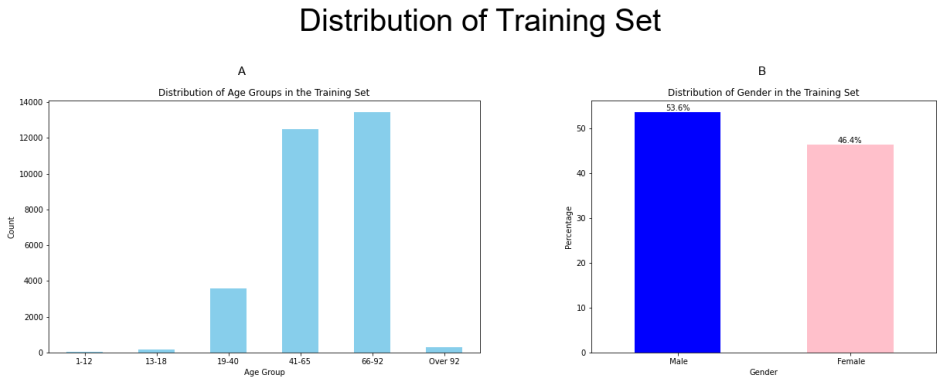

Figure 3.3. Demographic Analysis: age and gender distribution in training set. (A) Distribution of age groups in the training set, with age groups '1-12' '13-18', '19-40', '41-65, '66-92', and 'Over 92'. (B) Distribution of gender in the training set, with categories 'Male' and 'Female' represented as percentages.

As shown in the graphs, these two subsets are characterized by a modest number of elements with an age over 92 years. The analysis revealed that the PTB-XL dataset, one of the largest considering the entire dataset, contains a significant number of signals associated with an age of 300 years. The article explains that

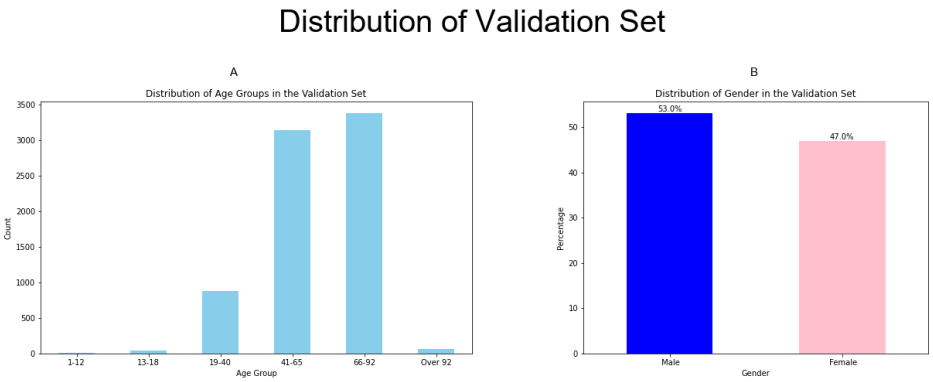

Figure 3.4. Demographic Analysis: age and gender distribution in validation set. (A) Distribution of age groups in the training set, with age groups '1-12' '13-18', '19-40', '41-65, '66-92', and 'Over 92'. (B) Distribution of gender in the training set, with categories 'Male' and 'Female' represented as percentages.

for patients with ECGs recorded at 90 years or older, the age is set to 300 years to comply with the standards of the Health Insurance Portability and Accountability Act (HIPAA). [46] For this reason, after representing the top 200 highest age values

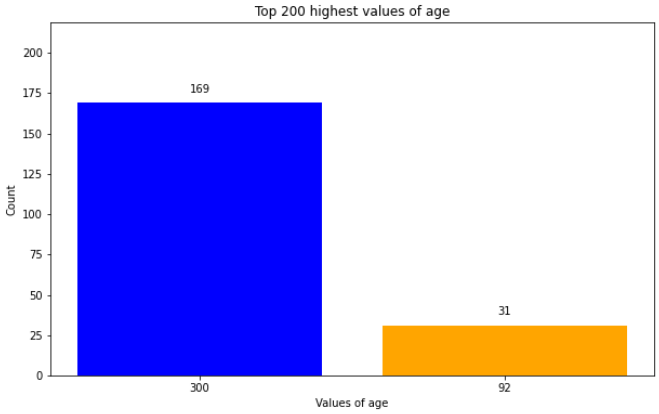

Figure 3.5. Graphical representation of the 200 highest age values.

in this dataset using bar charts in Fig. 3.5, it was decided to replace the value 300 with the first valid value greater than 90. This way, a reasonable age value is set, avoiding the exclusion of these elements from the dataset while respecting

the guidelines of the dataset creators. The analysis revealed that the first value greater than 90 within the dataset is 92. The decision was also made because, if the actual age were 98 years, replacing this value with 92 would still represent an acceptable and representative alteration of the signal. Finally, to get an idea of the representativeness of the individual classes in the entire dataset, bar charts were used to represent the classes and their total count in the validation and training sets, as shown in Fig. 3.6 and Fig. 3.7. As can be seen from the graph, the classes

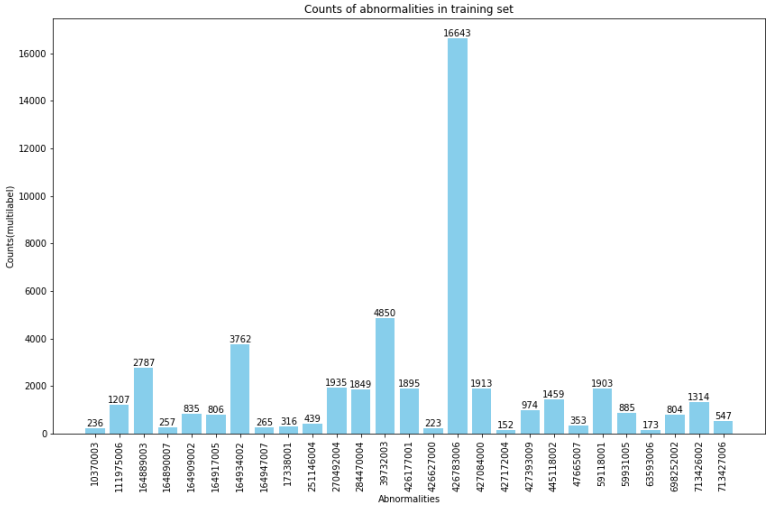

Figure 3.6. Abnormalities distribution in the training set.

have different sizes, further highlighting how the dataset itself is skewed towards the normal class, a factor to consider during the development of the classification model.

3.3.1 Pre-processing

The preprocessing phase initially involves the exclusion of all signals that are not associated with one of the 27 labels considered for the challenge. Subsequently, the sampling frequency of all training data consistent was set to 500 Hz to make all the samples consistent, and thus the PTB dataset was mainly downsampled.

Starting from the 12-lead electrocardiographic signal, shown in Fig. 3.8, a series of preprocessing operations are performed before being passed as input to the model.

The signals from leads III, aVR, aVL, and aVF are excluded from the model input, which brings the number of used leads to 8. This is because these four leads are linearly dependent on the others and can be calculated based on Einthoven's

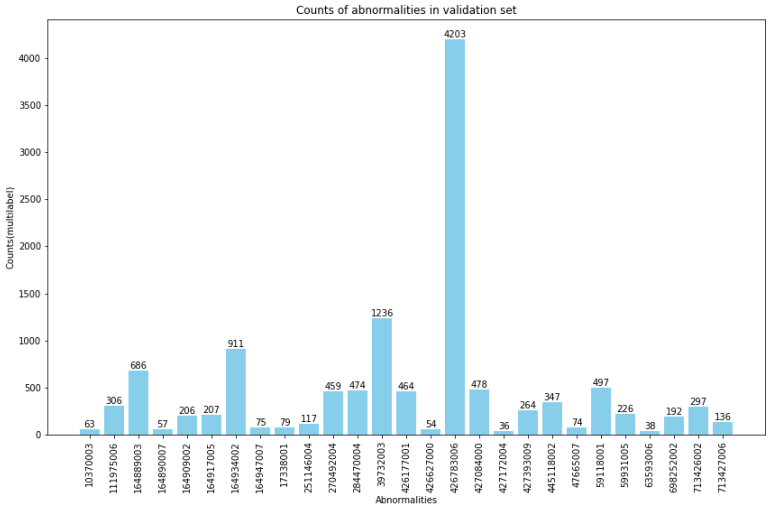

Figure 3.7. Abnormalities distribution in the validation set.

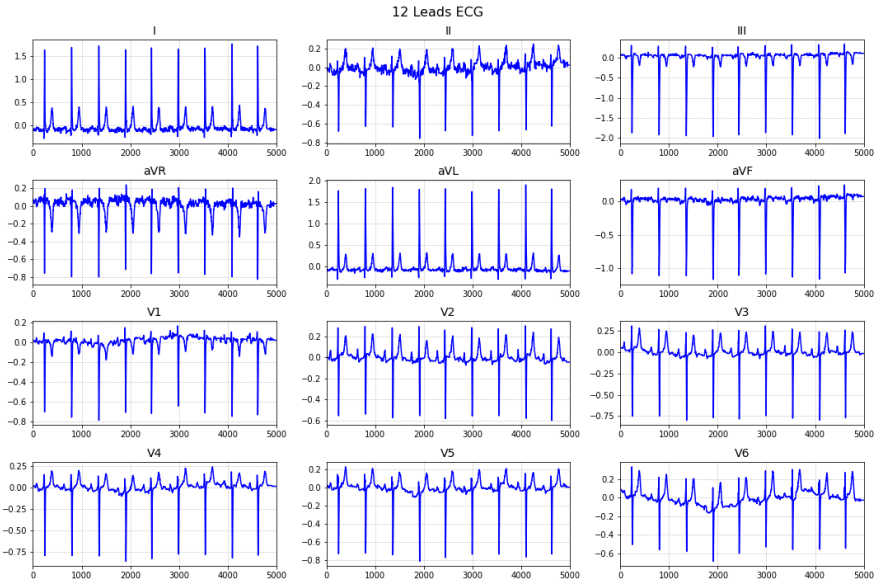

Figure 3.8. Example of a 12-lead electrocardiographic signal from the PTB-XL database of an 82-year-old woman, labeled with the classes Sinus rhythm, 1st degree AV block, Left anterior fascicular block, and Left axis deviation.

Law [29] and Goldberger's equations [5].

For the initial model considered, especially the single Squeeze and Excitation Residual Network (SE-ResNet), all input signals were fixed at 30 s in length. This

was done by truncating the part exceeding the first 30 s for longer signals and padding shorter signals with zeros. For the other ensemble model, to reduce the effect of padding on shorter signals, the input length was set to 10 s with the same preprocessing method. The idea of using two different lengths arises from the observation that some alterations in an ECG signal can occur throughout the signal duration, such as alterations affecting the heart rhythm and thus having a repeated shape, while others are related to a single beat. For example, arrhythmias like atrial fibrillation can manifest continuously and affect the entire 30 s segment, making an irregular repeated pattern evident.[18] Conversely, anomalies like an abnormal QRS complex can be isolated to a single beat and not frequently repeated. [47] Therefore, using different signal lengths allows capturing both long-term and short-term characteristics of the ECG signal, thus improving the model's ability to detect a variety of heart conditions. Once the leads to be used are identified, a biorthogonal wavelet transformation (bior2.6) is applied to reduce the noise in ECG signals. The numbers of vanishing moments for the decomposition and reconstruction filters were 2 and 6, respectively. The level of refinement was set to be 8, where the high-frequency coefficients in level 1, level 2, and level 8 were set to zero. Although the article [52] does not provide specific reasons for the choice of these parameters, it is likely that they were chosen empirically. Often, parameters like the number of vanishing moments and the level of refinement are chosen based on the specific characteristics of the ECG signals being analyzed. Researchers might experiment with different settings to find the optimal configuration that minimizes noise and preserves important signal features. [40]

The biorthogonal wavelet transformation is particularly useful for denoising ECG signals due to its ability to maintain good signal reconstruction quality while effectively eliminating noise. The biorthogonal wavelet allows for signal decomposition into components that capture both local and global characteristics, making it ideal for analyzing non-stationary signals like ECGs. Using the biorthogonal wavelet has several advantages. Firstly, it allows for a multi-resolution representation of the signal, meaning that the signal can be analyzed at different scales or levels of detail.

This is crucial for ECG signals, where significant events can occur at various frequencies and durations. At the end of the operations, the signals will be as shown in Fig. 3.9.

3.4 Experimental procedures

The workflow began with the identification and reproduction of one of the models proposed by the top five groups ranked in the PhysioNet Computing in Cardiology Challenge 2020. Subsequently, modifications were made using the initial model as

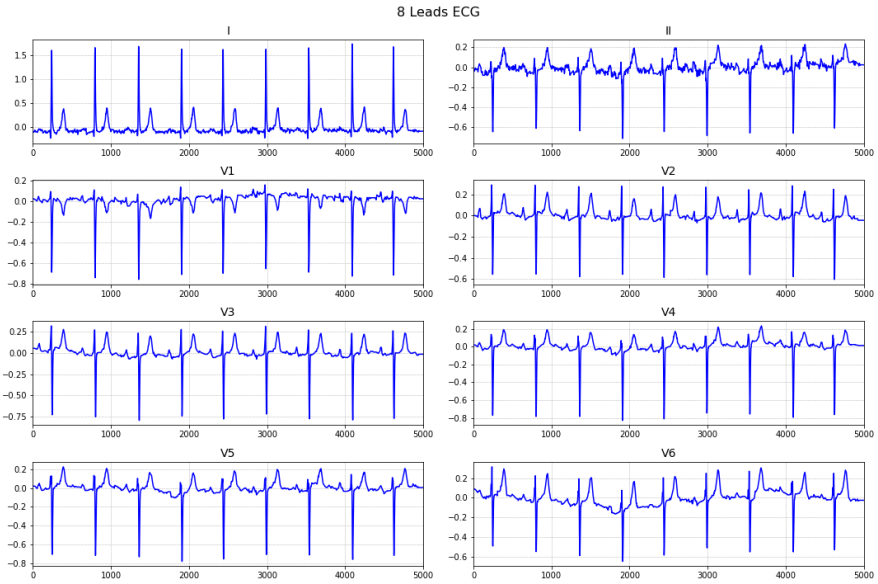

Figure 3.9. Example of a 8-lead electrocardiographic signal from the PTB-XL database of an 82-year-old woman obtained after steps of pre processing, labeled with the classes Sinus rhythm, 1st degree AV block, Left anterior fascicular block, and Left axis deviation.

a reference to improve performance and increase generalizability to new data. For each modification made to the model, a structured workflow was followed, which includes the following phases: integration of the modification into the original model, training of the new model, performance evaluation, comparison with the baseline, and retention of the modification only if the performance was improved compared to the previous model. Otherwise, the previous model was kept. This iterative approach ensured that each change made effectively contributed to the overall performance improvement. The workflow is shown in Fig. 3.10.

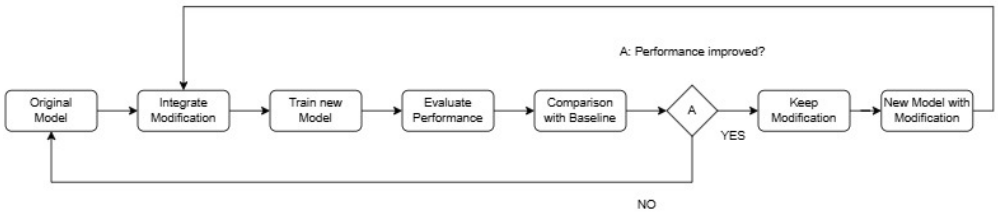

Figure 3.10. Example of workflow followed to integrate modifications to the model.

3.4.1 Choice of the model to improve

The first phase of the thesis work focused on analyzing the solutions proposed by the groups ranked in the top five positions at the Cardiology Challenge 2020 [53, 35, 21, 37, 51]. In addition to comparing the results obtained on the hidden test set and hidden validation set, the analysis focused on identifying the pipelines proposed by the different groups that had good performance on their own training set and validation set. Despite the non-uniform data division, the choice was made with the primary goal of obtaining a model capable of performing well on the training data, but at the same time not showing a significant loss of performance in the validation set, as this would indicate poor generalizability of the proposed model. The results were then compared, and subsequently, the model to be used and modified for the subsequent training was identified.

3.4.2 Training setup

All models were trained with a batch size of 16 for 21 epochs (except for the reproduction of the first ensemble model), while the parameters were optimized with the Adam optimizer. During training, the learning rate was set to 0.001 and reduced by a factor of 10 at the 13th epoch. The loss function used in all training sessions was created by the group and is called multi-label SignLoss. This function combines elements of Binary Cross Entropy Loss (BCE Loss) with additional penalties based on the difference between the model's predictions and the true labels. It defines as follow:

$$\text{Sign}(p) = \begin{cases} (y^2 - 2py + p^2) & \text{if } |y - p| < 0.5 \\ 1 & \text{if } |y - p| \geq 0.5 \end{cases} \tag{3.1}$$

$$\text{Loss} = \sum_{i=1}^{27} \text{Sign}(p_i) \cdot \text{BinaryCrossEntropyLoss}(p_i, y_i) \tag{3.2}$$

where y denotes the ground truth and p denotes the model's estimated probability for $y = 1$.
The difference, which indicates how much the predictions deviate from the true labels, is compared to a threshold value of 0.5. If this is greater, the applied loss function is the standard BCE, which penalizes the model for each error in the predicted probabilities. If it is less than or equal to 0.5, the loss function combines BCE Loss with a quadratic penalty, which accounts for the quadratic distance between the predicted probabilities and the true labels. Finally, the function sums all these penalties for each element of the batch and calculates the average over the entire batch. In this way, the SignLoss function not only measures the accuracy of the model's predictions but also applies additional penalties to further improve

precision. This is done, as highlighted by the authors, to address the problem of class imbalance. [52]

3.4.3 Classification model baseline

The proposed model aims to address this problem by combining the structure of a residual network (ResNet) with Squeeze and Excitation (SE) modules, which allow the model to capture the relative importance of each lead from multi-lead ECG signals. Specifically, the structure used is ResNet-34, a variant of the ResNet family with a depth of 34 layers for processing one-dimensional signals, such as time series. The model takes one-dimensional signals as input and processes them with a series of convolutional layers, batch normalization, and ReLU activation functions, followed by pooling operations to reduce the spatial dimensions of the data.

The structure consists of four main levels, each containing a specific number of blocks (SEBasicBlock) for a total of sixteen, and as the level increases, the number of convolutional filters increases to capture increasingly complex features of the input sequence. The SE layers work by compressing the signal information into a compact representation for each channel, called the squeeze phase, followed by an excitation phase where the information generated in the previous phase is used to capture dependencies between channels, and thus produce a weight for each of them, applied in the feature maps of the previous levels to recalibrate the importance of each channel of the original input, multiplying each channel by its corresponding weight. [24]
At the end of these operations, there is an adaptive pooling operation to reduce the signal dimensionality to a compact representation, which is then passed through a fully connected layer that transforms these features into a prediction for each class. Fig. 3.11 shows the architecture of this model in the case in which signals of 15000 samples are considered as input.

The use of adaptive pooling allows the model to handle signals of variable length by adapting them to a specific size regardless of the original input size, which is particularly useful in scenarios where input sizes may vary, without resizing to a fixed size and maintaining relevant information. This aspect is particularly important because the final proposed model is an ensemble of two models that handle input signals of different temporal lengths: the first takes input signals of 30 seconds (Model A), while the second, to reduce the effect of padding on shorter signals, handles input sequences of 10 seconds (Model B).
The ensemble phase considers the predictions generated by the two models, to which a sigmoid function is applied to generate probabilities. These probabilities of the signal belonging to individual classes are combined using a weight of 0.5 for each model, which implies that both predictions contribute equally to the final

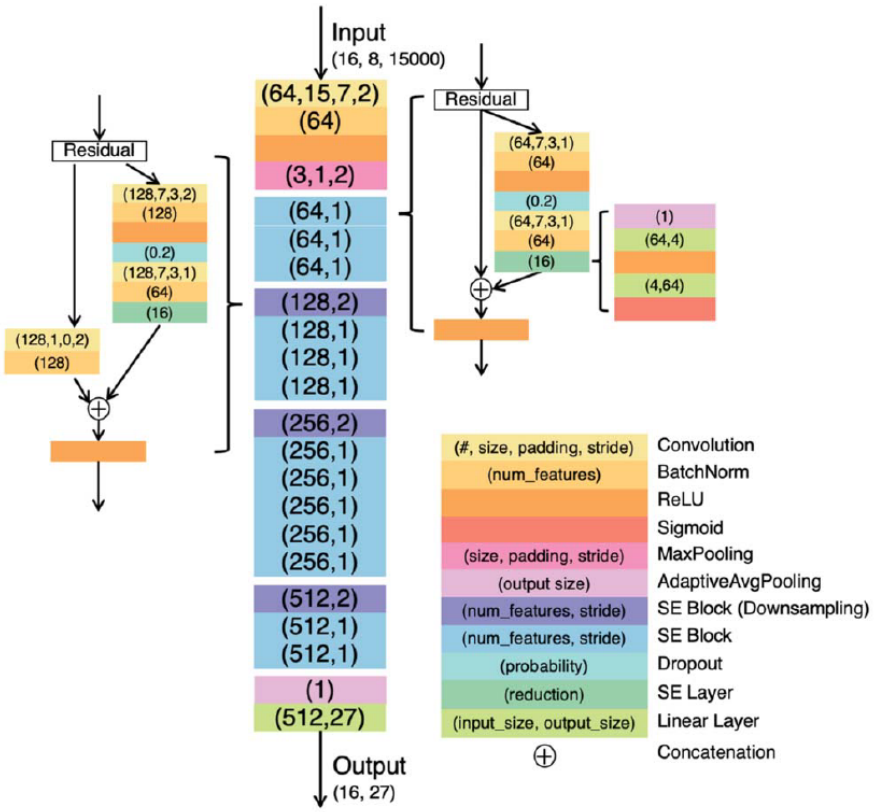

Figure 3.11. Architecture of the SE-ResNet model.

prediction. The final predictions are assigned to the signals if the probability value for a specific class exceeds the threshold value set at 0.36, obtained by the group on the validation set by varying different threshold values in steps of 0.01. The choice to merge the two outputs was therefore used to improve generalizability, while the use of two different lengths allows for greater information when the signal is longer than 10 seconds and is truncated to ensure the correct input to model B. Additionally, it provides a multiple view of the same data to the model.
Following the ensemble phase of the two models, two rules are applied to correct potential errors due to incorrect model behavior. Due to low performance by the models on the bradycardia class because of low representativeness, a rule-based model is applied in cascade to the ensemble for the recognition of the alteration; since this is a characteristic variation of the heart rhythm, the implementation of the Pan and Tompkins algorithm [44] allows for the identification of R peaks considering lead I and subsequently identifying the RR intervals to evaluate their duration and compare them with fixed values. Specifically, the model identifies if

the duration of the RR intervals is between 1 second and 1.6 seconds, then the specific interval can be defined as bradycardic. Consequently, if at least half of the identified intervals in the total analyzed intervals have a duration within that range, and if there are at least 6 RR intervals analyzed, then the model determines that the signal belongs to the bradycardia class. The purpose of this rule-based model is to validate or reject the prediction generated by the network. Therefore, it is called at the end of the ensemble phase, and if the output is true, then the bradycardia class is validated or added; otherwise, the network's prediction is set to 0 if it was initially set to 1 for the bradycardia class.
To avoid signals without a class, a check is implemented to verify the belonging of each signal to at least one of the classes. If this is not true, then the prediction is set to 1 for the normal class, as it is the most represented class. This is done to avoid invalidating the inference phase and the absence of predictions for the signals.
Fig. 3.12 represents the design of the entire proposed model.

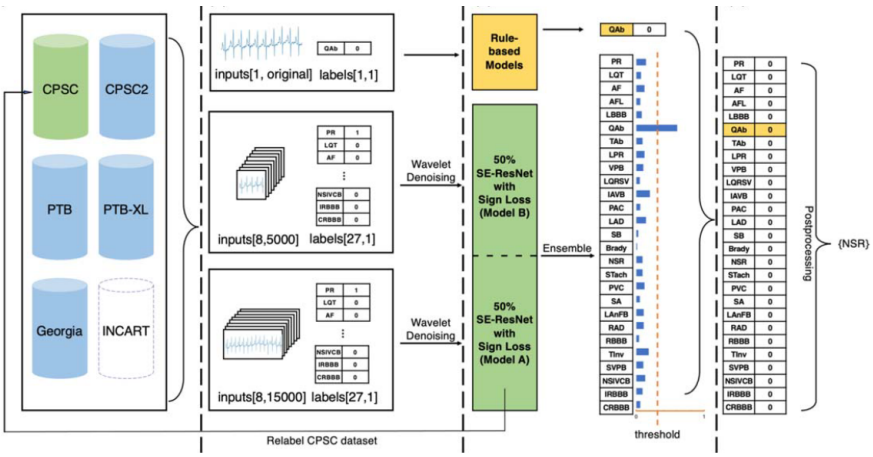

Figure 3.12. Design of the proposed model.

Starting from the codes made available by the group, an attempt was made to reproduce the proposed model. Regarding the training phase, a fine-tuning of model A was carried out using shorter input signals by considering the model B previous described. Model A was trained for a total of 18 epochs, and subsequently, starting from the weight vectors of the best epoch of model A, fine-tuning of model B was carried out for a total of 3 epochs, as stated in the article published by the group. [52] The fine-tuning phase aimed to exploit the knowledge already learned by the model during the initial training and readapt it to the new dataset with a shorter length.

The weight vectors of model A were made available and were therefore used

directly to complete the training phase. Additionally, the first model was validated by clinicians after performing inference for the CPSC dataset, and it was not possible to reproduce the model in its entirety. To further verify the reproducibility of the model, the previous model was decomposed into two separate branches, thus considering two separate and independent models. For this reason, to verify the declared metrics with those obtained, inference was performed with the weight vectors of the best epoch declared in the group's training code and made available. The workflow, therefore, only included the complete reproduction of model B, keeping the hyperparameters set by the group for the training phase unchanged, in order to verify the results obtained and compare them with those declared in the papers published by the group during the submission phase. Model B receives 10-second long 8-lead signals as input. Two successive training sessions were conducted to ensure reproducibility.

3.4.4 Modification of the baseline model

After a careful comparative analysis between the model used and the models classified in the first and second positions in the challenge respectively [35, 51], some differences emerged, including signal normalization, concatenation of covariates (such as gender and age), the use of 12 leads or the use of a single lead, and the inclusion of some temporal features characterizing the signals. The study examined these modifications sequentially and compared the results obtained with those obtained with the reproduction of model B. The first operation integrated into the signal preprocessing was normalization using the min-max scaling technique:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3.3}$$

where x is the original value and x' is the normalized value.
The purpose of data normalization was to reduce variability among the data, making them comparable on a common scale, and it was carried out by considering each lead individually and their respective maximum and minimum reference signal values.
Similarly, training was carried out considering all 12 leads, without excluding any of them. This choice was made with the aim of increasing the extractable information and providing a more accurate and complete view of the heart's electrical activity, so that some alterations that are not very visible and poorly represented with a smaller number of leads become more evident. This could result in a loss of information and the possibility of generating incorrect diagnoses. Infact, the information contained in 12-lead ECG signals differs depending on the lead considered in terms of intensity and amplitude values.

In parallel with the training considering the 12 leads, some clinical reports [9, 25] have shown that lead II of the electrocardiographic signal is the most used in hospitals for a quick and accurate diagnosis. For this reason, in addition to using 12 leads, it was decided to carry out training considering exclusively this lead, with the idea of verifying whether some alterations would become more evident. These training sessions, varying the number of leads used as input, were carried out with the aim of finding the best configuration to provide the model with the most information and ensure a more refined classification. At the same time, they aimed to highlight substantial differences in the model's recognition of individual alterations, as some of these concern a single heartbeat, and therefore it is possible that a single lead allows the model to be more selective.

3.4.5 Addition of features to the baseline model

In addition to the modifications related to the derivations considered as input signals, the covariates gender and age, obtained from the respective header files and appropriately encoded, were added to improve performance. This was done because some ECG signal alterations are more evident at certain ages and there are significant differences between genders. For example, the risk of developing certain arrhythmias, such as atrial fibrillation, increases with age. Additionally, it is known that some electrocardiographic characteristics, such as the QT interval, can vary between men and women, with women generally having a longer QT interval. [3] These physiological differences can influence the interpretation of ECG signals and the accuracy of the diagnosis. Therefore, including the covariates of gender and age allows for the creation of more precise and personalized models, improving the ability to detect and analyze rhythm alterations in both pathological contexts and subjects with normal rhythm.

The addition was done in steps: initially, three variables related to the presence or absence of gender in the header file were added, and subsequently, age was integrated with appropriate encoding, thus creating a vector of 5 variables to evaluate the presence or absence of this variable. Regarding the gender variable, it was encoded using the one-hot encoding technique. In this approach, the categorical variable is transformed into a binary vector, where each category is represented by a single active position in the vector. Specifically, gender was encoded by creating a three-element vector:

- The first element of the vector is set to 1 if the gender is female.
- The second element of the vector is set to 1 if the gender is male.
- The third element of the vector is set to 1 if the gender is absent or unspecified.

This representation allows machine learning models to treat gender as a numerical variable without implying any order or hierarchy among the categories, thus preserving the integrity of the categorical information. Following the creation of the binary vector for gender, two elements were added to integrate the age information using encoding techniques that enhance the model's ability to interpret this information. Specifically, age was encoded in two distinct ways to evaluate performance differences.

In the first approach, age is normalized relative to the maximum value as described in Section 3.3. This normalization transforms age into a continuous variable ranging from 0 to 1, preserving the relative scale and facilitating integration with the model's other numerical features.
In the second approach, age is divided into predefined groups to represent specific age ranges, each represented by a distinct numerical code. In this case, the age ranges could include categories such as children (1-12 years), adolescents (13-18 years), young adults (19-40 years), adults (41-65 years), and seniors (66 years and older). This encoding allows reflecting differences between specific age groups, which may have distinct impacts on ECG signal characteristics. [38, 6]
In cases where age is missing or invalid (e.g., if it is negative or zero), the second element of the vector is set to 1 to indicate the error or absence of data, while the first element remains 0. This handling of missing data is applied in both encoding approaches. The additional features thus obtained are concatenated with the main features extracted by the model before performing the classification task. This integration enriches the available information, improving the model's ability to detect and analyze the correlations between the covariates and the ECG signal alterations that might not be fully captured by the convolutional layers, as the header file is not provided as direct input. For the processing of this five-element vector, two different components were designed as described in Fig. 3.13.

A. In the first case, the additional features are passed through a first linear layer that increases their dimensionality to 64, followed by a ReLU activation function, and then reduced to 16 dimensions by a second linear layer, also followed by a ReLU activation function.

B. In the second case, the additional features are passed through a single linear layer that transforms them into a 10-dimensional vector.

Two different processing units were tested to extract additional information from these covariates with two different final dimensions to understand how these affect the model's performance. The main difference between the two additional layers lies in the depth and complexity of the transformations of the additional features. The final choice will therefore depend on the results obtained during the validation phase, taking into account the complexity of the additional features

and computational efficiency. The design choices for the linear and ReLU layers, as well as the output dimensions, were inspired by the approaches used by other groups working on similar problems and participating in the PhysioNet Challenge in 2020 and 2021. By reviewing their methodologies, strategies that have been effective in processing similar types of features were adopted. This comparative analysis guided the decisions, although it did not follow a strict scientific criterion.

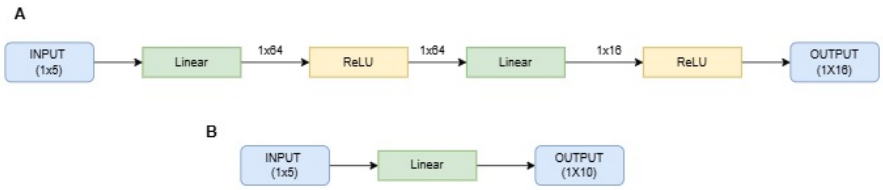

Figure 3.13. Description of the two additional layers that process the covariates. **A.** Layer that generates 16 additional features. **B.** Layer that generates 10 additional features.

In addition to the covariates obtainable from the header files of individual 12-lead ECG signals, an experiment was conducted by concatenating seven features derived from the temporal and nonlinear domains to the output features from the model.

After performing the preprocessing described in Subsection 3.3.1 lead II of each signal is isolated, as it is most commonly used in clinical settings. On this lead, operations are performed using specific processing tools to assess Heart Rate Variability (HRV). [43] HRV is considered a measure of neurocardiac function that reflects heart-brain interactions and the dynamics of the autonomic nervous system. It is, therefore, an index of the heart's ability to respond to physiological and psychological variations within the organism and the functional status of the sympathetic and parasympathetic systems.

The main feature extracted is the R peaks, which characterize an electrocardiographic signal. From these peaks, RR intervals are calculated by subtracting consecutive R peaks identified earlier. Once the intervals are extracted, their mean value in seconds is calculated by dividing the number of samples by the sampling frequency, and the standard deviation is also calculated as a measure of variability.

$$\text{Mean RR} = \frac{\sum_{i=1}^{N} RR_i}{N} \tag{3.4}$$

$$\text{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{N} (RR_i - \text{Mean RR})^2}{N}} \tag{3.5}$$

where RR_i represents an interval in samples, and N is the total number of intervals identified on each individual signal.

From the RR intervals, nonlinear characteristics are also extracted. In this case, nonlinearity is discussed because the relationship between the variables cannot be plotted as a straight line. These metrics allow for the quantification of the unpredictability of a time series of intervals. For example, a random series of RR intervals, a normal series, and a totally periodic series can have the same standard deviation value, but their underlying organization could be completely different.

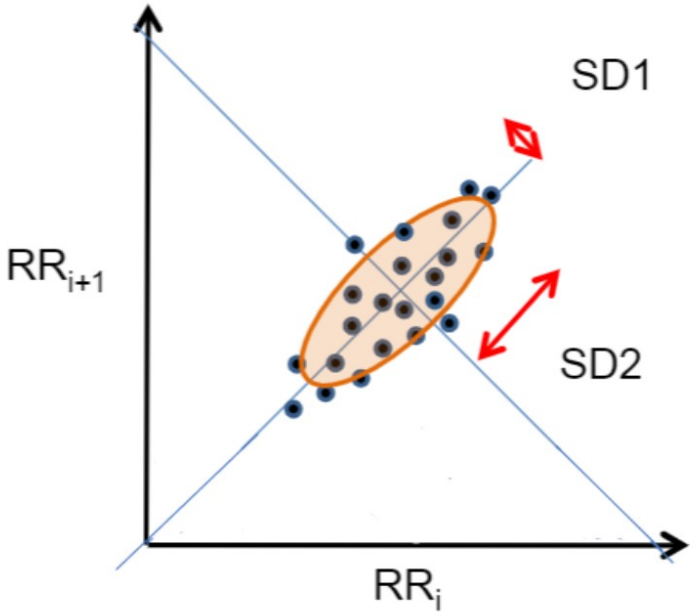

Figure 3.14. Example of Poincaré plot.

Using the RR intervals, the Poincaré plot is derived, where each RR interval is represented as a function of its previous interval. [36] This is a scatter plot that allows for the quantification of short-term and long-term variability by measuring the lengths of the two axes.

From the Fig. 3.14, the values of SD1, represented by the short axis, which measures short-term HRV in milliseconds, and the values of SD2, represented by the long axis of the ellipse, which measures long-term HRV in milliseconds, are extracted. From these values, for each signal, the SD1/SD2 ratio is calculated, which measures the unpredictability of the time series of RR intervals.

$$\mathrm{SD1} = \frac{1}{\sqrt{2}} \times \mathrm{SD}_{\Delta} \tag{3.6}$$

$$\mathrm{SD2} = \sqrt{2 \times \mathrm{SD}_{\mathrm{RR}}^2 - \mathrm{SD1}^2} \tag{3.7}$$

$$\mathrm{SD}_{\Delta} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N-1} (RR_{i+1} - RR_i)^2} \tag{3.8}$$

where N is the total number of RR intervals and $\mathrm{SD}_{\mathrm{RR}}$ is the standard deviation of RR intervals. The values of SD1 and SD2 are also used to calculate the Cardiac Sympathetic Index (CSI) and the Cardiac Vagal Index (CVI), where both represent the balance between sympathetic and parasympathetic tone.
Specifically, the first parameter is defined as:

$$\mathrm{CSI} = \frac{4 \times SD1}{4 \times SD2} \tag{3.9}$$

While the second is defined as:

$$\mathrm{CVI} = \log_{10}(4 \times SD1 \times 4 \times SD2) \tag{3.10}$$

These indices provide information on the autonomic regulation of the heart and highlight the differences between individual alterations in the ECG signal, allowing for the distinction between various cardiac conditions and the identification of specific patterns of irregularity in heart rate variability. [14] In this case, for each signal, a vector of 7 elements is generated and added to the 5 elements generated through the encoding of gender and age. These are directly concatenated to the features extracted from the model. In this case as well, training was performed to evaluate the performances.

3.4.6 Dataset reorganization

From the preliminary analysis of the available datasets described in the Section 3.3, a significant imbalance emerged, with a high representation of signals labeled as normal sinus rhythm, SNOMED code 426783006. To try to mitigate this problem, which could negatively affect the performance of a potential model, an analysis was conducted on the signals labeled with this class. In particular, boxplots were created to represent some features extracted from the signals for each databases, including power, heart rate, energy, standard deviation and mean value of RR intervals, number of beats, and Shannon entropy. [49] The extraction of these features was carried out starting from lead II and after applying the pre-processing

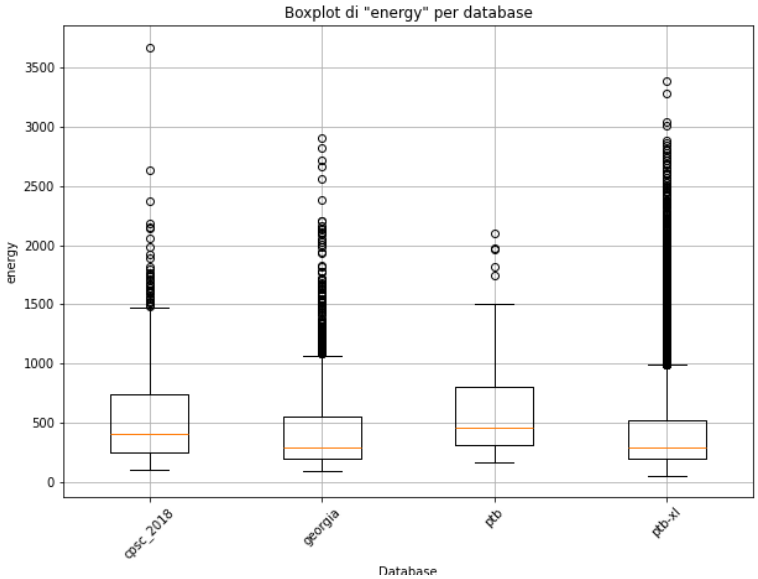

Figure 3.15. Boxplot of the energy feature for signals labeled as normal sinus rhythm. The black circles represent the outliers for each database.

described in the Subsection 3.3.1 to homogenize the dataset, using specific tools for processing electrocardiographic signals.

Starting from the created boxplots, outliers were analyzed for each feature. In particular, the quantities of these outliers were identified, and subsequently, training was carried out by identifying the feature with the highest number of outliers and removing the specific signals associated with it. Specifically, the outliers of the energy feature were removed as shown in Fig. 3.15, for a total of 1568 signals. This operation resulted in training set and validation set sizes of 28913 and 7235, respectively. The decision to exclude certain signals was made to try to homogenize the distribution, in relation to a specific parameter, for the most represented class in this dataset. Following this exclusion of some signals, a new training was carried out while keeping the modifications described in the previous sections unchanged, which achieved the best performances. In parallel, while keeping the model with the best performance unchanged, all outliers of all features related to the boxplots of the normal sinus rhythm class were removed. The aim was to eliminate all possible sources of noise belonging to signals of the same class but very dissimilar to each other in terms of representation and extracted features. As a result, the new sizes for the training set are 19836, while for the validation set they are 4986.

To further reduce the size of the most represented class, an additional attempt, different from the previous ones described before, was made to identify and select

only the signals associated with that specific class, but with the unique label of normal sinus rhythm. The reason for this choice lies in the fact that some of the signals with the normal sinus rhythm class were simultaneously associated with other classes representing certain alterations. For this reason, an effort was made to have the normal sinus rhythm label expressed uniquely. An example of the selection process is shown in Fig. 3.16.

This operation also led to a reduction in the sizes of the two datasets used, reducing the training set to 20836 elements and the validation set to 5211 elements. Again, a new training was carried out with these new specifications, keeping unchanged the characteristics that allowed achieving the best metrics.

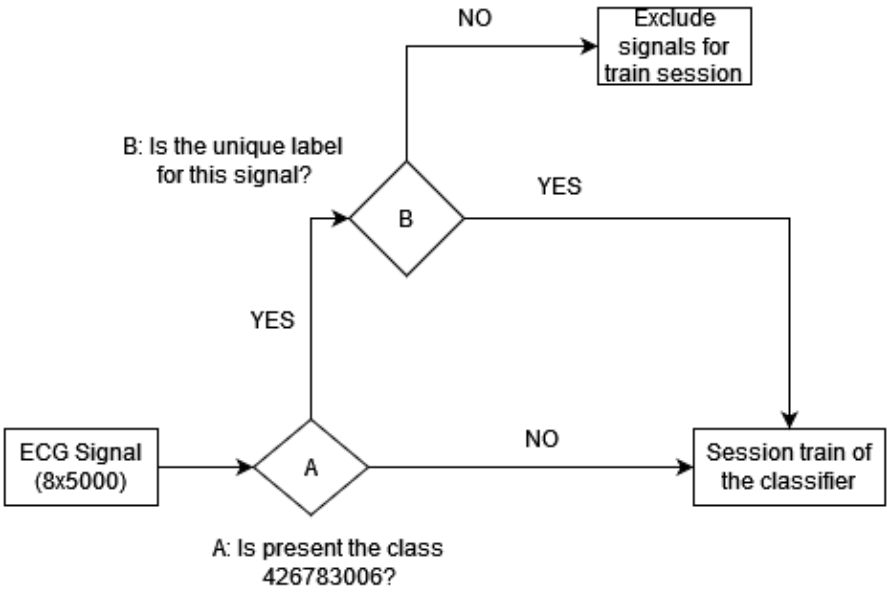

Figure 3.16. Example of pipeline before starting the train session.

3.4.7 Model decomposition

The issue of different numerosity for each classes in training sets necessitated the decomposition of the main model into more specific sub-models for certain classes that were grouped together in the global model. In particular, a sub-model was created to distinguish between bradycardia and sinus bradycardia, two very similar alterations but characterized by different shapes, which negatively affected the performance of the global model. A similar operation was carried out by grouping the classes T wave abnormal, T wave inverse, and Prolonged QT interval.

In this case, an additional step was introduced to improve the model's specificity and discriminate between similar classes. Starting from a classifier that discriminates a reduced number of classes, where some of them are grouped, if the output class belongs to those grouped, a second, more specific classifier is called. The architecture of the models added in cascade was kept identical to that of the general model, with the only difference being that the number of classes to be discriminated was reduced. The adoption of sub-models for an excessive number of groupings would have resulted in a significant increase in computational costs.

The attempt aimed to capture the characteristics of simpler models, with a smaller number of signals and elements in the training datasets. The workflow followed in the case of the sub-model that discriminates between bradycardia and sinus bradycardia is described in Fig. 3.17.

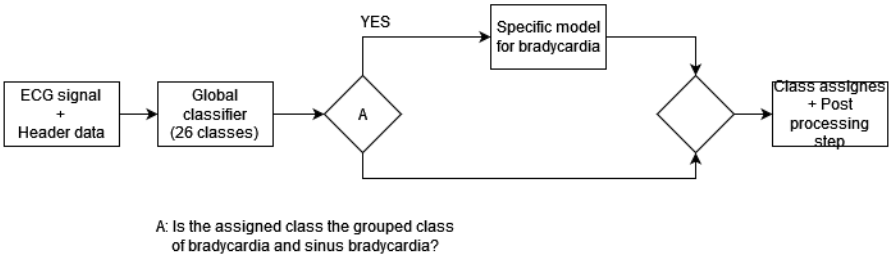

Figure 3.17. Example of pipeline by considering the submodel to discriminate bradycardia class and sinus bradycardia class.

The insertion of a large number of sub-models in cascade, capable of discriminating a reduced number of classes, would have led to a significant increase in computational resources.

Therefore, the problem of high class imbalance and the large presence of elements labeled as normal sinus rhythm was addressed by developing a classifier that distinguished only between two classes, with the possibility of assigning both simultaneously. The model architecture was kept similar to that described in the previous steps, but it was changed from a 27-class classifier to a 2-class classifier. Specifically, the two identified classes are healthy ECG and altered ECG. In this context, a relabeling operation was carried out in which all signals belonging to the normal sinus rhythm class were assigned to class 0, while all other classes were grouped into class 1. If both classes were present, the signals were labeled with both classes 0 and 1. This approach introduced an additional phase to differentiate between two classes with similar frequencies, thus reducing the imbalance that had appeared in previous models.

To assign the classes, different threshold values of 0.5, 0.55, 0.6, and 0.7 were tested, with the aim of identifying the optimal value by evaluating the performance

of each individual value. Subsequently, after class assignment through thresholding, a post-processing operation was carried out. In this context, if the probabilities of belonging to both classes were below the chosen threshold, an attempt was made to assign directly the class. Both solutions were tested: one in which the explicitly set class was the healthy one and the other in which the assigned class was the altered one. The metrics obtained from each of the two solutions were compared to identify which post-processing improved the model's accuracy the most. The lack of class assignment is managed through a check on a vector whose sum must be greater than or equal to 1 for the signal to have an associated label. An example workflow is shown in the Fig. 3.18, where, at the end of the thresholding operations, no class is assigned to the signal; therefore, the post-processing phase and the manual assignment of the healthy class are carried out.

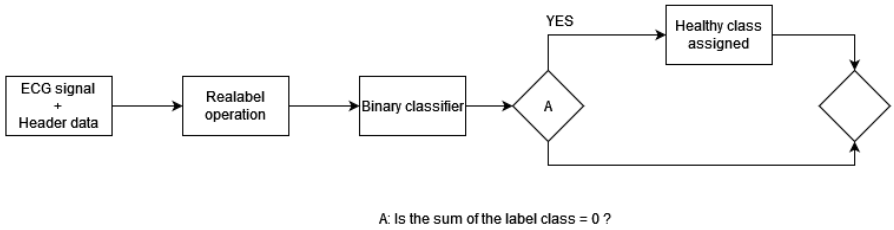

Figure 3.18. Example of pipeline for the binary classifier in which the post processing is to set the absent class to the healthy one.

Following this model, a second classifier capable of differentiating between the remaining 26 classes, excluding all signals labeled as normal sinus rhythm, was trained and added. This exclusion operation led to a new distribution of labels in the training and validation sets, as described in Fig. 3.19, and resulted in a change in their sizes: the training set consisted of 13529 elements, while the validation set had 3341 elements. As shown in Fig. 3.19, the removal of signals belonging to the normal class from the two datasets reduced the representativeness of some classes. In particular, it can be observed that the prolonged PR interval class (LPR, SNOMED code 164947007) has only 15 elements in the training set and 6 in the validation set, which significantly affects the final performance of the model. For this reason, two parallel training sessions were conducted to evaluate the performance and compare which approach offered the best results. In the first case, the training was conducted without making any changes to the base model used so, to be precise it was decided to maintain the threshold value for assigning a class equal to 0.36. In the second case, a weight was assigned to the loss function for each label in the entire dataset. The assigned weight was calculated based on the logarithm of the inverse of the frequency of each individual class, as proposed by the group in their code. In this way, the less represented classes, with a much

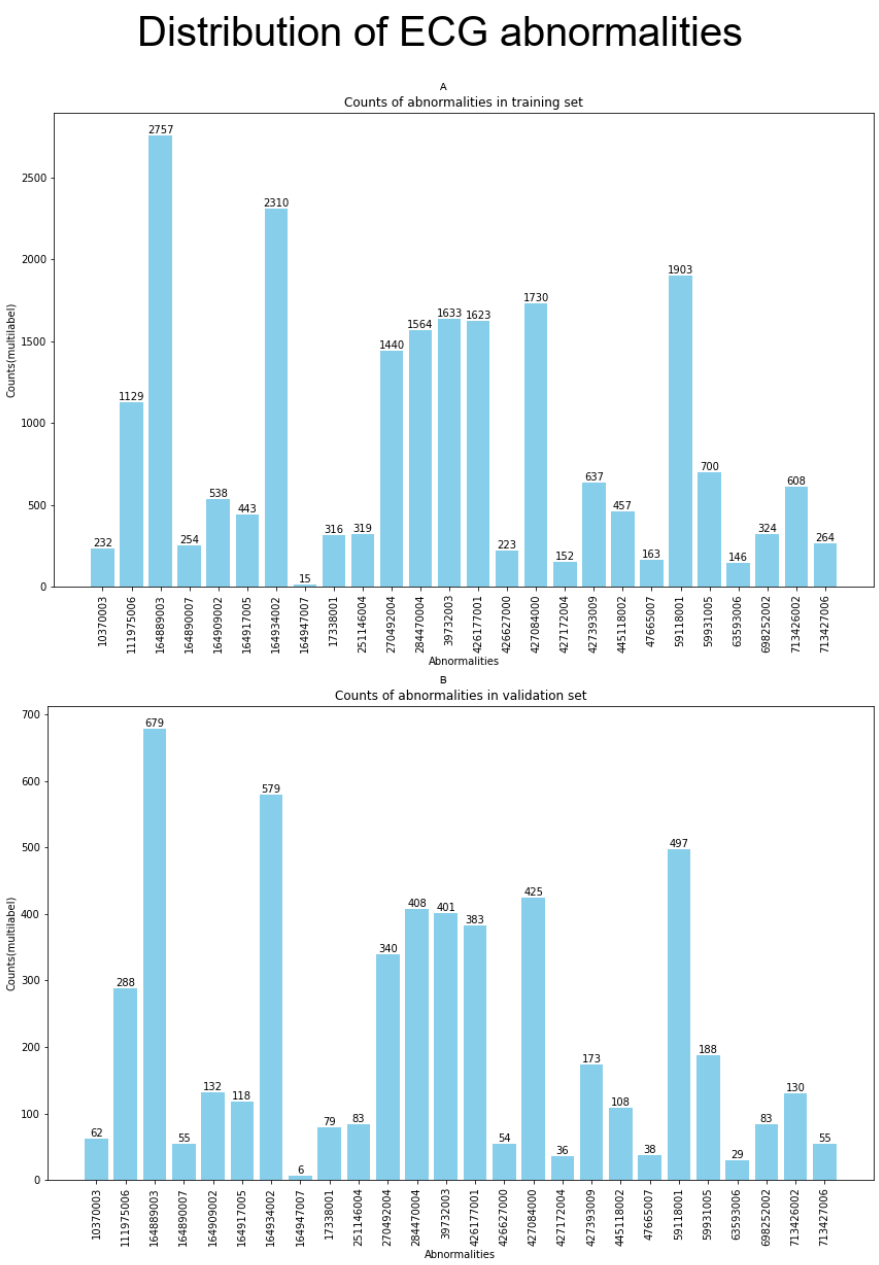

Figure 3.19. Distribution of the abnormalities in the training set and validation set after the exclusion of all elements in which the normal class was present.

smaller number of elements, were assigned a greater weight, which influenced the calculation of the loss function. The choice of the pipeline to follow was made by comparing the metrics obtained from the two different training sessions.

The main model has been divided into two sub-models specialized in different tasks, with the aim of addressing the problem of the high number of elements classified as normal rhythm. Specifically, the input signals are initially processed by a first model that classifies and distinguishes between normal rhythm and altered rhythm. If the signal is classified by the first model as belonging to the second class, it is subsequently sent to a second model, whose task is to distinguish between 26 different overall classes. The idea behind this approach was to organize the two models in a cascade, thus generating a final prediction that takes into account the succession of the models, thereby improving the specificity of the classification for the less represented classes.

To further increase the accuracy of the classifier, the original idea proposed by the group was integrated, which involves combining the outputs to generate a prediction from two signals of different lengths. In this context, the two cascade models were implemented through an ensemble of two sub-models, which handle signals of lengths equal to 5000 and 15000 samples, respectively. The final prediction for each signal, as well as for both models, is thus obtained by combining the results of the two sub-models, to which an equal weight of 0.5 is assigned. To determine the optimal configuration to use, an ensemble of the two sub-models was performed using signals with 12 and 8 leads as input. The Fig. 3.20 shows the final workflow in the specific case where the input signals are 8 leads. Furthermore, all post-processing operations adopted in the previous models have been kept unchanged.

3.4.8 Model variation

To compare the behavior of the baseline model with more recent models in the literature, a new network was identified to be used with the same data, in order to have a comparison and understand if the performance was in line with what was being obtained. The network used is an ECGNet proposed by Jin et al. (2024) [27]. The peculiarity is that the model was designed to be interpretable and therefore, in addition to identifying possible alterations, it is able to highlight the anomalous regions on the ECGs, thus helping clinicians to quickly locate problematic areas. The model was trained on a dataset created specifically for this network and contains signals from different Chinese patients between January 2017 and December 2021. It is designed to identify 6 alterations of the electrocardiographic signal, including: normal sinus rhythm (NSR), atrial fibrillation (AF), sinus tachycardia (ST), sinus bradycardia (SB), atrial premature contraction, and ventricular premature contraction. Since the model was used for comparison, the number of

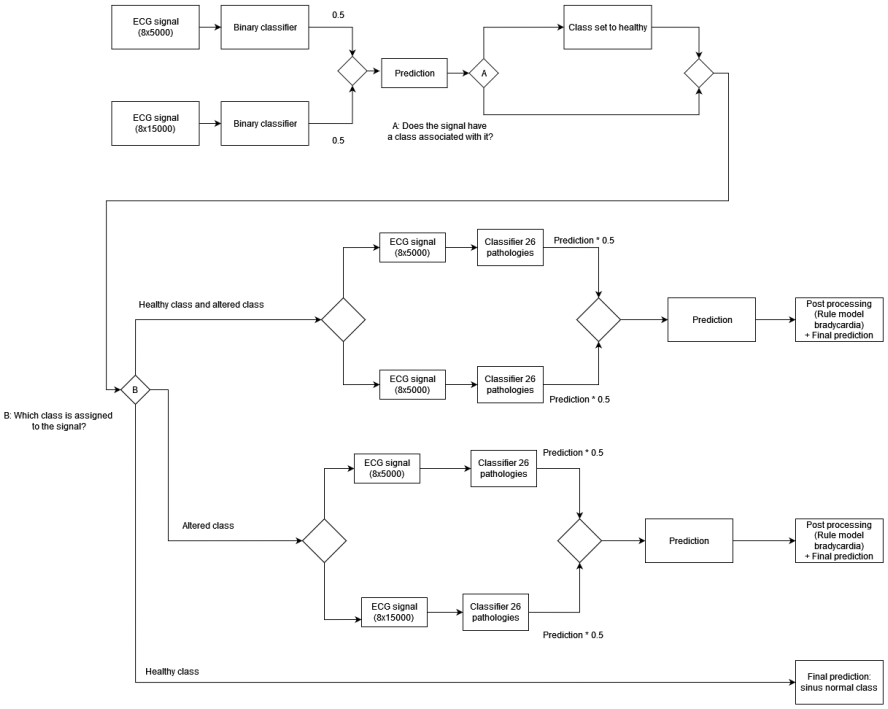

Figure 3.20. Example of the final pipeline considering the two models and their specific ensemble.

classes was modified, going from the original 6 classes to the 27 identified in the Cardiology Challenge. This was necessary to understand if the baseline had results in line with more specific and recent models.

The ECGNet is a network integrated into a telemedicine system to provide an assisted solution even outside the hospital. For this reason, the input signal considers only one lead, instead of the 8 previously considered, specifically lead II is used as input. The model uses a combination of Convolutional Neural Networks, Residual blocks, and Bidirectional Long-Short Term Memory (BiLSTM) layers to automatically extract features from ECG signals. The Residual blocks are responsible for extracting short-term dependence features, while the BiLSTM layer extracts long-term dependence features. Subsequently, an Attention (ATT) layer is used to enhance beneficial features and indicate the model's focus on the input features. Finally, a multi-label classifier processes the deep features to generate the prediction results. At the end of the predictions, a post-processing module, based on medical knowledge, is added. This module analyzes the outputs and corrects some predictions by extracting characteristic and discriminatory features of these cardiac arrhythmias, such as heart rate or the number of R peaks. The architecture is shown in Fig. 3.21.

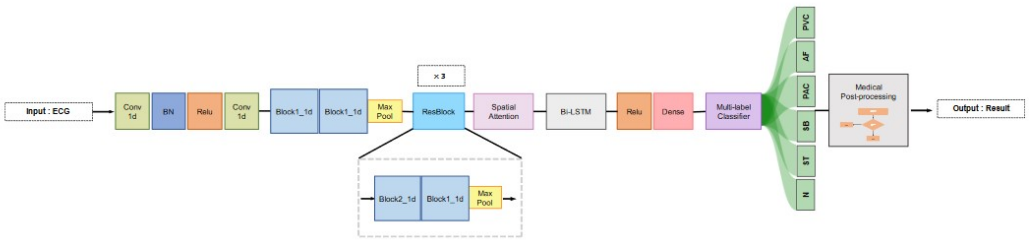

Figure 3.21. Architecture of the original ECGNet.

However, since this model was designed to identify 6 different cardiac arrhythmias, but the comparison is made on the 27 considered in the challenge, the post-processing module was not integrated into the final model. Instead, only the architecture was used to evaluate its performance. This model was therefore trained considering two types of input signals: 8 leads, like the models described so far, and a single lead to evaluate its performance while respecting the original architecture. The choice to test this model on the data of the challenge was made because, as stated by the authors, after a validation phase acquiring information from three expert cardiologists, the results obtained highlighted that the proposed diagnostic system has superior diagnostic performance over that of clinicians.

3.5 Training and evaluation of models

To evaluate the performance of a classification model in a multi-class and multi-label context and to maintain the comparison with the obtained results, a new scoring function proposed by the challenge organizers on the validation set was considered. This was necessary to identify whether certain modifications to the model improved the declared results. Additionally, this metric was evaluated on the additional test set to have a global assessment of the model's performance on this dataset after making modifications.

The metrics will therefore be calculated on the validation set and test set at the end of each training.

3.5.1 Challenge metric and performance evaluation

The new metric proposed by the organizers of the Cardiology Challenge 2020 aims to better reflect clinical reality, a limitation encountered when using traditional metrics such as the area under the curve (AUC). [3] It assigns partial credits to incorrect diagnoses that actually result in outcomes or treatments similar to the

real diagnoses, and similarly highlights how some incorrect diagnoses are more harmful and should be treated accordingly. Additionally, it reflects the fact that it is less harmful to confuse certain classes compared to others because the responses can be similar or identical. For the challenge metric, it starts with the creation of a modified confusion matrix $A = [a_{ij}]$, where a_{ij} is the normalized number of recordings in a database that were classified as belonging to class c_i but actually belong to class c_j (where c_i and c_j may be the same class or different classes). Given that each recording can have one or more labels associated with it, and the classifier can generate multiple class memberships, then for a generic signal k, let x_k be the set of actual classes and y_k be the set of predicted classes for that specific recording. Defining:

$$a_{ij} = \sum_{k=1}^{n} a_{ijk} \tag{3.11}$$

$$a_{ijk} = \begin{cases} \frac{1}{|x_k \cup y_k|}, & \text{if } c_i \in x_k \text{ and } c_j \in y_k \\ 0, & \text{otherwise} \end{cases} \tag{3.12}$$

where n represents the total number of recordings and the quantity $|x_k \cup y_k|$ is the number of distinct classes with a positive label and/or classifier output for recording k.

To assign credits to the obtained predictions, a reward matrix $W = [w_{ij}]$ is defined, where w_{ij}, defined by cardiologists based on the similarity between ECG signal alterations, are the reward for a positive classifier output for class c_i with a positive label c_j. The reward matrix is shown in Fig. 3.22. The reward matrix assigns the highest values along its diagonal, giving full credit for correct classifier outputs, partial credit for incorrect classifier outputs, and no credit for labels and classifier outputs not represented in the weight matrix.

The metric is thus defined as a weighted sum of the rewards from the reward matrix and the number of recordings classified as belonging to the specific class c_{ij}, representing a generalized version of accuracy:

$$s_{\text{unnormalized}} = \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} a_{ij} \tag{3.13}$$

where m represents the number of diagnoses considered.

To make the comparison easier, this score is normalized so that a classifier which always outputs the real classes or class receives a score of 1, while a classifier which always outputs the normal class receives a score of 0, defining it as an inactive classifier:

$$s_{\text{normalized}} = \frac{s_{\text{unnormalized}} - s_{\text{inactive}}}{s_{\text{true}} - s_{\text{inactive}}} \tag{3.14}$$

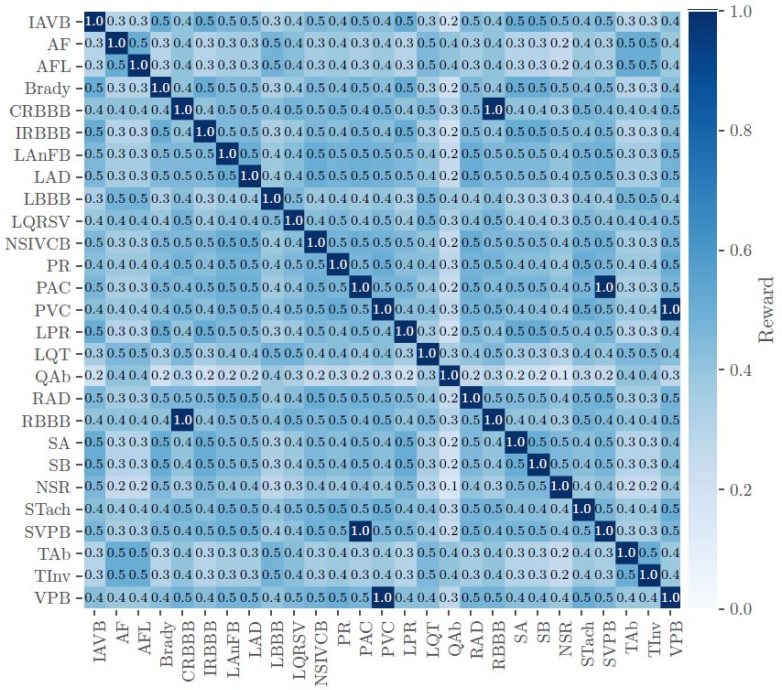

Figure 3.22. Reward matrix W used for scoring diagnoses in the Challenge is depicted with rows and columns labeled by the abbreviations for the diagnoses. Off-diagonal entries with a value of 1 indicate similar diagnoses that are scored equivalently to the same diagnosis.

where s_{inactive} is the score for the inactive classifier and s_{true} is the score for the ground-truth classifier.

Since some of the 27 classes are considered similar during the evaluation and calculation of metrics, the problem shifts from the expected 27 classes to 24. This is represented in the Fig. 3.22, where some scores equal to 1 (same anomaly) are also reported outside the main diagonal, indicating that these two alterations can be considered the same diagnosis.

Specifically, the following pairs of classes are treated as equivalent:

- Right bundle branch block (RBBB, SNOMED code 59118001) and Complete right bundle branch block (CRBBB, SNOMED code 713427006).
- Supraventricular premature beats (SPVB, SNOMED code 63593006) and Premature atrial contraction (PAC, SNOMED code 284470004).
- Ventricular premature beats (VPB, SNOMED code 17338001) and Premature ventricular contractions (PVC, SNOMED code 427172004).

Therefore, during the model evaluation, signals with one of the diagnoses contained in these pairs will be relabeled with one of the two, thus making the task a 24-class classification. For each specific class and considering the binary classifier, accuracy is evaluated as the ratio of the number of recordings for which all predicted labels match the true labels to the total number of recordings. This can be expressed as:

$$\text{Accuracy} = \frac{\text{Number of recordings with all correct labels}}{\text{Total number of recordings}}$$

Furthermore, to compare results and behavior on individual alterations, the F1 score has been taken into consideration, defined as:

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{3.15}$$

where TP are the number of examples correctly classified as positive, FP are the number of examples incorrectly classified as positive, and FN are the number of examples incorrectly classified as negative.

And in particular, taking into account different classes, the macro averaged F1 score has been calculated, that is computed using the arithmetic mean (unweighted mean) of all the per-class F1 scores. This method treats all classes equally regardless of their support values. Defined as:

$$\text{Macro F1 Score} = \frac{\sum_{i=1}^{m} \text{F1 Score}_i}{m} \tag{3.16}$$

3.6 Clinical feedback and model modifications

The importance of having a clinical opinion in this type of project is very significant and allows for the inclusion of purely health-related information into a primarily engineering vision. During the project, it was possible to have a discussion with experts in the field, particularly with a cardiologist specializing in electrophysiology/arrhythmology who performs activities in electrostimulation, electrophysiology, and clinical arrhythmology and possesses extensive knowledge of the electrocardiographic signal and its alterations.

From this meeting, it emerged that some of these electrocardiographic signal alterations affect a single beat and not the entire signal trace. For this reason, the first modification that was integrated was to perform two model trainings with shorter duration signals as input, so that evaluations could be made on individual alterations that are more evident in a single beat and for which it is possible, through the isolation of the latter, to assess their presence and/or absence. Specifically, two trainings were conducted with input signals of 500 samples,

corresponding to approximately 1 second of signal, and 1000 samples, corresponding to 2 seconds of signal. These trainings were carried out while maintaining the previous modifications and excluding the normal sinus rhythm class to exclusively discriminate between individual alterations. The idea, once the performance of these models was obtained, was to give greater weight when performing inference if the metrics obtained on alterations recognizable through a single beat in the clinical setting were higher than those considering input signals of 5000 samples. Among the diagnoses recognizable through a single beat are: complete and incomplete bundle branch blocks, left and right axis deviations, inverted T wave, and abnormal Q wave.

Furthermore, from the meeting, it emerged that some of the 27 classes were similar to each other and that one of these was not very functional to maintain recognition. Therefore, these similar classes were grouped, and the classes deemed not important to recognize for a clinical task were excluded from the training. Specifically, the total number of classes was reduced from 27 to 20, that are represented in the Tab 3.2. This was achieved by merging the classes bradycardia and sinus bradycardia, abnormal T wave and inverted T wave, supraventricular premature beats and premature atrial contraction, ventricular premature beats and premature ventricular contractions, left axis deviation and right axis deviation, and prolonged PR interval and 1st degree atrioventricular block, and excluding the nonspecific intraventricular conduction disorder class during the evaluation of the metrics. In this case as well, the modifications were integrated during the perfoming of inference when the metrics are evaluated.

Both modifications made to the model were carried out to integrate clinical knowledge into the proposed model to perform a classification task that best reflects the hospital setting.

Diagnosis
Atrial fibrillation
Atrial flutter
Bradycardia/Sinus bradycardia
Complete right bundle branch block/Right bundle branch block
Incomplete right bundle branch block
1st degree AV block/Prolonged PR interval
Left anterior fascicular block
Left axis deviation/Right axis deviation
Left bundle branch block
Low QRS voltages
Pacing rhythm
Premature atrial contraction/Ventricular premature beats/Supraventricular premature beats
Premature ventricular contractions
Prolonged QT interval
Q wave abnormal
Sinus arrhythmia
Sinus rhythm
Sinus tachycardia
T wave abnormal/T wave inversion
Nonspecific intraventricular conduction disorder

Table 3.2. Diagnoses considered after the grouping suggested during the meeting with the cardiologist.

Chapter 4

Results

Tab 4.1 shows the Challenge metric results for the top five groups in the Cardiology Challenge, comparing performance on the training, hidden validation, and hidden test datasets. The values highlighted in yellow represent promising results obtained by the third group, maintaining performance in line with the other groups on the hidden test set, which is why this solution was selected. It is important to note that the chosen solution, consisting of an ensemble, was validated by clinicians in the first phase, especially the model outputs were subjected to clinical review by two cardiologists. This aspect has a considerable impact on the choice, as the other solutions did not undergo clinical validation. Additionally, although the other results appear less affected by overfitting, since the hidden test set was not released, the path that showed the best results on their own training sets was preferred.

	Challenge metric		
Group	Training set	Hidden Validation set	Hidden Test set
1° group	0.675	0.587	0.533
2° group	0.684	0.672	0.520
3° group	0.885	0.682	0.514
4° group	0.724	0.640	0.485
5° group	0.629	0.609	0.437

Table 4.1. Challenge metrics for the five top ranked groups in the Cardiology Challenge across training, validation, and test sets.

Tab 4.2 presents a comparison of the new score challenge metric values on the validation set, obtained by attempting to replicate the results of the third-place group using different weight vectors provided in their submission file. The results show that Model B, using the $best_w$ weight vector, produced the similar results,

Model (weight vectors)	Challenge Metric Validation	Declared by the Group 3
Ensemble (current w+best_w)	No information	No information
Ensemble (best_w_15k8L + best_w)	0.628	0.663
Model A (current_w)	0.800	No information
Model A (best_w_15k8L)	0.190	0.674
Model B (best_w)	0.663/0.665	0.674

Table 4.2. Comparison of challenge metric values on validation by trying to repeat the code with the different weight vectors provided by the third-ranked group in their submission zip file.

highlighted in yellow, both in the code replication and in the results declared by the group.

Tab 4.3 presents the Challenge metric results on the validation and test sets after integrating additional variables (covariates) such as gender and age. Various solutions are compared: G represents the division of age into groups, while N indicates the normalization of age relative to the maximum. The numbers 2 and 3 indicate the type of variable processing: with number 2, the variables are processed by a ReLU layer, generating 16 additional features, while with number 3, they are processed by a linear layer, generating 10 features.

Challenge Metric		
	Validation set	Test set
Baseline (no modifications)	0.664	0.394
Gender 2 + Age (N)	0.673	0.414
Gender 2 + Age (G)	0.667	0.395
Gender 3 + Age (N)	0.675	0.399
Gender 3 + Age (G)	0.669	0.398

Table 4.3. Results of the challenge metric on the validation set after the concatenation of covariates to the output of the model. The letter G represents the solution in which the age is subdivided into groups, while the letter N represents the solution in which the age is normalized by the maximum. The number 2 represents the solution in which the variables are elaborated by the ReLU layer, resulting in 16 additional features generated, while the number 3 represents the solution in which the variables are elaborated by the Linear layer, resulting in 10 additional features.

Fig. 4.1 and Fig. 4.2 show the trend of the Challenge metric on the validation set and the test set, respectively, after implementing various modifications to the baseline model, which already includes the covariates gender and age processed by the linear layer.

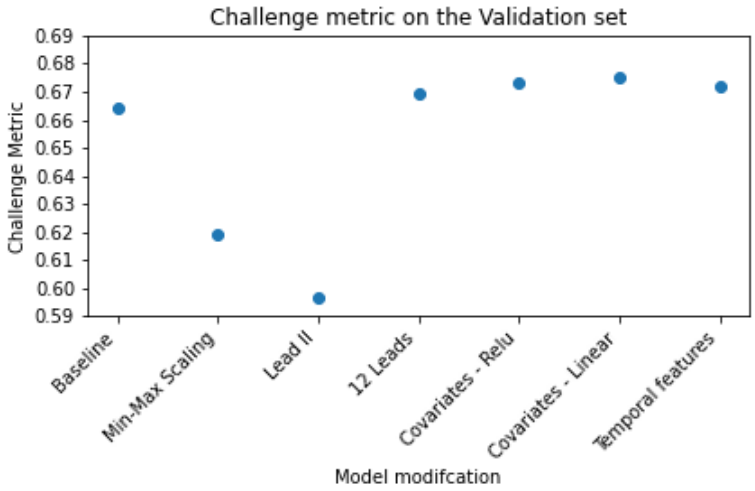

Figure 4.1. Comparison of values of challenge metric on the validation set after different modifications. Each model starts from the model that has integrated the covariates elaborated by the Linear layer.

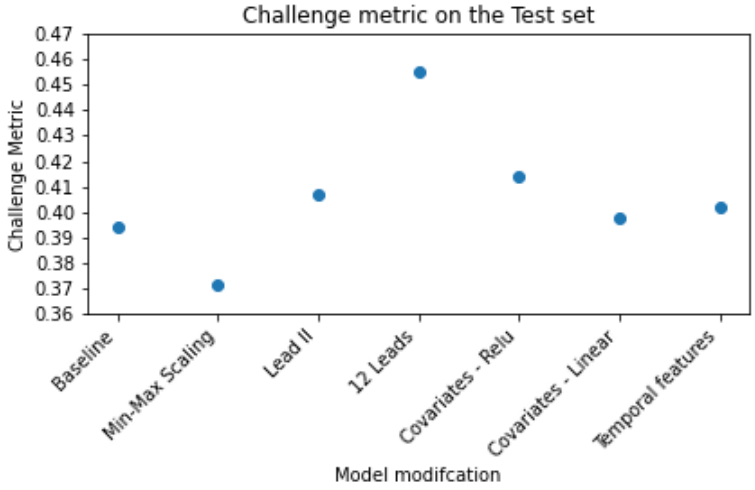

Figure 4.2. Comparison of values of challenge metric on the test set after the different modifications. Each model starts from the model that has integrated the covariates elaborated by the Linear layer.

Tab 4.4 compares the Challenge metric results obtained by the new model with integrated covariates against the ECGNet model on validation and test sets, using both 8-lead and single-lead ECG. The results highlight that the model with

Challenge Metric		
	Validation set	Test set
8 Leads (Covariates model)	0.675	0.398
8 Leads (ECGNet)	0.502	0.357
1 Lead (Covariates model)	0.590	0.447
1 Lead (ECGNet)	0.482	0.419

Table 4.4. Comparison of results obtained by the new model considered with the model in which the covariates are integrated and two different inputs: 1 lead and 8 leads.

covariates achieves superior performance compared to ECGNet, both with 8 leads and 1 lead. Tab 4.5 reports the Challenge metric values obtained after various attempts to balance the dataset. Three solutions were experimented with: the removal of outliers based on signal energy, the removal of outliers across all features, and the classification of all elements plus elements where the single associated class is normal sinus rhythm.

Challenge Metric		
	Validation set	Test set
Covariates model	0.675	0.398
No outliers energy	0.666	0.377
No outlier of all features	0.618	0.404
Single class normal sinus rhythm	0.508	0.377

Table 4.5. Challenge metric values obtained after the different attempts done to balance the dataset.

Tab 4.6 presents the results in terms of Accuracy and MacroF1 score for a binary model using 12-lead ECG. Various configurations were tested: the use of signals with a length of 5000 or 15000 samples, and two approaches for handling the missing class: as altered ("A") or as healthy ("S"). The results show that the ensemble combination, which integrates models with different signal lengths, leads to improved performance, with the configuration handling the missing class as healthy ("S") achieving the best F1 score results on both the validation and test sets. Table 4.7 compares the Challenge metric values for two cascade models, where the first stage is the binary model from Tab 4.6, using the configuration with a 0.5 threshold and handling the missing class as healthy ("S"). The first cascade model uses 12-lead ECG for both the binary model and the subsequent 26-class model, while the second cascade model replaces the 12 leads in the second

stage with 8-lead ECG.

	Validation		Test	
	Accuracy	F1	Accuracy	F1
12 leads + 5000 samples + A	0.840	0.925	0.882	0.830
12 leads + 5000 samples + S	0.840	0.925	0.881	0.829
12 leads + 15000 samples + A	0.841	0.925	0.877	0.825
12 leads + 15000 samples + S	0.842	0.925	0.876	0.825
Ensemble + S	0.851	0.930	0.890	0.843
Ensemble + A	0.847	0.918	0.885	0.837

Table 4.6. Accuracy and MacroF1 score obtained considering the binary model with 12 leads. 'A' indicates handling the missing class as altered, while 'S' indicates handling the missing class as healthy. The threshold to assign the class is set to 0.5. The ensemble phase is the ensemble between the model which input's signals have a length of 5000 samples and the model in which input's signals have a length of 15000 samples.

Challenge Metric		
	Validation set	Test set
Cascade model (12 leads + 12 leads)	0.613	0.432
Cascade model (12 leads + 8 leads)	0.617	0.450

Table 4.7. Challenge metric values comparing the two ensemble cascade models. The first model includes a binary model consisting of an ensemble with 12 leads of two different lengths (15000 samples and 5000 samples), followed by a 26-class model considering the 12 leads and the ensemble of two different lengths. The second model considers the 26-class model with 8 leads and the ensemble of two different lengths.

The presented results reflect the implementation of clinical knowledge acquired during a meeting with a cardiologist. Tab 4.8 reports the attempts to reduce the length of the ECG signals expressed in samples (500 and 1000), using the initial configuration with 5000 samples as a reference. Tab 4.9 shows the results obtained by grouping some similar classes, as suggested by the cardiologist. The integration of these modifications led to a notable improvement for both the covariates model and the ensemble model.

Fig. 4.3 and Fig. 4.4 show the F1 score values for all pathologies on both the validation and test sets, comparing the model that achieved the best performance and the implemented solution that achieved consistent performance on the

Challenge Metric		
	Validation set	Test set
5000 samples	0.675	0.398
1000 samples	0.663	0.430
500 samples	0.616	0.414

Table 4.8. Comparison of challenge metric values using different signal lengths excluding the normal sinus rhythm class and considering the model with 26 classes.

Challenge Metric		
	Validation set	Test set
Covariates model	0.675	0.398
Covariates model + class grouped	0.692	0.471
Ensemble Model	0.617	0.450
Ensemble Model + class grouped	0.644	0.504

Table 4.9. Comparison of results by considering the ensemble model and the best model with the integration of the covariates between the normal solution with 27 classes and the other case in which the classes are grouped.

hidden test set. In some cases, the F1 score is 0 due to the absence of representative elements for that class. The two figures compare the results obtained before and after integrating the clinical information provided by the cardiologist. It can be observed how the grouping of some similar pathologies, suggested during the meeting, influenced the F1 values, improving the model's performance in several cases.

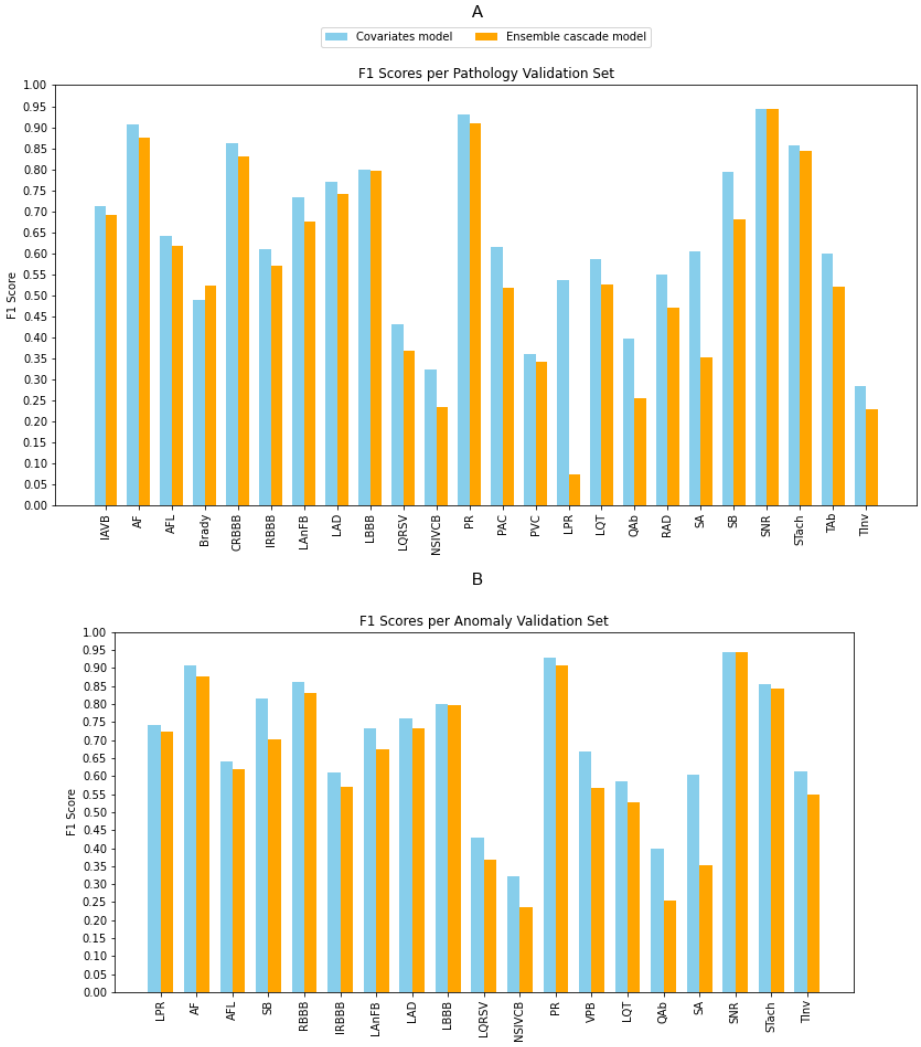

Figure 4.3. F1 score for each diagnoses in validation set. **A**. F1 score for diagnoses in validation set considering the model with only alterations and the ensemble cascade model before grouping some pathologies following the meeting with the specialist. **B**. F1 score for diagnoses in validation set considering the model with only covariates and the ensemble cascade model after grouping some alterations following the meeting with the specialist.

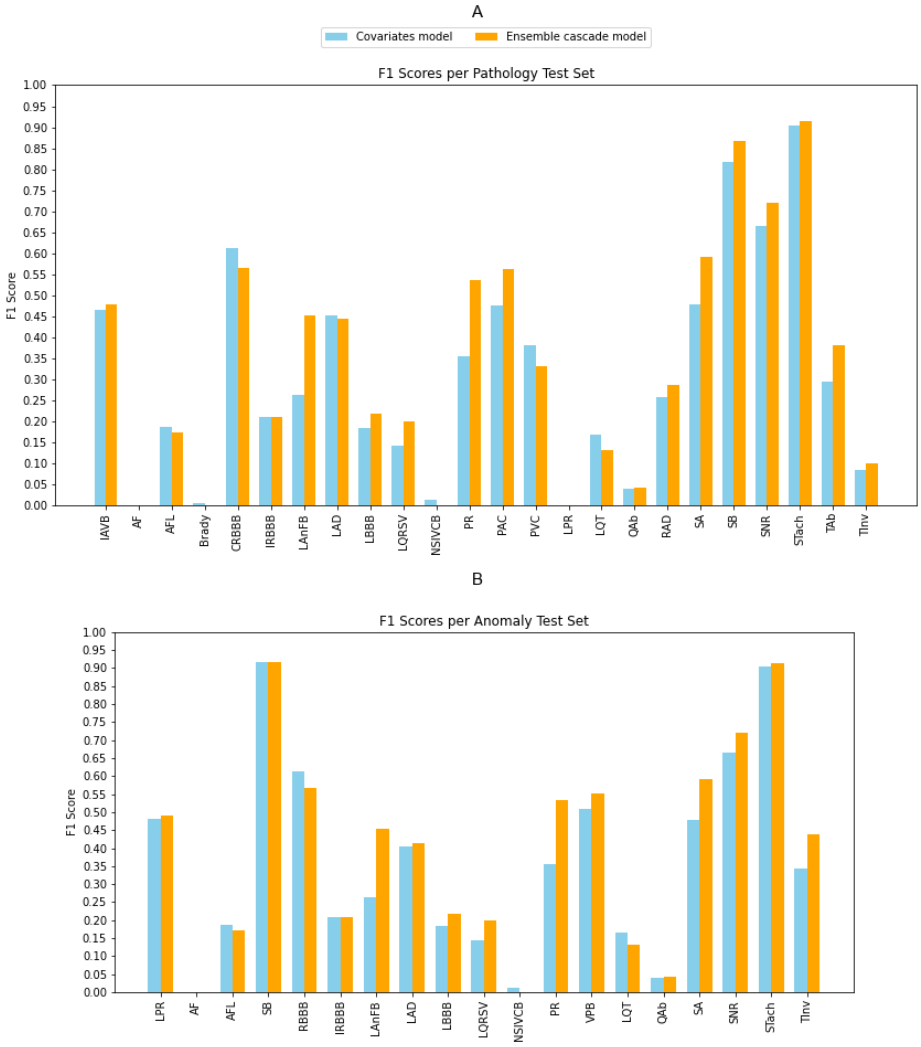

Figure 4.4. F1 score for each diagnoses in test set. **A**. F1 score for diagnoses in test set considering the model with only alterations and the ensemble cascade model before grouping some pathologies following the meeting with the specialist. **B**. F1 score for diagnoses in test set considering the model with only covariates and the ensemble cascade model after grouping some alterations following the meeting with the specialist.

Chapter 5

Discussion

Identification of the starting pipeline

The choice of the model to use as a baseline during the thesis project and its reproduction represented a crucial starting point for identifying the necessary steps to improve the system's performance. From the comparison of the results obtained by the top five groups in the 2020 Cardiology Challenge [53, 35, 21, 37, 51], it emerged that generalization ability was a common challenge for all proposed solutions, with performance on the hidden test set around 50%, depending on the metric used for the challenge, as shown in Tab 4.1. This motivated the decision to evaluate the proposed models on all subsets of data used for evaluation, including not only the hidden test set but also the results obtained on the training and validation sets, despite the division of the latter being at the discretion of each participating group.

The choice thus fell on the solution proposed by the third-ranked group [52], as the declared metrics, particularly on the training and validation sets, appeared promising and constituted a solid starting point. Additionally, from a technical perspective, the solution was easily understandable. However, reproducing this model encountered difficulties related to what was reported in the provided code. To ensure exact reproducibility, the code uploaded on the Challenge page was executed without any modifications, in order to compare the declared results with those obtained after reproduction. Since the metrics declared by the group referred exclusively to the validation set, all comparisons were made considering specifically the validation set used for model training.

In the case of the final solution the provided code already included the weight vectors to be used. As for model A, it was used for a relabel of the CPSC dataset, subsequently validated by clinicians. Indeed each training started from the relabeled dataset, considering the new classes assigned to the signals for that specific

database. However, reproducing the provided code did not yield the results declared by the group in its paper. From the analysis of the Tab 4.2, it emerges that the possible cause of this lack of reproducibility could be attributed to an incorrect declaration of the weight vectors for the ensemble. Even in the simple reproduction of the single model, without integrating the ensemble phase, a disparity of about 0.1 was noted between the obtained result and the declared one, thus substantially impacting subsequent modifications to the model.

To address the issue of non-reproducibility, the main model was decomposed into sub-models, verifying each for code errors that might prevent full reproduction. This process identified that the ensemble phase between two signals of different lengths was missing, with only 10-second-long signals being input into the network. However, the results from the two code executions were consistent with the declared results, confirming the methodology's correctness.

Modification of the starting pipeline

To improve classifier performance, several modifications were introduced, including additional pre-processing steps and changes to the network architecture to handle different input types. For instance, signals were normalized to scale data between 0 and 1 while preserving important features like signal peaks. [16] These modifications were evaluated using the challenge metric on the validation and test sets to assess generalization.

Among the implemented modifications, the one that had a significant impact on the model's performance was the addition of covariates, such as gender and age, as supplementary inputs. These variables were concatenated to the features extracted by the model before generating the prediction and assigning the class. This choice was motivated by the awareness that age affects the progression of certain arrhythmias and was inspired by an article published by the group from which the reference solution was taken. [52] In that article, it was highlighted that the main difference compared to the solution ranked second in the Challenge was precisely the inclusion of covariates in the model. [50]

To evaluate the effect of the depth of the additional layer architecture and the processing mode of these variables, two types of flows were used to handle the encoded covariates as shown in Tab 4.3, in which it is possible to understand that the best perfomance are obtained by the solution that linearly combines the features and normalizes age by the maximum value rather than dividing it into groups. Regardless of the method used for processing, the addition of this information highlighted how crucial the patient's clinical history is for correct class prediction. The clinical history represents an essential source of information that allows contextualizing the data collected during instrumental analyses, offering a more complete view of the patient's health status. It includes details on medical

history, chronic conditions, ongoing pharmacological treatments, and previous cardiac events, which can significantly influence the presentation of arrhythmias. [45] Such information provides a context in which to interpret physiological signals, allowing the model to recognize patterns that could be associated with specific clinical conditions, thus improving prediction accuracy.

From this consideration emerges the importance of having not only the signals and associated classes but also variables that describe the patient's social and clinical condition. Currently, the only information available in the header files concerned gender and age, but it would be useful to have other relevant information that could affect the presence of electrocardiographic anomalies. For example, it might be useful to generate variables indicating the presence or absence of risk factors, such as smoking habits, weight, or height of the patient. [15]

These considerations are supported and were confirmed during a meeting with a cardiologist, during which it became clear how parameters related to the patient's life can influence the presence of electrocardiographic anomalies. To improve the model's ability to correctly classify the 27 cardiac arrhythmias, the covariates appropriately encoded within the features generated as output by the neural network were kept constant. Starting from this basic configuration, various pre-processing techniques were experimented with the aim of increasing the model's performance. However, none of these methodologies led to a significant improvement in performance metrics, as highlighted in the previous section.

The ineffectiveness of these pre-processing strategies, shown in Fig. 4.1 and Fig. 4.2 can be attributed to two main reasons. Firstly, the model architecture configuration was optimized by the group from which the solution was derived specifically for input signals acquired from 8 leads. Consequently, any variation in the dimensions or shape of the input, even when accompanied by code modifications to adapt the model, did not allow the architecture's full potential to be exploited. This structural rigidity implies that the 8-lead input, for which the model was designed, constitutes the optimal format on which the architecture can express its maximum capabilities. Secondly, the addition of pre-processing techniques, such as normalization, although theoretically promising for improving signal quality and reducing variability, did not facilitate the model's training in effectively distinguishing the 27 classes. On the contrary, these techniques may have further complicated the signal representation, introducing additional complexity that the model was unable to adequately handle.

Therefore, the choice to maintain the 8-lead model, along with the application of wavelet denoising as the only pre-processing technique, is confirmed as the most balanced solution to preserve input quality and make the best use of the implemented architecture. Attempts to modify the input or add further pre-processing stages following wavelet denoising did not bring significant improvements in terms of performance, suggesting that the current model configuration may already be

close to a local optimization, where further interventions do not produce tangible benefits.

The integration of temporal features with those generated by the neural network and with the encoded covariates allowed for results comparable to those obtained with the best model. However, the evaluation metrics could have been significantly better if all the signals used had been adequately cleaned. In such a case, it would have been possible to more precisely extract the parameters necessary for the calculation of temporal features, as highlighted in the Fig. 5.1.

In particular, it was observed that on some signals, the tools used to identify the RR peak indices did not achieve the expected results, making errors in identifying some peaks and failing to locate clearly present peaks. This highlights the need to review the methodologies employed for peak identification, despite using solutions currently implemented even in modern devices for processing electrocardiographic signals. It is important to emphasize that, despite the promising results obtained with the addition of these features, the significant increase in computational costs made their definitive implementation impractical. In fact, training the model required considerable times, with durations reaching up to 24 hours. Therefore, in light of the current limitations in terms of efficiency and computational resources, the integration of these temporal features was not implemented in a stable manner.

Influence of dataset quality and analysis output with its implications

In all the analyses conducted, the evaluation phase was followed by a careful analysis and comparison between the model's predictions and the actual classes. This phase led to the emergence of two main critical issues. The first concerns the high representativeness of the sinus rhythm class, which resulted in frequent misclassification of signals that do not belong to this class but were still assigned to it due to its numerical predominance in the dataset. The second issue that emerged is the model's difficulty in distinguishing between some classes that, although considered different during labeling, are actually very similar in their manifestation on the electrocardiographic trace.

For example, the classes inverted T wave and abnormal T wave often created confusion in the classifier. After a more in-depth literature review and consultation with cardiology experts, it emerged that the abnormal T wave class represents a broader set of alterations that also include the inverted T wave, thus making it understandable the model's difficulty in differentiating these two types of signals. [31] This overlap between classes suggested the need to reconsider their definition or, alternatively, to improve the quality of the dataset to minimize these ambiguities.

In light of these observations, it was deemed necessary to focus particularly on the quality of the starting dataset, in order to identify and mitigate any sources

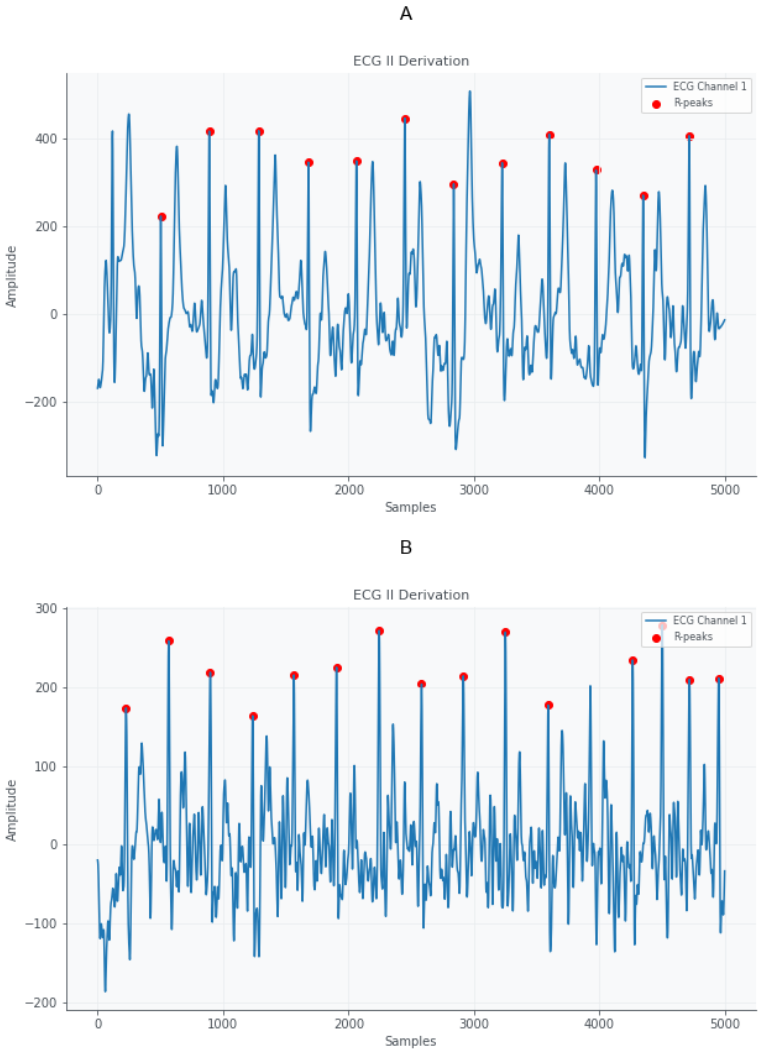

Figure 5.1. Examples of signals in which R-peak detection is compromised by low signal quality. **A**. 10-second ECG trace from the PTB-XL database labeled as normal sinus rhythm (SNOMED code 426783006). **B**. 10-second ECG trace from the Georgia database labeled as 1st degree AV block (SNOMED code 270492004).

of noise that could negatively affect the model's performance. A thorough review of the labeled classes, along with the elimination of problematic or ambiguous signals, was considered crucial to improve the overall accuracy of the classification system. In parallel with these considerations, the decision was made to train a new model based on a completely different architecture from the one previously used. This was done with the aim of verifying whether the performance of the initial

model was in line with those of more recent and advanced models. Additionally, it was sought to determine whether the limitations encountered in performance were attributable to the computational constraints of the original architecture or whether the main cause was to be found in the dataset used for training. This dual strategy, improving the quality of the dataset and introducing an alternative architecture, allowed for a deeper exploration of the potential causes of the difficulties encountered, offering a more comprehensive perspective to address the challenges of classifying cardiac arrhythmias. The choice to implement ECGNet as the new model is based on the inclusion of BiLSTM layers, known for their ability to handle complex sequential data and capture long-term dependencies. [12] The analysis of ECG signals, characterized by their temporal nature, requires models capable of storing and processing information distributed over long temporal sequences, in order to correctly identify variations in heart rhythm. LSTMs, thanks to their unique architecture, are particularly suitable for this type of task, as they allow relevant information to be maintained for extended periods, reducing the risk of losing critical details. Additionally, LSTMs excel in recognizing complex patterns within the data, such as those associated with different cardiac arrhythmias, thus improving classification accuracy. The use of bidirectional LSTM layers allows for a more comprehensive analysis of the signals, as it enables the model to consider the temporal context both forward and backward, further enhancing the ability to capture relevant features. The results obtained, as shown in the Tab 4.4, indicate that the use of the ECGNet architecture, both in the case of input from a single lead and using 8 leads, produced comparable, if not slightly inferior, performance to the architecture used so far. These results suggest that the main problem encountered does not lie so much in the model itself, but rather in the intrinsic limitations of the dataset used for training. In particular, the dataset proved to be not sufficiently robust, characterized by suboptimal variability and an inadequate number of samples to ensure effective learning and appropriate generalization. This limits the model's ability to accurately distinguish between the numerous arrhythmia classes, negatively affecting overall performance. The comparison between the two different architectures highlighted the importance of intervening on the dataset to improve the model's classification performance. One of the first issues addressed was class imbalance, characterized by a high number of signals labeled as normal sinus rhythm. This imbalance led to an overrepresentation of this class, with the consequence that many signals were erroneously classified as belonging to normal sinus rhythm, even when they were not. To address this issue, it was initially decided to work on this subset of signals, aiming to make the dataset more homogeneous and representative of the different classes. The goal was to create a distribution of signals for the sinus rhythm class that was as uniform and consistent as possible, eliminating signals that could introduce noise or ambiguity in the classification. However, these attempts to clean and homogenize

the subset of signals labeled as sinus rhythm did not produce the expected results. In particular, the idea of removing signals that deviated from the distribution of a specific feature was based on the assumption that a class with more homogeneous signals would facilitate the model's learning process, allowing it to better identify the distinctive characteristics of each arrhythmia. The ultimate goal was to generate a distribution of signals related to the sinus rhythm class that was entirely free of anomalies or significant variations, which could confuse the model.

If this approach had shown an improvement in performance, the same process could have been extended to the other more numerous classes in the dataset, simultaneously addressing two issues: class imbalance and the presence of signals labeled with the same class but significantly different from each other. This strategy could have led to a reduction in intra-class variance and an improvement in the model's ability to discriminate between different arrhythmias. Unfortunately, however, the results did not show an overall improvement in the model's classification performance or its ability to distinguish between classes, as shown in Tab 4.5. For this reason, this approach was not pursued further.

Another attempt to manage the size of the sinus rhythm class was to select only the signals whose label was solely sinus rhythm, excluding those where the normal class coexisted with other labels. The goal was to reduce the complexity of the normal class, hoping to improve the model's accuracy in this specific category. However, training the model on this subset revealed a significant limitation: the model learned to correctly classify signals with the sole label sinus rhythm but lost the ability to correctly recognize signals where sinus rhythm was present along with other labels. This issue led to a decrease in performance metrics for the multilabel classification task, highlighting that excessive simplification of the dataset, while facilitating the classification of single labels, compromises the model's ability to handle more complex and realistic tasks, such as correctly identifying label combinations.

Introducing the final model and clinician influence: encouraging performance on datasets

Having identified the problem of imbalance and recognizing the difficulty of addressing it with a classifier capable of distinguishing between 27 classes, the idea emerged to decompose the initial model into two distinct sub-models: the binary model and the 26 classes model. The strategy of the binary model proved particularly promising, as the results obtained using 12-lead input signals showed an accuracy of 85% and an F1 score of 93%, as shown in Tab 4.6. The latter, in particular, is a significant metric, especially in a multi-label classification context, despite the approach being binary. Additionally, this model can generate a third class where both categories coexist, increasing reliability in distinguishing between

healthy and altered heart rhythms, or in simultaneously detecting both conditions. The accuracy of this binary model could, in some cases, ensure a level of confidence comparable to that of a diagnosis made by an expert cardiologist. This aspect is particularly relevant, as the project aims to achieve a level of reliability that reduces the need for manual review of labeled traces by clinicians, thus allowing for efficient integration of deep learning models into the diagnostic process.

Subsequently, following this binary model, an 8-lead classifier was introduced for differentiating the remaining 26 classes. At this stage, an interesting behavior was noted: while the binary model's performance was slightly superior with 12-lead signals, the second model showed a slight improvement when 8-lead signals were used as shown in Tab 4.7. The analysis of the outputs confirmed that the first model significantly reduced misassignments to the sinus rhythm class, demonstrating how the preliminary phase, focused on distinguishing between healthy and altered, contributes to an overall improvement in accuracy.

Regarding the second model, it was observed that the particularly imbalanced dataset significantly affects the final performance, despite the effectiveness of the previous binary model. Therefore, it might be worth considering modifications to the second model to improve differentiation between individual alterations. One possible modification could be to work not with 26 distinct classes but rather to identify macro-classes that group alterations based on their type and presentation on the electrocardiographic trace. For example, an approach could be to divide the alterations into five main macro-classes: rhythm, conduction abnormalities, acute coronary syndromes, axis deviations, and chamber enlargements. [22] Within these macro-classes, various pathologies characterized by similar alterations could be grouped. Once the macro-classes are identified, specific sub-models could be trained to distinguish a smaller number of alterations, thus improving overall classification accuracy. For example, in the rhythm category, alterations such as bradycardia, tachycardia, and atrial fibrillation could be grouped, which share substantial differences in rhythm and graphical representation of the electrocardiographic signal. This approach could lead to more accurate classification, as the original 26-class task would be broken down into tasks with a smaller number of labels, allowing the models to specialize more in a limited number of classes.

A crucial aspect that emerged during this project is the importance of having a clean and balanced dataset, so that the model can learn as much information as possible in an equitable manner for each class. In this way, the utility of the designated classifier could find greater resonance in the clinical field, as if the initial steps already show excellent performance, the clinician's work would be further eased, allowing the first phase of alteration identification to be managed by an automatic classification tool. The availability of a carefully labeled and balanced dataset thus implies a necessary data cleaning phase, with the aim of making the

training dataset homogeneous while maintaining the different diagnostic methodologies adopted by clinicians depending on the country of origin. This aspect is particularly relevant, as some articles published by groups participating in the Cardiology Challenge 2020 have highlighted how some of the 27 considered alterations present different diagnostic criteria depending on the country of origin. For example, regarding left axis deviation (LAD), the presence of QRS complex inversion in lead II is a requirement for diagnosis in UK manuals, but not in some Chinese texts. [10] Despite the low representativeness of some classes, the decomposition of the model into two parts ensures good accuracy in the first step. The direct comparison with the model that achieved the best performance shows that the results on the most represented individual labels are in line, with better performance on the healthy class, especially in the hidden test set as shown in Fig. 4.4. The presented results in Fig. 4.3 and Fig. 4.4 are promising: although there is a slight decrease in the F1 metric for almost all classes using the ensemble model on the validation set, the performance on the test set is better. This highlights a greater generalization capability of the model. Consequently, it can be stated that the solution consisting of the ensemble and the cascade of two models, unlike the simpler model, has demonstrated superior generalization capability on unseen data. This is a crucial objective in the development of automatic classification systems in the clinical field. The integration of clinical knowledge into the project has had a significant and positive impact on the overall performance of the model, demonstrating the importance of a multidisciplinary approach in the design of complex classification systems. In particular, the choice to group certain classes and exclude one resulted in a noticeable improvement in performance, as shown in Tab 4.9. This strategy confirmed that the poor representativeness of some classes negatively affects the performance of the cascade ensemble model. This finding highlights how the accuracy of a classification model depends not only on the algorithm used but also on the correct selection and aggregation of classes, processes that require a deep understanding of the clinical application domain. Similarly, the decision to reduce the duration of the signals to improve the recognition of electrocardiographic alterations, characterized by repetitiveness along the entire trace, did not produce the desired results in terms of performance, as shown in Tab 4.8. The intent was to focus feature extraction on a single beat, with the aim of identifying distinctive traits capable of discriminating between different classes. This strategy, although based on the clinical logic that some anomalies manifest repetitively and are therefore easily identifiable by clinicians observing a single P-S complex, proved insufficient due to the random nature with which other anomalies may appear in a single beat. The integration of clinical knowledge has proven crucial for refining the model, increasing its sensitivity to the peculiarities of each class and improving its discriminative ability. This result underscores the importance of the contribution of clinical experts in multi-label classification projects in

the cardiology field, where the complexity of the task requires a deep understanding of the specific characteristics of different pathologies. The inclusion of such knowledge has facilitated the understanding of the distinctive peculiarities of the classes, allowing informed decisions regarding their aggregation or exclusion from the model, and demonstrating how an interdisciplinary collaborative approach can lead to significant improvements in the performance of deep learning models.

Chapter 6

Conclusion

In the course of this thesis, it has emerged that an automatic support capable of recognizing and identifying alterations in the electrocardiographic signal can constitute a valuable additional tool for clinicians. The usefulness of a classification model lies in its ability to accelerate diagnoses and lighten the workload of specialists, but it is equally important to emphasize that such solutions must offer adequate performance to minimize mislabeling, effectively serving as an aid for experts in the field.

The results obtained with the binary model proposed in this thesis are encouraging: the model demonstrated an accuracy of about 92% in recognizing the presence of anomalies in an ECG trace, maintaining such high performance even on new and unseen data, which highlights a good generalization ability. However, although these results are promising, the results of the 27-class classification, although in line with those obtained by the best participants in the 2020 Cardiology Challenge, show an accuracy of about 50% on an unseen test set. This underscores the complexity of the challenge in creating a model capable of distinguishing between very similar classes or evaluating signals belonging to the same class but characterized by different parameters and features.

The successful completion of the classification task and, therefore, the effectiveness of the developed model depend on several factors, including the choice of model architecture, parameter selection, and the quality of the data used. Errors in the labels assigned during the labeling phase or misclassified signals can negatively affect the model's performance. The main objective of this thesis was to develop a classifier capable of distinguishing between 27 electrocardiographic alterations using databases from different sources, in order to ensure high accuracy and good generalization ability.

In this context, it is crucial to have a clean and accurate training dataset that takes into account the variability of sources and the different sizes of the data, especially in multi-label and multi-source tasks. It is essential to avoid the deep

learning model from learning specific characteristics of individual alterations from larger datasets, as this could compromise the model's ability to generalize to new data. Additionally, differences in the reliability of training labels may not be random but depend on the specific dataset.

A possible approach to address this issue is the creation of models capable of accounting for differences between datasets, a methodology known as domain adaptation. This approach was proposed by the fifth-place participants [21] in the 2020 Cardiology Challenge and implemented through a technique called "Joint Learning and Unlearning". [4] The use of an adversarial multi-task approach, aimed at simultaneously minimizing the accuracy in domain prediction and maximizing task accuracy, has proven effective in MRI image segmentation problems [42], thus suggesting the opportunity to explore the application of this methodology to ECG data as well.

The combination of these elements, from the rigorous selection of training data to the exploration of advanced learning techniques, represents a promising path towards the development of increasingly accurate and generalizable models, capable of effectively supporting clinicians in the diagnosis and treatment of electrocardiographic anomalies.

In this perspective, it might be useful to consider creating a clean dataset, possibly supervised by clinicians, that contains correctly labeled signals and is homogeneously representative of all alterations, in order to increase the performance of the subsequent model. This would allow for a more precise solution to be cascaded to the binary model, while avoiding a substantial decrease in performance. As emerged during this thesis, another crucial aspect concerns the need for close interdisciplinary collaboration between engineers, data scientists, and clinicians. This cooperation is essential not only to ensure the quality and reliability of the dataset but also to ensure that the decisions made by the model are clinically relevant and easily interpretable by healthcare professionals. Involving clinical experts in the process of labeling and validating ECG signals could significantly improve the quality of the training data and, consequently, the model's performance.

Furthermore, it is important to consider how the developed model can be integrated into clinical decision support systems, capable not only of accurately classifying electrocardiographic alterations but also of offering transparent and understandable explanations of the predictions made. This would help to promote the adoption of such technologies in the clinical field, improving doctors' trust in artificial intelligence solutions. Finally, validating the model in real clinical environments is a crucial step to verify its effectiveness on unseen data from different patient populations. This continuous validation process is essential to ensure that the model not only maintains high levels of accuracy but also meets clinical standards and the expectations of professionals in the field.

Bibliography

[1] PhysioBank, PhysioToolkit, and PhysioNet components of a new research resource for complex physiologic signals, 2000.

[2] U.R. Acharya, H. Fujita, S.L. Oh, Y. Hagiwara, J.H. Tan, M. Adam, and R.S. Tan. Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals. *Applied Intelligence*, 49:16–27, 2018.

[3] Erick A. Perez Alday, Annie Gu, Amit Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Qiao Li, Ashish Sharma, Gari D. Clifford, and Matthew A. Reyna. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. In *Proceedings of the PhysioNet/Computing in Cardiology Challenge*, Online, 2020. PhysioNet.

[4] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[5] Zachary D. Goldberger Ary L. Goldberger and Alexei Shvilkin. Goldberger's Clinical Electrocardiography. https://www.sciencedirect.com/book/9780323401692/goldbergers-clinical-electrocardiography, 2017.

[6] Zachi I. Attia, Paul A. Friedman, Peter A. Noseworthy, Francisco Lopez-Jimenez, Dorothy J. Ladewig, Gaurav Satam, Patricia A. Pellikka, Thomas M. Munger, Samuel J. Asirvatham, Christopher G. Scott, Rickey E. Carter, and Suraj Kapa. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. *Circulation: Arrhythmia and Electrophysiology*, 12, 9 2019.

[7] H. Benhar, A. Idri, and J. L Fernández-Alemán. Data preprocessing for heart disease classification: A systematic literature review. *Computer Methods and Programs in Biomedicine*, 195:105635, 10 2020.

[8] Carl Boettiger. An introduction to docker for reproducible research, with examples from the R environment. 10 2014.

[9] Jonathan James Hyett Bray, Elin Fflur Lloyd, Firdaus Adenwalla, Sarah Kelly, Kathie Wareham, and Julian P.J. Halcox. Single-lead ECGs (alivecor) are a feasible, cost-effective and safer alternative to 12-lead ECGs in community diagnosis and monitoring of atrial fibrillation. *BMJ Open Quality*, 10, 3 2021.

[10] WB Chen, XL Pan, XH Wan, XF Lu, C Liu, SJ Hu, XX Kang, and J Yang. *Diagnosis*. People's Medical Publishing House, People's Republic of China, 8th edition, 2013.

[11] Zahra Ebrahimi, Mohammad Loni, Masoud Daneshtalab, and Arash Gharehbaghi. A review on deep learning methods for ECG arrhythmia classification. 2020.

[12] Enes Efe and Emrehan Yavsan. AttBiLFNet: A novel hybrid network for accurate and efficient arrhythmia detection in imbalanced ECG signals. *Mathematical Biosciences and Engineering*, 21:5863–5880, 2024.

[13] M. I. Ferrer. The significance of axis deviation., 1972.

[14] Sara Fezzotti. Analisi della variabilità del ritmo cardiaco per il monitoraggio del sistema nervoso in condizioni di stress. Master's thesis, Università Politecnica delle Marche, Ancona, Italy, 2023.

[15] Anne Groot, Michiel L. Bots, Frans H. Rutten, Hester M. Den Ruijter, Mattijs E. Numans, and Ilonca Vaartjes. Measurement of ECG abnormalities and cardiovascular risk classification: A cohort study of primary care patients in the Netherlands. *British Journal of General Practice*, 65:e1–e8, 1 2015.

[16] Ma Guanglong, Wang Xiangqing, and Yu Junsheng. ECG signal classification algorithm based on fusion features. In *Journal of Physics: Conference Series*, volume 1207. Institute of Physics Publishing, 4 2019.

[17] Utkarsh Gupta, Naveen Paluru, Deepankar Nankani, Kanchan Kulkarni, and Navchetan Awasthi. A comprehensive review on efficient artificial intelligence models for classification of abnormal cardiac rhythms using electrocardiograms. *Heliyon*, 10, 3 2024.

[18] Cecilia Gutierrez and Daniel G Blanchard. Atrial fibrillation: Diagnosis and treatment, 2011.

[19] John R Hampton. ECG FACILE traduzione della 5 edizione inglese di: The ECG made easy.

[20] A.Y. Hannun, P. Rajpurkar, M. Haghpanahi, G.H. Tison, C. Bourn, M.P. Turakhia, and A.Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69, 2019.

[21] Hosein Hasani, Adeleh Bitarafan, and Mahdieh Soleymani Baghshah. Classification of 12-lead ECG signals with adversarial multi-source domain generalization. In *Computing in Cardiology*, volume 2020-September. IEEE Computer Society, 9 2020.

[22] Robert Herman, Anthony Demolder, Boris Vavrik, Michal Martonak, Vladimir Boza, Viera Kresnakova, Andrej Iring, Timotej Palus, Jakub Bahyl, Olivier Nelis, Monika Beles, Davide Fabbricatore, Leor Perl, Jozef Bartunek, and Robert Hatala. Validation of an automated artificial intelligence system for 12-lead ECG interpretation. *Journal of Electrocardiology*, 82:147–154, 1 2024.

[23] Jelle C.L. Himmelreich and Ralf E. Harskamp. Diagnostic accuracy of the pmcardio smartphone application for artificial intelligence–based interpretation of electrocardiograms in primary care (AMSTELHEART-1). *Cardiovascular Digital Health Journal*, 4:80–90, 6 2023.

[24] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7132–7141. IEEE Computer Society, 12 2018.

[25] Rui Duarte Id, Angela Stainthorpe Id, James Mahon, Janette Greenhalgh Id, Marty Richardson Id, Sarah Nevitt Id, Eleanor Kotas, Angela Boland, Howard Thom, Tom Marshall Id, Mark Hall Id, and Yemisi Takwoingi Id. Lead-I ECG for detecting atrial fibrillation in patients attending primary care with an irregular pulse using single-time point testing: A systematic review and economic evaluation. 2019.

[26] Selina Jarvis and Selva Saman. Cardiac system 1: Anatomy and physiology. *Margate Health Consortium*, 2024.

[27] Yanrui Jin, Zhiyuan Li, Mengxiao Wang, Jinlei Liu, Yuanyuan Tian, Yunqing Liu, Xiaoyang Wei, Liqun Zhao, and Chengliang Liu. Cardiologist-level interpretable knowledge-fused deep neural network for automatic arrhythmia diagnosis. *Communications Medicine*, 4, 2 2024.

[28] Ho Keun Kim and Myung Hoon Sunwoo. An automated cardiac arrhythmia classification network for 45 arrhythmia classes using 12-lead electrocardiogram. *IEEE Access*, 12:44527–44538, 2024.

[29] Paul Kligfield. The centennial of the Einthoven electrocardiogram, 2002.

[30] Bradley P. Knight, Gregory F. Michaud, S. Adam Strickberger, and Fred Morady. Electrocardiographic differentiation of atrial flutter from atrial fibrillation by physicians. *Journal of Electrocardiology*, 32:315–319, 10 1999.

[31] Weiqin Lin, Swee Guan Teo, and Kian Keong Poh. Electrocardiographic T wave abnormalities. *Singapore Medical Journal*, 54:606–610, 11 2013.

[32] Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, Jianqing Li, and Eddie Ng Yin Kwee. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8:1368–1373, 8 2018.

[33] Marios Loukas, Pamela Youssef, Jerzy Gielecki, Jerzy Walocha, Kostantinos Natsis, and R. Shane Tubbs. A glimpse of our past: History of cardiac anatomy: A comprehensive review from the Egyptians to today. *Journal of Cardiac History*, 2024.

[34] David M. Mirvis and Ary L. Goldberger. Electrocardiography. In *Fundamental Principles*, pages 126–165. Publisher Name, City, Country, 2024.

[35] Annamalai Natarajan, Yale Chang, Sara Mariani, Asif Rahman, Gregory Boverman, Shruti Vij, and Jonathan Rubin. A wide and deep transformer neural network for 12-lead ECG classification. In *Computing in Cardiology*, volume 2020-September. IEEE Computer Society, 9 2020.

[36] Suraj K. Nayak, Arindam Bit, Anilesh Dey, Biswajit Mohapatra, and Kunal Pal. A review on the nonlinear dynamical system analysis of electrocardiogram signal, 2018.

[37] Maximilian P. Oppelt, Maximilian Riehl, Felix P. Kemeth, and Jan Steffan. Combining scatter transform and deep neural networks for multilabel electrocardiogram signal classification. In *Computing in Cardiology*, volume 2020-September. IEEE Computer Society, 9 2020.

[38] Gabriel Ott, Yannik Schaubelt, Juan Miguel Lopez Alcaraz, Wilhelm Haverkamp, and Nils Strodthoff. Using explainable AI to investigate electrocardiogram changes during healthy aging-from expert features to raw signals. *PLoS ONE*, 19, 4 2024.

[39] JaeYeon Park, Kichang Lee, Noseong Park, Seng Chan You, and JeongGil Ko. Self-Attention LSTM-FCN model for arrhythmia classification and uncertainty assessment. *Applied Soft Computing*, 132:110797, 2023.

[40] Ziran Peng and Guojun Wang. Study on optimal selection of wavelet vanishing moments for ECG denoising. *Scientific Reports*, 7, 12 2017.

[41] Matthew A. Reyna, Nadi Sadr, Erick A. Perez Alday, Annie Gu, Amit J. Shah, Chad Robichaux, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Sardar Ansari, Hamid Ghanbari, Qiao Li, Ashish Sharma, and Gari D. Clifford. Issues in the automated classification of multilead ECGs using heterogeneous labels and populations. *IEEE Journal of Biomedical and Health Informatics*, 24(12):3543–3551, 2021.

[42] Z. Shang, Z. Zhao, H. Fang, S. Relton, D. Murphy, Z. Hancox, R. Yan, and D. Wong. Deep discriminative domain generalization with adversarial feature learning for classifying ECG signals. In *2021 Computing in Cardiology (CinC)*, pages 1–4, Piscataway, NJ, 2021. IEEE.

[43] Reena Tiwari, Ravindra Kumar, Sujata Malik, Tilak Raj, and Punit Kumar. Analysis of heart rate variability and implication of different factors on heart rate variability. *Current Cardiology Reviews*, 17, 1 2021.

[44] Willis J Tompkins. A real-time QRS detection algorithm, 1985.

[45] Kuo Kun Tseng, Jiaqian Li, Yih Jing Tang, Ching Wen Yang, and Fang Ying Lin. Healthcare knowledge of relationship between time series electrocardiogram and cigarette smoking using clinical records. *BMC Medical Informatics and Decision Making*, 20, 7 2020.

[46] Patrick Wagner, Nils Strodthoff, Ralf Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7, 12 2020.

[47] Chandra Wijaya, Andrian, Mawaddah Harahap, Christnatalis, Mardi Turnip, and Arjon Turnip. Abnormalities state detection from P-Wave, QRS Complex, and T-wave in noisy ECG. In *Journal of Physics: Conference Series*, volume 1230. Institute of Physics Publishing, 9 2019.

[48] Yong Xia, Yueqi Xiong, and Kuanquan Wang. A transformer model blended with CNN and denoising autoencoder for inter-patient ECG arrhythmia classification. *Biomedical Signal Processing and Control*, 105:105271, 2023.

[49] Can Ye, B. V.K. Vijaya Kumar, and Miguel Tavares Coimbra. Heartbeat classification using morphological and dynamic features of ECG signals. *IEEE Transactions on Biomedical Engineering*, 59:2930–2941, 2012.

[50] Z. Zhao, D. Murphy, H. Gifford, S. Williams, A. Darlington, S. D. Relton, H. Fang, and D. C. Wong. Analysis of an adaptive lead weighted ResNet for multiclass classification of 12-lead ECGs. *Physiological Measurement*, 43, 3 2022.

[51] Zhibin Zhao, Hui Fang, Samuel D. Relton, Ruqiang Yan, Yuhong Liu, Zhijing Li, Jing Qin, and David C. Wong. Adaptive lead weighted ResNet trained with different duration signals for classifying 12-lead ECGs. In *Computing in Cardiology*, volume 2020-September. IEEE Computer Society, 9 2020.

[52] Zhaowei Zhu, Xiang Lan, Tingting Zhao, Yangming Guo, Pipin Kojodjojo, Zhuoyang Xu, Zhuo Liu, Siqi Liu, Han Wang, Xingzhi Sun, and Mengling Feng. Identification of 27 abnormalities from multi-lead ECG signals: An ensembled SEResNet framework with sign loss function. *Physiological Measurement*, 42, 6 2021.

[53] Zhaowei Zhu, Han Wang, Tingting Zhao, Yangming Guo, Zhuoyang Xu, Zhuo Liu, Siqi Liu, Xiang Lan, Xingzhi Sun, and Mengling Feng. Classification of cardiac abnormalities from ECG signals using SEResNet. In *Computing in Cardiology*, volume 2020-September. IEEE Computer Society, 9 2020.