Classification and Analysis of 12-Lead Electrocardiograms

by

Alexander William Wong

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

© Alexander William Wong, 2021

Abstract

The electrocardiogram is the standard tool for detecting cardiac abnormalities, such as atrial fibrillation, irregular complexes, and heart blocks. However, the interpretation of this data is an unsolved problem with discrepancies among panels of cardiologists and automated analysis requiring additional human over-reading. This thesis explores the classification of 12-lead ECGs to a set of 27 diagnoses as defined in the *PhysioNet/CinC 2020* Challenge.

I propose three approaches, starting with manual feature engineering and classification using shallow gradient boosted tree ensembles. Our second approach uses a deep learning approach by combining fixed and variable length autoencoders to learn the features, followed by a multi layer perceptron (MLP) classifier. Our third approach combines the deep autoencoders and our shallow decision tree ensembles by training the shallow gradient boosted trees with both the manually extracted features as well as the bottleneck dimension representation of the 12-lead ECG record. I empirically evaluate our different approaches using a weighted classification scoring function using repeated random subsampling of the publicly available challenge dataset. This thesis concludes with future ways to approach the multi-channel signal classification problem that addresses some of the limitations discovered in the prior approaches. Our best model, using the averaged top 1000 manually engineered features with autoencoder embeddings, attains a mean test split challenge metric of 0.4366 with an overall mean classification accuracy of 30.7%.

Preface

Chapter 3 contains an adapted version of “Multilabel 12-Lead Electrocardiogram Classification Using Gradient Boosted Tree Ensemble” [55], published in *Computing in Cardiology (CinC) 2020* conference under the PhysioNet Challenge track. This work contains manuscript revisions from Dr. Abram Hindle and Dr. Sunil Vasu Kalmady.

Chapter 4 contains an adapted version of “Multilabel 12-Lead Electrocardiogram Classification Using Beat to Sequence Autoencoders” [56], submitted to the *International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2021* and has manuscript revisions from Dr. Abram Hindle and Dr. Sunil Vasu Kalmady. Figure 4.1, which showcases an overview of the methodology, is contributed by Amir Salimi.

Acknowledgements

I am grateful for Dr. Abram Hindle, for his invaluable mentoring, guidance, and support throughout my academic journey. I am grateful to Dr. Nilanjan Ray for his critical and insightful advice regarding meta analysis of neural networks. I am thankful to Dr. Eleni Stroulia, who contributed her time and efforts in getting me accepted into the program. I appreciate the support of Dr. Padma Kaul, who welcomed me to the Canadian VIGOUR Centre (CVC) and facilitated my transition into cardiac research.

This research is supported in part by an Alberta Excellence Graduate Scholarship and by a Graduate Research Assistant Fellowship provided by the CVC.

I am grateful to my parents, Andrew and Amy, for the privileged life and many opportunities that I have as a result of their hardship and sacrifices. I am especially thankful for my partner, Bonnie, who has provided endless support and instilled in me the three traits that continue to guide my life: to remain genuine; to be ambitious; and to stay grateful.

Contents

1 Introduction **1**
1.1 Contributions . . . 2
1.2 Thesis Organization . . . 3

2 Background Information **4**
2.1 Cardiac Physiology . . . 4
2.1.1 Anatomy and Electrical Conduction System . . . 4
2.1.2 Electrocardiogram Tracing . . . 6
2.1.3 Electrode Placement and Cardiac Axis . . . 7
2.2 PhysioNet/CinC 2020 Challenge Overview . . . 9
2.2.1 Public Dataset . . . 9
2.2.2 Diagnosed Labels . . . 10
2.2.3 Scoring Function . . . 27
2.3 Related Work . . . 28

3 Gradient Boosting Tree Ensemble **34**
3.1 Introduction . . . 36
3.1.1 Dataset & Scoring Criteria . . . 36
3.2 Methodology . . . 36
3.2.1 Signal Pre-processing . . . 37
3.2.2 Feature Engineering . . . 38
3.2.3 Classification . . . 39
3.3 Results . . . 40
3.4 Discussion & Future Work . . . 42
3.5 Conclusion . . . 43

4 Beat to Sequence Autoencoders **44**
4.1 Introduction . . . 46
4.2 Related Work . . . 47
4.2.1 Challenge Dataset and Task Specification . . . 47
4.3 Methodology . . . 47
4.3.1 Signal Preprocessing . . . 48
4.3.2 Heartbeat Autoencoder . . . 50
4.3.3 Embedding Sequence Autoencoder and Classifier . . . 51
4.4 Results and Discussion . . . 53
4.5 Conclusion . . . 54

5 Autoencoder Embeddings with Improved Tree Ensemble **55**
5.1 Methodology . . . 56
5.2 Results . . . 58
5.2.1 Averaged vs Labelwise Feature Selection . . . 59
5.2.2 Effectiveness of Feature Selection . . . 60
5.2.3 Adding Embeddings vs Without Embeddings . . . 61

5.3 Discussion 62
5.4 Conclusion 64

6 Conclusion **67**
6.1 Future Work 68

References **70**

Appendix A XGBoost Classifiers with Autoencoder Embeddings **76**
A.1 Label-wise Feature Importances 76
A.2 Label-wise F_1 Scores 76

List of Tables

2.1 Evaluated SNOMED CT codes with definition, count and percentage in dataset. 11

4.1 Evaluated labels, count and percentage in dataset. 48
4.2 Heartbeat autoencoder neural network architecture. 51
4.3 Sequence autoencoder and classifier architecture. 52

5.1 Test split classification metrics mean ($\bar{x}$) and standard deviations (σ) for all experiment configurations. Bolded value indicates largest mean for metric category. 59

List of Figures

2.1 Anterior, anatomical view of the structures of a human heart. 5
2.2 Anterior, anatomical view of the conduction system of a human heart. 6
2.3 Normal sinus rhythm electrocardiogram (ECG) tracing with PQRST peaks annotated. 7
2.8 Instance of ECG record with 1st degree av block. Long PR interval greater than 200ms is observed. 10
2.9 Instance of 12-lead ECG with atrial fibrillation. Irregular, fibrillatory p-waves found in all leads. 12
2.10 Instance of 12-lead ECG with atrial flutter. Rapid P waves exist. 12
2.11 Instance of 12-lead ECG with bradycardia. Patient has a mean resting heart rate of 57 beats per minute. 13
2.12 Instance of 12-lead ECG with complete right bundle branch block. QRS complex exceeds 120 ms (4 boxes) and contains an anterior skew on the right end of the QRS complex. Characteristic “bunny ears” trait emphasized in lead V2. 13
2.13 Instance of 12-lead ECG with incomplete right bundle branch block. The QRS complex is within 100 to 120 ms (2.5-3 boxes) and contains an anterior skew on the right end of the QRS complex. 14
2.14 Instance of 12-lead ECG with left anterior fascicular block. Lead I is positive while leads II and III are negative, indicating left axis deviation. Leads I and aVL show qR complexes, with rS complexes appearing in leads II, III, an aVF. 15
2.15 Instance of 12-lead ECG with left axis deviation. Lead I is positive while leads II and III are negative, indicating left axis deviation using the simple leads I, II, III method. Using the hexaxial axis estimation approach, we see that leads aVL and III have the greatest QRS complex magnitude with aVL being positive and III being negative, giving an electrical axis between $-30°$ and $-60°$, confirming left axis deviation. 16
2.16 Instance of 12-lead ECG with left bundle branch block. The QRS complexes exceed 100ms (2.5 boxes) and contain a posterior or leftward skew on the tail halves of the QRS complexes. 16
2.17 Instance of 12-lead ECG with low QRS voltages. All limb leads have QRS complex amplitudes of less than 0.5mV and all precordial leads have QRS complex amplitudes of less than 1.0mV. 17
2.18 Instance of 12-lead ECG with nonspecific intraventricular conduction disorder. QRS complexes exceed 100ms (2.5 squares) but do not exhibit a posterior/anterior skew characteristic on the tail half of the QRS complexes. 18

2.19 Instance of 12-lead ECG with pacing rhythm. Characteristic vertical lines appearing at the P wave onset suggests atrial pacing with normal conduction to the ventricles. 19
2.20 Instance of 12-lead ECG with premature atrial contraction. Instances of P waves occurring within preceding beat T wave in lead II are emphasized with black squares. 19
2.21 Instance of 12-lead ECG with premature ventricular contraction. Premature ventricular contraction occurs at the 4.5 second mark. 20
2.22 Instance of 12-lead ECG with prolonged PR interval. Patient record contains PR intervals exceeding 200ms (5 squares) but is also diagnosed as having normal sinus rhythm. 21
2.23 Instance of 12-lead ECG with prolonged QT interval. Patient T waves end past the halfway point of the preceding and upcoming R peaks. 21
2.24 Instance of 12-lead ECG with abnormal Q wave. Patient Q waves exceed 40ms wave duration in leads I, III, aVR, aVL, V2, V3, V4, V5, V6 and have Q wave amplutides exceeding 25% of the QRS complex amplitude in precordial leads V2-V6. 22
2.25 Instance of 12-lead ECG with right axis deviation. A negative lead I with positive leads II and III suggest right axis deviation. Using the hexaxial system, leads II, III, and aVF all have the largest QRS amplitudes (each approximately 0.3mV), resulting in an cardiac axis estimate between 60° and 120°. This overlaps with our right axis deviation criteria where the axis exists between 90° and 180°. 23
2.26 Instance of 12-lead ECG with sinus arrhythmia. Recorded heart rate decreases over time, or distance between R peaks increases over time. 23
2.27 Instance of 12-lead ECG with sinus bradycardia. Upright P wave in lead II exists, with a mean heart rate of 59 beats per minute. 24
2.28 Instance of 12-lead ECG with normal sinus rhythm. No visible defects or abnormalities indicated. 25
2.29 Instance of 12-lead ECG with sinus tachycardia. The average heart rate is 101 beats per minute. 25
2.30 Instance of 12-lead ECG with supraventricular premature beats. Four instances of premature ventricular beats are observed in this record. 26
2.31 Instance of 12-lead ECG with abnormal T wave. 26
2.32 Instance of 12-lead ECG with T wave inversion. Leads with T waves of incorrect direction are observed in leads II, III, aVR (should be inverted), aVF, and V3-6. 27
2.4 Cardiac conduction correlated to ECG tracing waves. 30
2.5 Placement of 12-lead ECG for precordial electrodes V1-V6. . . 31
2.6 Hexaxial (Cabrera) system for determining cardiac axis. Six frontal planes are viewed in the sequence aVL, I, aVR, II, aVF, and III. 32
2.7 Age and sex distribution of the PhysioNet/CinC 2020 public electrocardiogram dataset. 32
2.33 PhysioNet/CinC 2020 Challenge [40] provided reward matrix W for the diagnoses evaluated in the classification task. 33

3.1 Overview of manual feature engineering and gradient boosting decision tree classifier methodology. 37

3.2 Decision tree derived counts of important features by feature category. 40
3.3 Manual feature engineering with gradient boosting decision tree classification metrics summary on validation set. 41
3.4 Manual feature engineering with gradient boosting decision tree label-wise F_1 scores on validation set. 41

4.1 Methodology overview. From (1) $\mathcal{L}$-sized ECG signal, we (2) extract $\mathcal{M}$ heartbeat windows prior to (3) heartbeat & (4) sequence autoencoding. Finally we (5) pass our sequence embedding to the classifier to output our 27 predictions. 46
4.2 Evaluation scoring function weights per label. 49
4.3 Signal processing from the raw signal to the annotated intermediary signal, to output heartbeat windows. Window 4 is dropped due to the cutoff threshold. Window 6 is dropped due to insufficient window size. Window 2 is the most abnormal heartbeat with the minimum outlier factor of the given windows. Only 3 of the 12 leads (`II`, `V3`, and `V6`) are shown for clarity. 50
4.4 Test set F_1 scores of all 27 labels compared with our prior XGBoost ensemble method. Mean values annotated. 51
4.5 Classification metric summary of 20 experiments compared to XGBoost ensembles. Mean values annotated. 52

5.1 Test split classification metrics of XGB ensemble using all configurations of engineered features and autoencoder embeddings. 65
5.2 Wilcoxon signed-rank test P-values, comparing test split challenge metric distributions of all 6 experiment configurations. . 66

A.1 Feature importances of Configuration 4: “Top 1000 Features with Embeddings”, utilization over 20 experiments. The 27 diagnosed labels are displayed separately, showcasing feature derived lead (alternatively Meta or Autoencoder), and category. 77
A.2 Feature importances of Configuration 6: “Top 100 Features with Embeddings”, utilization over 20 experiments. The 27 labels are displayed separately, showcasing feature derived lead (alternatively Meta or Autoencoder), and category. 78
A.3 Test split label-wise F_1 scores for each of the 10 configurations over the 20 independent repeated random subsampling iterations. 79

Glossary

1st degree av block (IAVB)

Diagnosis; a delay in the electrical impluses from the atria, through the atrioventricular node, to the ventricles.

atrial flutter (AFL)

Diagnosis; the occurrence of rapid beats of the upper chambers (atria) of the heart.

atrial fibrillation (AF)

Diagnosis; the occurrence of chaotic or irregular beats of the upper chambers atria of the heart.

bradycardia (Brady)

Diagnosis; sinus rhythm is below the normal range relative to patient age, typically under 60 beats per minute in adults.

complete right bundle branch block (CRBBB)

Diagnosis; full action potential block to right bundle branch, QRS duration exceeding 120 ms with QRS complex rightward skew.

electrocardiogram (ECG)

A non-invasive tool for measuring the electrical activity of the heart.

incomplete right bundle branch block (IRBBB)

Diagnosis; a delay or blockage along the right side pathway that electrical impulses travel to trigger a heart beat.

left bundle branch block (LBBB)

Diagnosis; action potential block to left bundle branch, QRS duration exceeding 100ms with QRS complex leftward skew.

left axis deviation (LAD)

Diagnosis; cardiac axis exists between -30° and -90°.

left anterior fascicular block (LAnFB)

Diagnosis; a defect in the anterior half of the left bundle branch, related but distinct from left bundle branch block.

low QRS voltages (LQRSV)

Diagnosis; QRS complex amplitudes < 0.5 mV in all limb leads and < 1 mV in all precordial leads.

nonspecific intraventricular conduction (NSIVCB)

Diagnosis; QRS complex durations over 100ms without anterior/posterior skew characteristic.

pacing rhythm (PR)

Diagnosis; cardiac pacing stimuli is delivered using external means, such as a pace maker.

premature atrial contraction (PAC)

Diagnosis; atrium beating prematurely, P wave occuring within the T wave of the preceding beat.

premature ventricular contractions (PVC)

Diagnosis; ventricular contraction occurred prior to expected sinoatrial node action potential.

prolonged QT interval (LQT)

Diagnosis; inadequate recovery/repolarization of the heart after each beat, T waves ending beyond the midway point of an RR interval.

prolonged PR interval (LPR)

Diagnosis; delayed conduction through the atrioventricular node, PR interval exceeding 200ms.

Q wave abnormal (QAb)

Diagnosis; Q wave duration exceeds 40 ms or amplitude exceeding 25% of the QRS complex amplitude.

right bundle branch block (RBBB)

Diagnosis; action potential block to right bundle branch, QRS duration exceeding 100ms with QRS complex rightward skew.

right axis deviation (RAD)

Diagnosis; the net direction of the depolarization wave of the heart is between $+90°$ to $+180°$.

sinus tachycardia (STach)

Diagnosis; sinus rhythm is above the normal range relative to patient age, typically over 100 beats per minute in adults.

sinus rhythm (SNR)

Diagnosis; normal, healthy function of the heart.

sinus bradycardia (SB)

Diagnosis; subtype of bradycardia, sinoatrial node firing fewer than 60 times per minute (in typical adults).

sinus arrhythmia (SA)

Diagnosis; change in the beat-to-beat variation over time, irregular heart rate.

supraventricular premature beats (SVPB)

Diagnosis; atrial contractions triggered through invalid conduction, such as non-sinoatrial node origin.

Systematized Nomenclature of Medicine (SNOMED)

A systematic, computer-processable set of medical terminology, definitions, and synonyms.

T wave inversion (TInv)

Diagnosis; T wave misaligned with QRS complex duration or not upright in leads I, II, V3-6 or not inverted in lead aVR.

T wave abnormal (TAb)

Diagnosis; T wave missing asymmetry, misaligned with QRS complex duration, low amplitude.

ventricular premature beats (VPB)

Diagnosis; see premature ventricular contractions (PVC).

Chapter 1

Introduction

Heart and cardiovascular diseases are the global leading cause of death, with 80% of cardiovascular disease related deaths due to heart attacks and strokes [51]. The electrocardiogram (ECG), when correctly interpreted, is the primary tool in our ongoing efforts to detect cardiac abnormalities and screen vulnerable members of our society for heart related issues [46]. An ECG works by recording electrical activity corresponding to the heartbeat muscle contractions using non-invasive electrodes placed on the surface of the skin [11]. Although computerized interpretations of ECGs are in widespread use, automated approaches have not yet matched the quality of an expert cardiologist reference, leading to poor patient outcomes or even fatality [13].

Multiple configurations of ECG machines exist ranging from consumer portable ECG devices such as the single lead AliveCor KardiaMobile and six lead KardiaMobile 6L variant [3], the single lead Apple Watch [5] and the three lead QardioCore devices [42], to the cardiologist focused devices built by General Electric [1] and Koninklijke Philips [2]. The focus of this research applies to the 12-lead ECG, as it is the standard hospital setting device used by cardiologists for evaluating heart disorders [28].

In this thesis, I discuss approaches for the multi-label, multi-class classification of ECG records using a combination of deep learning and traditional machine learning methods. I explore in-depth the following predictions:

- Despite the overwhelming popularity of deep learning classifiers, I predict that shallow learning methods such as gradient boosted decision trees can remain a viable and sensible choice for the ECG classification task, outperforming a deep learning autoencoder model on summary classification metrics such as F-measure and weighted accuracy (see Scoring Function, Section 2.2.3).

- When working with gradient boosted decision trees, I predict that regularization of the input feature space and appropriately selecting the important features for the classifier are more effective than incorporating deep learning autoencoder embeddings for improving the challenge classification score.
- I predict that naively joining deep learning autoencoder embeddings with manually engineered features for decision tree classifiers will improve the summary classification metrics in the ECG classification task.

1.1 Contributions

My contributions to this thesis include:

- I defined a methodology and engineered the experiment for the classification of 12-lead ECGs using manual feature extraction techniques and an ensemble of gradient boosting trees and publish a submission to the *PhysioNet/CinC 2020 Challenge* [40]. This attempt had an official phase challenge validation score of 0.476 and test score of −0.080, ranking our attempt at 36 of 41 successful entries (Chapter 3).
- I developed a deep learning approach using autoencoders to generate representations of the ECG heart beat and sequence of heart beat embeddings for the classification of 12-lead ECGs (Chapter 4). Because the official test set records are unavailable to the public, I utilize a monte carlo repeated random subsampling approach, running 20 experiments where the publicly available data is split into 80% training, 10% validation, and 10% testing sets. Our beat to sequence autoencoder classifiers attain an average test split challenge score of 0.248, with worse overall

classification performance compared to the shallow machine learning approach, but slightly improved label-wise specificity on incomplete right bundle branch block, left anterior fascicular block, pacing rhythm, and right axis deviation.

- I created a hybrid shallow/deep machine learning approach for 12-lead ECG classification by fusing together the manually engineered features with the autoencoder sequence embedding representation of the record. I fix the shortcomings of the prior challenge submission attempt, opting for feature selection for each diagnosis classifier rather than overall importances of all labels. The best approach, "Top 1000 Features with Embeddings", selects 1000 features by importance for each classifier and attains a test split challenge score of 0.4366.

1.2 Thesis Organization

This work is organized into the following chapters: Chapter 2 describes the characteristics of an ECG, the dataset of ECG records used in our analysis and algorithm training, and the different classification labels that our algorithm predicts probabilities for. Chapter 3 contains an approach for the classification of ECG records using manual feature extraction and a gradient boosted decision tree ensemble. Chapter 4 contains a deep learning classification approach using stacked autoencoders to learn an embedding representation of heartbeats and the ECG signal. Chapter 5 fuses the autoencoder and decision tree ensemble into one hybrid model and showcases the results in comparison to the prior two methods. Additional improvements are made to address shortcomings, notably in the feature selection process for the label-wise classifiers. Finally, Chapter 6 proposes future research directions and concludes the thesis.

Chapter 2

Background Information

In this chapter, I give a brief overview of the anatomy of a human heart. Next, I describe the characteristics of a standard 12-lead ECG and the notable waves in a ECG signal. I then give an overview of the *PhysioNet/CinC 2020 Challenge* task/objective, provided dataset of ECG records, and definitions for the diagnoses we are tasked to predict.

2.1 Cardiac Physiology

2.1.1 Anatomy and Electrical Conduction System

A high level overview of the primary valves and chambers within the heart can be found in Figure 2.1. The upper chambers of the heart, consisting of the right and left atriums, work in cooperation with the lower chambers of the heart, consisting of the left and right ventricles [4]. The right ventricle pushes blood into the pulmonary artery which connects to the lungs to return oxygenated blood [4]. The oxygenated blood returns to the heart through the pulmonary veins and enters into the left atrium [4]. The left atrium collects and pumps the oxygenated blood into the left ventricle through the mitral valve [4]. The left ventricle pumps the oxygenated blood out of the heart to the rest of the body through the aorta [4]. The deoxygenated blood is collected back into the heart through the superior and inferior vena cava and into the right atrium [4]. The cycle repeats as the right atrium pumps the blood into the right ventricle through the tricuspid valve, allowing the lungs to once again oxygenate the blood [4].

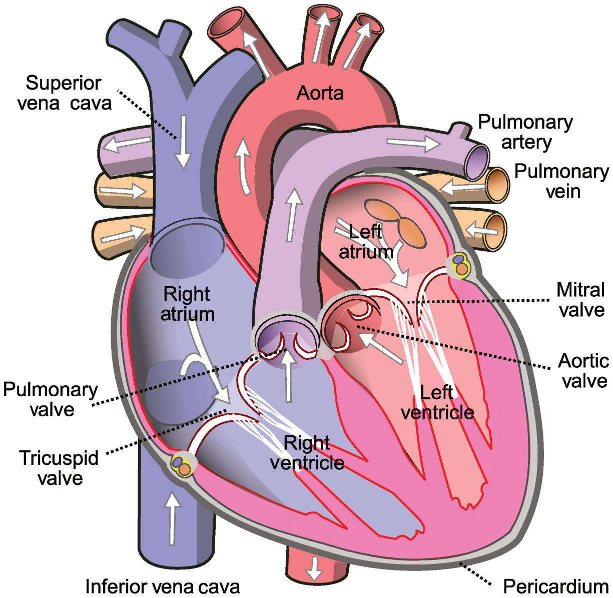


Figure 2.1: Anterior, anatomical view of a human heart. The primary valves and chambers of the heart are annotated, with arrows indicating the direction of blood flow due to the contractions of the cardiac chambers. Image licensed `CC BY-SA 3.0` from Wikipedia user Wapcaplet, source: `https://en.wikipedia.org/wiki/Heart#/media/File:Diagram_of_the_human_heart_(cropped).svg`.

Figure 2.2 shows an overview of the primary structures relevant to the cardiac conduction cycle.

Starting at rest the sinoatrial node, also known as "the pacemaker of the heart", initiates the sinus rhythm action potential which travels across the right and left atria [9]. Once the action potential reaches the atrioventricular node, "there is a delay of approximately 100ms that allows the atria to complete pumping blood before the impluse is transmitted to the atrioventricular bundle" [9]. Once the delay finishes, the impluse sweeps across the atrioventricular bundle, right and left bundle branches, and to the Purkinje fibers [9]. Ventricular contraction occurs, then the impulse dissipates at the contractile fibers of the ventricle causing the ventricles to repolarize in preparation for the next heartbeat [9].

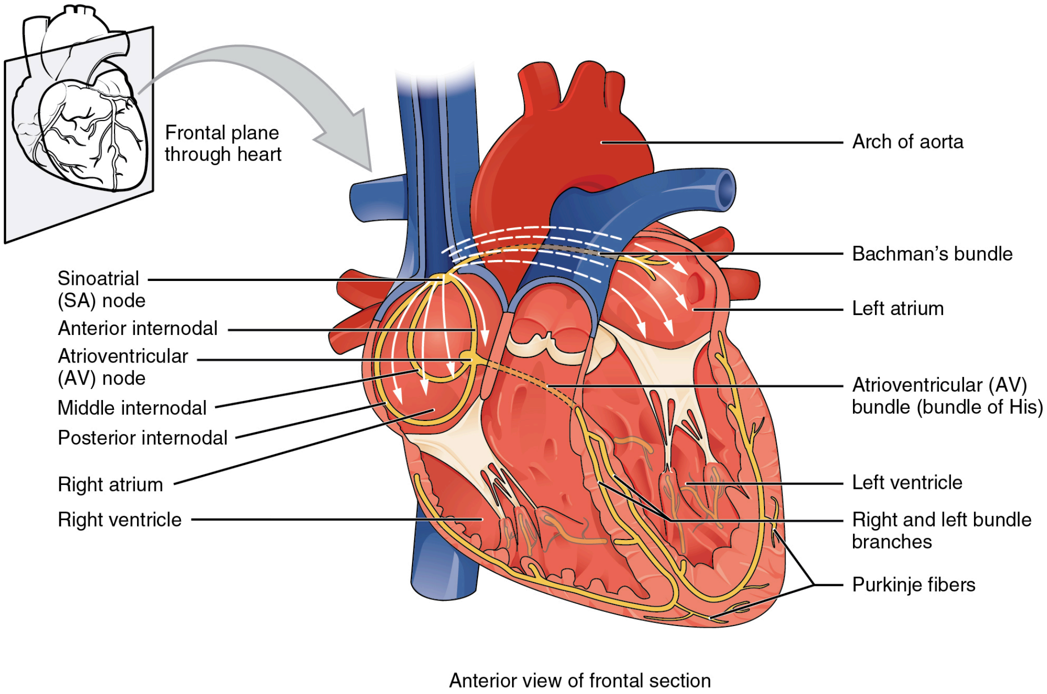


Figure 2.2: Anterior, anatomical view of the conduction system of a human heart. The conducting components of the heart begin with the sinoatrial node and include the internodal pathways, the atrioventricular node, the atrioventricular bundle, the right and left bundle branches, and the Purkinje fibres [9]. Image licensed `CC BY 4.0` from Betts *et al* [9] on the OpenStax platform, source: `https://openstax.org/books/anatomy-and-physiology/pages/19-2-cardiac-muscle-and-electrical-activity#fig-ch20_02_02`.

2.1.2 Electrocardiogram Tracing

Within a typical ECG, there are five peaks per beat, labeled PQRST respectively, that define the major components of a heartbeat as shown in Figure 2.3. The P wave represents the sinoatrial node initiating an impulse action potential and marks the start of a heartbeat [9]. The PR segment, which starts after the P wave and ends before the QRS complex, represents the delay between the atrial contraction and the propagation of the signal through the atrioventricular bundle [9]. The QRS complex is the notable large spike in the ECG, which represents the electrical impluse traveling through the atrioventricular bundle and bundle branches to the Purkinje fibers [9]. The ST segment, starting after the QRS complex and ending before the T wave begins, is the phase of the ECG where the actual ventricle contraction occurs [9]. Af-

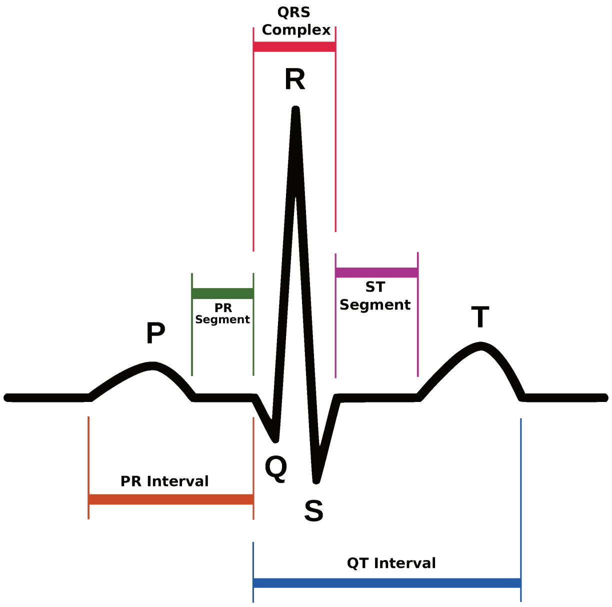


Figure 2.3: Normal sinus rhythm ECG tracing showing the PQRST peaks, the P wave, QRS complex, and T wave along with the PR and QT intervals, plus the PR and ST segments. Image belonging to `public domain` attributed to Anthony Atkielski, source: `https://en.wikipedia.org/wiki/Electrocardiography#/media/File:SinusRhythmLabels.svg`.

ter the ventricle contraction completes, the impulse dissipates and allows the ventricle muscles to relax and repolarize, manifesting as the ECG T wave [9]. A step by step diagram indicating the different ECG tracing components and the corresponding cardiac diagram can be found in Figure 2.4.

2.1.3 Electrode Placement and Cardiac Axis

The 12-lead ECG is derived from 10 electrodes placed on the surface of the skin [16]. The positioning of the leads is critical to ensure that the proper traces are recorded. See Figure 2.5 for a guide to placing the 6 precordial (front of the heart) leads V1 to V6. The remaining four leads are extremity or limb electrodes: Right Arm (RA) should be placed anywhere between the right shoulder and right elbow; Right Leg (RL) should be placed below the right torso and above the right ankle; Left Arm (LA) should be placed between the left shoulder and left elbow; and Left Leg (LL) should be placed below the left torso and above the left ankle [16].

When reading a 12-lead ECG, the leads I, II, II, aVR, aVL, and aVF are derived from the limb leads [23].

$$V_W = \frac{1}{3}(RA + LA + LL) \tag{2.1}$$

Equation 2.1 shows a common virtual electrode, known as the Wilson's central terminal, defined by averaging three of the limb leads (RA, LA, LL) together [23].

$$I = LA - RA \tag{2.2}$$

$$II = LL - RA \tag{2.3}$$

$$III = LL - LA \tag{2.4}$$

The unused limb lead RL does not show up in the ECG readings and is considered a neutral grounding lead for minimizing artifacts [16]. See Equation 2.2 for lead I, Equation 2.3 for lead II, and Equation 2.4 for lead III [23].

$$aVR = \frac{3}{2}(RA - V_W) = RA - \frac{1}{2}(LA + LL) \tag{2.5}$$

$$aVL = \frac{3}{2}(LA - V_W) = LA - \frac{1}{2}(RA + LL) \tag{2.6}$$

$$aVF = \frac{3}{2}(LL - V_W) = LL - \frac{1}{2}(RA + LA) \tag{2.7}$$

The augmented limb leads aVR, aVL, and aVF are derived from the same three electrodes as leads I, II, and III but rely on Wilson's central terminal as their negative pole. See Equation 2.5 for lead aVR, Equation 2.6 for lead aVL, and Equation 2.7 for lead aVF [23]. The remaining precordial leads V1 to V6 shown in the ECG are the directly measured signals from the electrodes.

An ECG's cardiac axis refers to the average direction of the wave of ventricular depolarization measured from the reference point of lead I on a standard 12-lead ECG [35]. One simple estimation of the cardiac axis is done by inspecting the magnitude of the R-peaks on leads I, II, and III [15]. In a normal ECG, leads I, II are positive while lead III may be positive or negative. There is right axis deviation if lead I is negative and lead III is positive (lead II may be positive or negative). Left axis deviation exists if lead I is positive while leads II and III are negative.

Alternatively, the Cabrera system or hexaxial reference system, can be used to logically derive the heart's electrical axis [30, 49]. By viewing the six frontal planes in the sequence aVL, I, aVR, II, aVF, and III, we check the maximal amplitude of the ECG vector (positive or negative) and use it to derive the cardiac electrical axis. Figure 2.6 shows the hexaxial reference system and the mapping to the six derived limb leads. A normal axis is a value between $-30°$ to $90°$, left axis deviation is between $-30°$ to -90, right axis deviation is between $90°$ to $180°$. Values that exceed the defined ranges are classified as extreme axis deviations.

2.2 PhysioNet/CinC 2020 Challenge Overview

This chapter summarizes the task of multi-label, multi-class classification of ECGs as proposed by Perez Alday *et al.* [40] in the *PhysioNet/CinC 2020 Challenge*.

2.2.1 Public Dataset

The challenge provided a public collection of 43,101 labelled ECG records for training. These public records were sourced from multiple locations, including:

1. The China Physiological Signal Challenge (**CPSC**) 2018 [31] corpus of data, containing 10,330 recordings.
2. The St. Petersburg Institute of Cardiological Technics (**INCART**) database of 12-lead arrhythmias [48], containing 74 recordings.
3. The Physikalisch Technische Bundesanstalt (PTB) contributed datasets **PTB** Diagnostic ECG Database [12] and the more recent **PTB-XL** database [53], containing a combined total of 22,353 records.
4. The Georgia 12-lead ECG Challenge (**G12EC**) database, containing 10,344 records.

A separate hold-out set of ECG records is sourced from an undisclosed organization containing patients geographically distinct from the publicly available data. This hold-out set of data is used for the official challenge test phase only and is not available to researchers. In my analysis, I do not make any distinction between the different ECG source locations and combine all of the records into a single repository. Two ECG records (`Training_2/Q0400`, `Training_2/Q2961` from the CPSC2018 tranche of signals) containing no activity on any leads are excluded from analysis.

Each ECG record also contains additional metadata in the form of the patient age and biological sex. There are 187 records that contain an undefined age. Of the remaining records, the mean patient age is 60.3. The ECG records contained primarily male and female sex, with 46.9% of the patients in the dataset as female, 53.1% of the patients as male, and one record that did not have the sex metadata specified. A histogram showcasing the distribution of the age and sex in our dataset can be found in Figure 2.7.

2.2.2 Diagnosed Labels

Each record in our available dataset is labelled with at least one numerical Systematized Nomenclature of Medicine (SNOMED) clinical term (CT). For the purpose of the challenge, a subset of 27 codes/labels have been selected for classification. All other codes are ignored in this study. Please refer to Table 2.1 for the labels, abbreviations, counts, and proportion within the dataset.

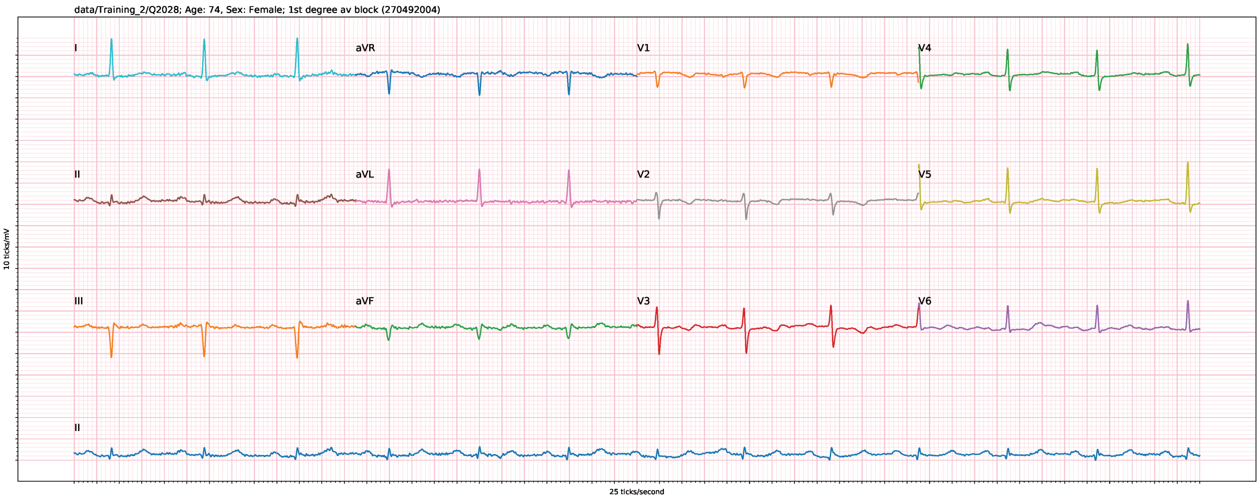

data/Training_2/Q2028; Age: 74, Sex: Female; 1st degree av block (270492004)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.8: Instance of ECG record with 1st degree av block. Long PR interval greater than 200ms is observed.

Table 2.1: Evaluated SNOMED CT codes with definition, count and percentage in dataset.

SNOMED	Abbr.	Diagnosis	Count (%)
270492004	IAVB	1st degree av block	2394 (5.6%)
164889003	AF	atrial fibrillation	3475 (8.0%)
164890007	AFL	atrial flutter	314 (0.7%)
426627000	Brady	bradycardia	288 (0.7%)
713427006	CRBBB	complete right bundle branch block	683 (1.6%)
713426002	IRBBB	incomplete right bundle branch block	1611 (3.7%)
445118002	LAnFB	left anterior fascicular block	1806 (4.2%)
39732003	LAD	left axis deviation	6086 (14.1%)
164909002	LBBB	left bundle branch block	1041 (2.4%)
251146004	LQRSV	low QRS voltages	556 (1.3%)
698252002	NSIVCB	nonspecific intraventricular conduction	997 (2.3%)
10370003	PR	pacing rhythm	299 (0.7%)
284470004	PAC	premature atrial contraction	1729 (4.0%)
427172004	PVC	premature ventricular contractions	188 (0.4%)
164947007	LPR	prolonged PR interval	340 (0.7%)
111975006	LQT	prolonged QT interval	1513 (3.5%)
164917005	QAb	Q wave abnormal	1013 (2.4%)
47665007	RAD	right axis deviation	427 (1.0%)
59118001	RBBB	right bundle branch block	2402 (5.6%)
427393009	SA	sinus arrhythmia	1240 (2.9%)
426177001	SB	sinus bradycardia	2359 (5.5%)
426783006	SNR	sinus rhythm	20846 (48.4%)
427084000	STach	sinus tachycardia	2402 (5.6%)
63593006	SVPB	supraventricular premature beats	215 (0.5%)
164934002	TAb	T wave abnormal	4673 (10.8%)
59931005	TInv	T wave inversion	1112 (2.6%)
17338001	VPB	ventricular premature beats	365 (0.8%)

1st degree av block (IAVB) This is when there is an abnormally long delay between the electrical impulse from the atria, through the ventricular node, to the ventricles. On the ECG, this is detected by the presence of a PR interval longer than 200ms [17] or by the existence of a notched (bimodal) P-wave in leads I, II, III, and aVF [7]. See Figure 2.8 for an example of an ECG record containing this ailment.

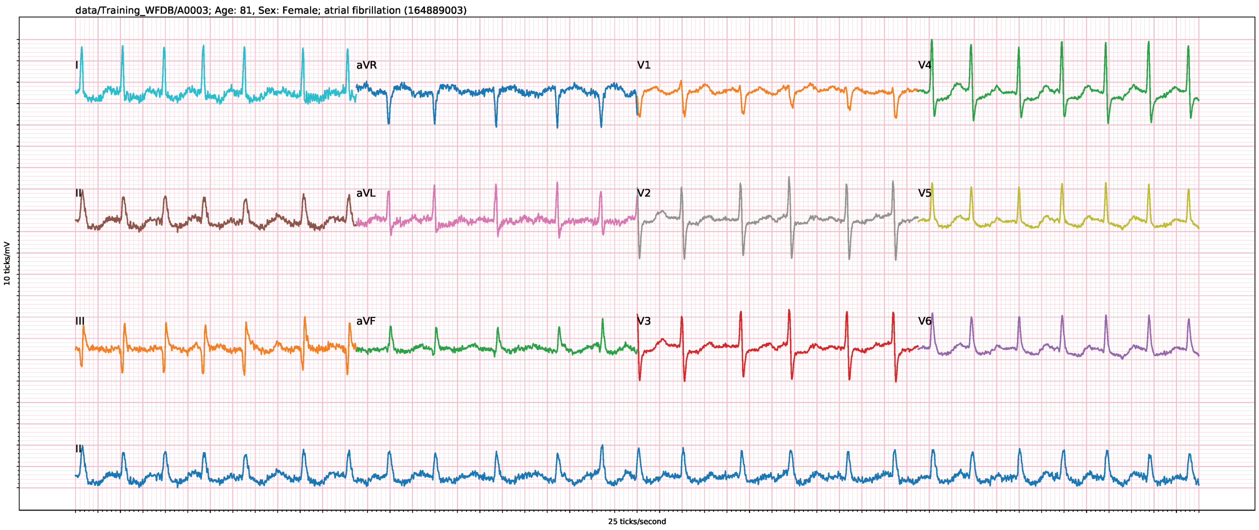

data/Training_WFDB/A0003; Age: 81, Sex: Female; atrial fibrillation (164889003)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.9: Instance of 12-lead ECG with atrial fibrillation. Irregular, fibrillatory p-waves found in all leads.

atrial fibrillation (AF) The manifestation of AF may take on multiple forms, ranging from the absence of P waves, irregular heartbeat rhythm, or fibrillatory (rapid, fluttering) P-waves [41, 37]. Figure 2.9 shows an ECG instance with this disorder.

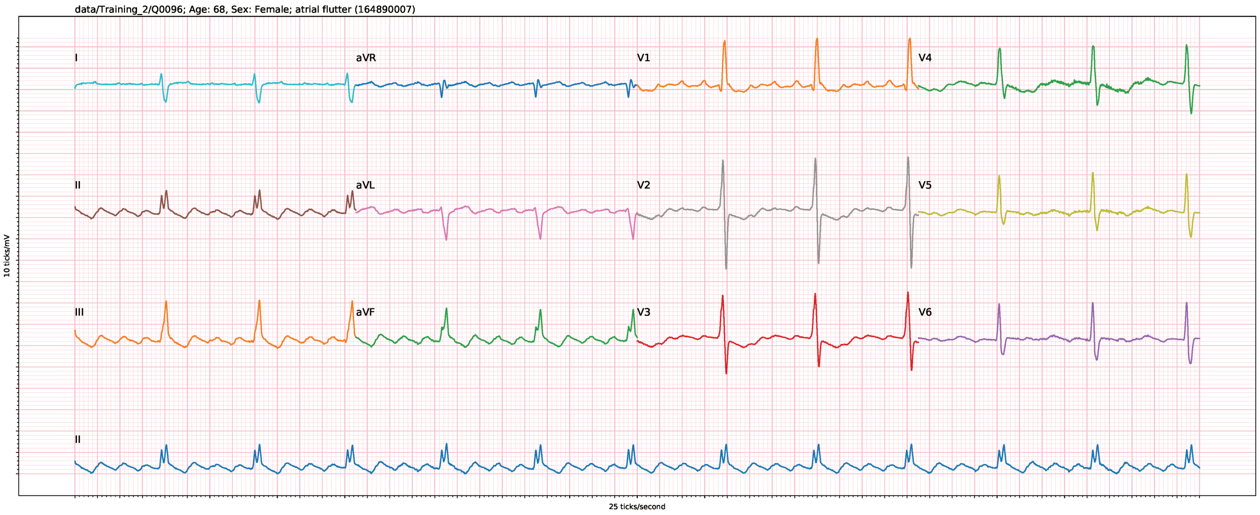

data/Training_2/Q0096; Age: 68, Sex: Female; atrial flutter (164890007)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.10: Instance of 12-lead ECG with atrial flutter. Rapid P waves exist.

atrial flutter (AFL) This disorder is indicated by the presence of atrial rhythms at a constant rate ≥ 100 beats per minute [44]. Figure 2.10 shows an ECG instance with rapid and well-formed P waves.

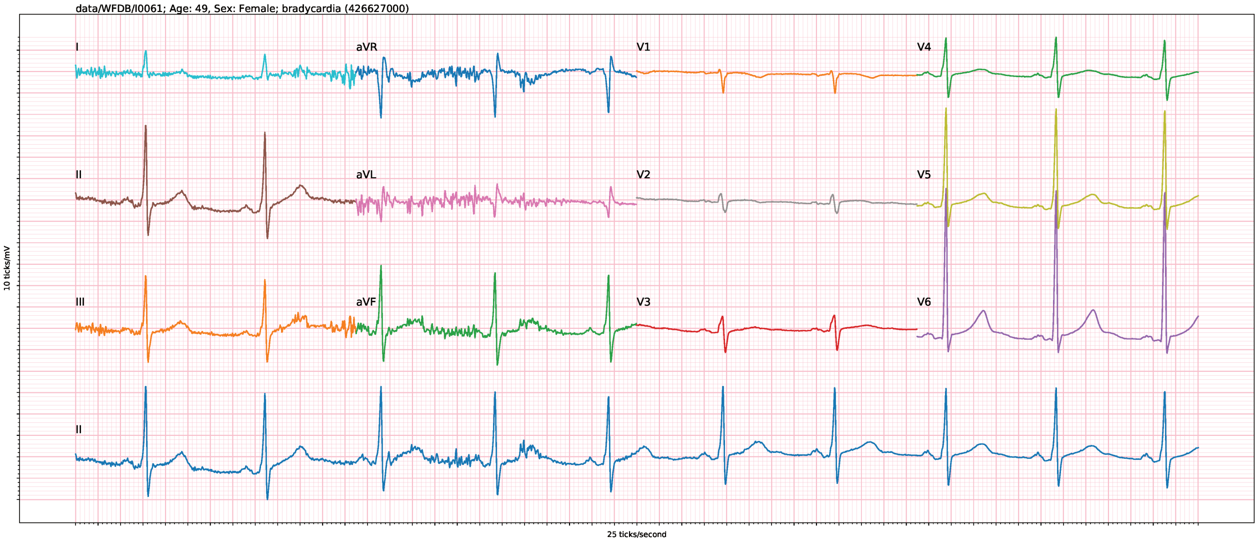

data/WFDB/I0061; Age: 49, Sex: Female; bradycardia (426627000)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.11: Instance of 12-lead ECG with bradycardia. Patient has a mean resting heart rate of 57 beats per minute.

bradycardia (Brady) A patient has bradycardia if their sinus rhythm is below the normal range of for the age of the patient. For adults, this is a resting heart rate below 60 beats per minute. An example of a patient with bradycardia is shown in Figure 2.11.

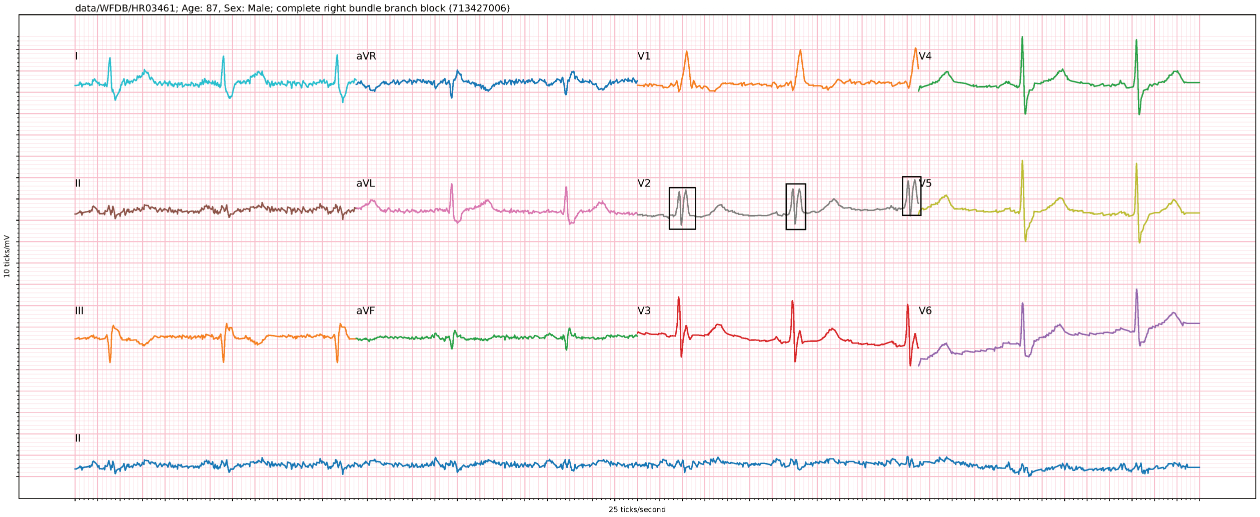

data/WFDB/HR03461; Age: 87, Sex: Male; complete right bundle branch block (713427006)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.12: Instance of 12-lead ECG with complete right bundle branch block. QRS complex exceeds 120 ms (4 boxes) and contains an anterior skew on the right end of the QRS complex. Characteristic "bunny ears" trait emphasized in lead V2.

complete right bundle branch block (CRBBB) A complete right bundle branch block is when the QRS duration exceeds 120 ms and the terminal halves of the QRS are skewed rightward and anteriorly, indicating that the right ventricle is being depolarized after the left ventricle [57]. An example ECG record is shown in Figure 2.12.

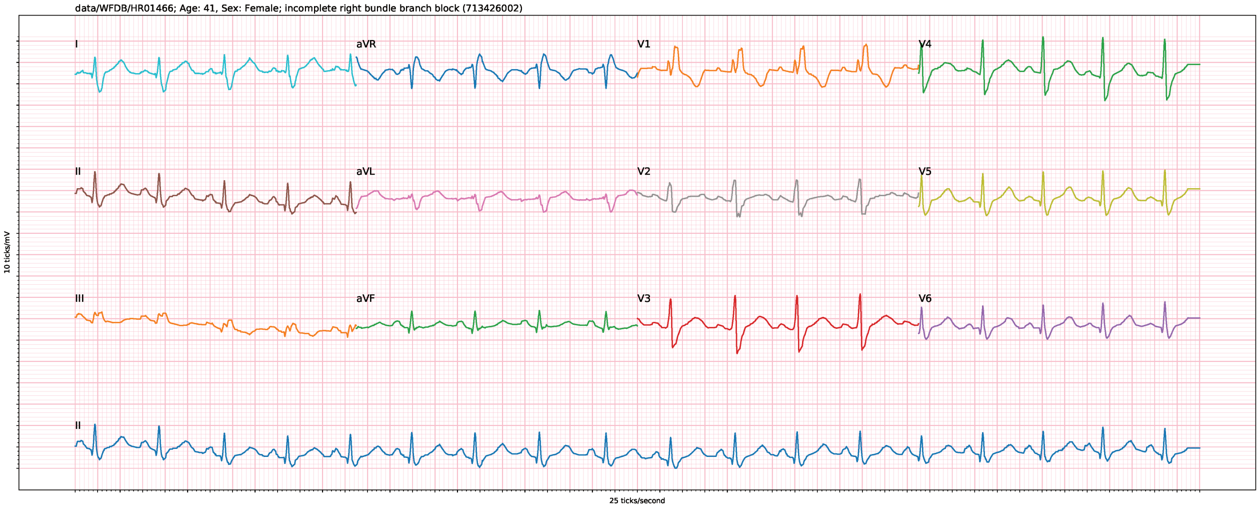

data/WFDB/HR01466; Age: 41, Sex: Female; incomplete right bundle branch block (713426002)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.13: Instance of 12-lead ECG with incomplete right bundle branch block. The QRS complex is within 100 to 120 ms (2.5-3 boxes) and contains an anterior skew on the right end of the QRS complex.

incomplete right bundle branch block (IRBBB) An incomplete right bundle branch block has the same terminal QRS features as CRBBB but has a QRS duration of 100 to 120 ms [57], as shown in the example record at Figure 2.13.

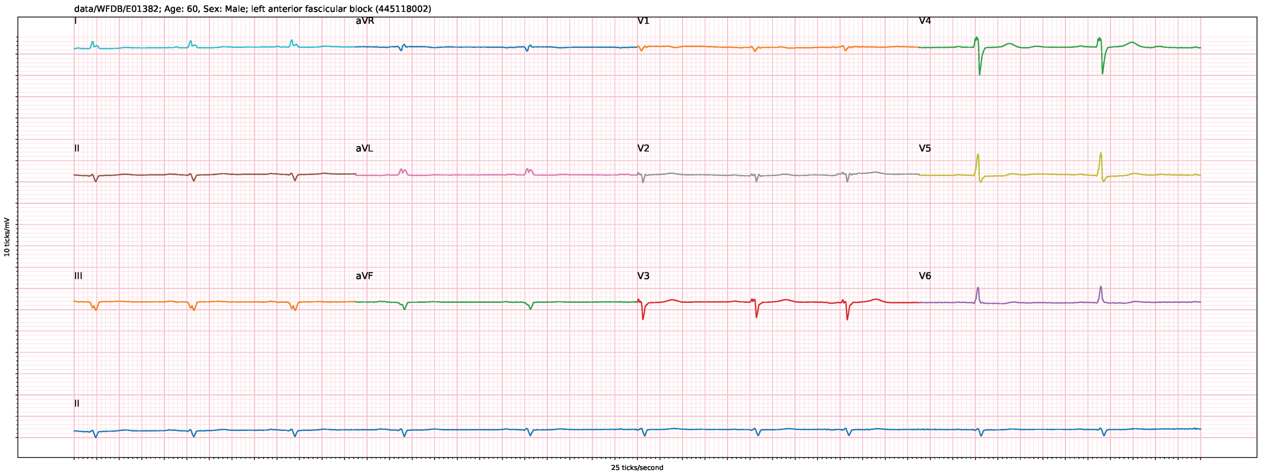

data/WFDB/E01382; Age: 60, Sex: Male; left anterior fascicular block (445118002)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.14: Instance of 12-lead ECG with left anterior fascicular block. Lead I is positive while leads II and III are negative, indicating left axis deviation. Leads I and aVL show qR complexes, with rS complexes appearing in leads II, III, an aVF.

left anterior fascicular block (LAnFB) This common intraventricular defect occurs when the anterior fascicle within the left bundle branch is blocked or otherwise unable to respond to action potential stimuli [57]. The criteria includes: left axis deviation (within −45° to −90°); small Q waves with large R waves (‘qR complexes’) in leads I and/or aVL; small R waves with deep S waves (‘rS complexes’) in leads II, III, and aVF; slightly prolonged QRS duration (80-110ms), usually a poor R progression in leads V1, V2, and V3 with deeper S waves in leads V5 and V6. See Figure 2.14 for an example.

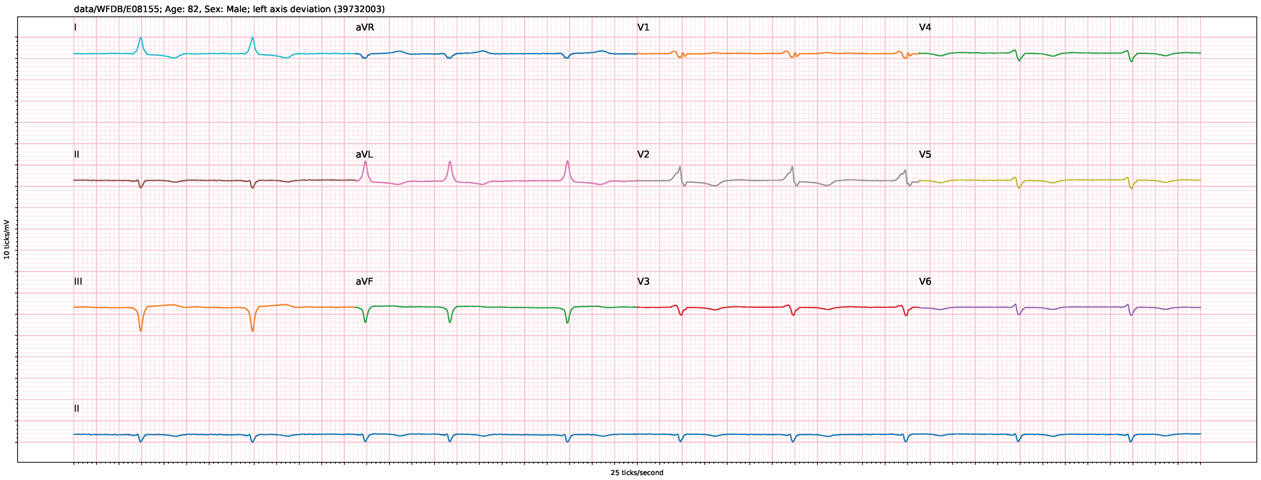

data/WFDB/E08155; Age: 82, Sex: Male; left axis deviation (39732003)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.15: Instance of 12-lead ECG with left axis deviation. Lead I is positive while leads II and III are negative, indicating left axis deviation using the simple leads I, II, III method. Using the hexaxial axis estimation approach, we see that leads aVL and III have the greatest QRS complex magnitude with aVL being positive and III being negative, giving an electrical axis between $-30°$ and $-60°$, confirming left axis deviation.

left axis deviation (LAD) Diagnosed when the cardiac axis exists between $-30°$ and $-90°$. See Figure 2.15 for an example ECG record, interpreted using the lead I, II, III approach as well as the hexaxial reference system.

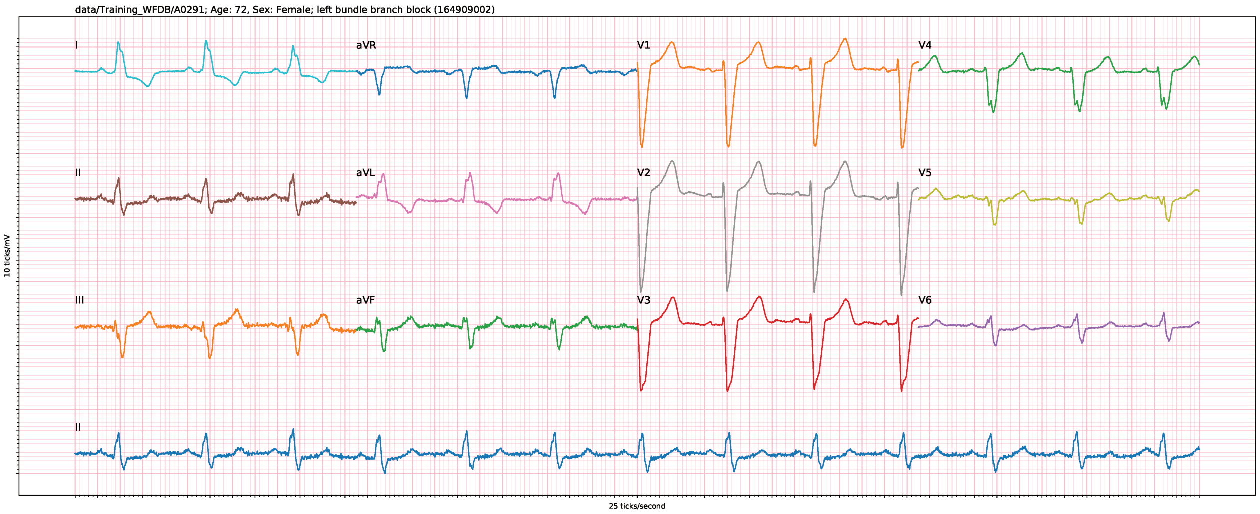

data/Training_WFDB/A0291; Age: 72, Sex: Female; left bundle branch block (164909002)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.16: Instance of 12-lead ECG with left bundle branch block. The QRS complexes exceed 100ms (2.5 boxes) and contain a posterior or leftward skew on the tail halves of the QRS complexes.

left bundle branch block (LBBB) A left bundle branch block is defined by a QRS complex exceeding 100ms with a posterior/leftward skew in the second half of the QRS complex [57]. An example record is shown in Figure 2.16. For the purposes of this challenge, no distinction is made between complete left bundle branch block (> 120ms) and incomplete left bundle branch block (between 100ms and 120ms).

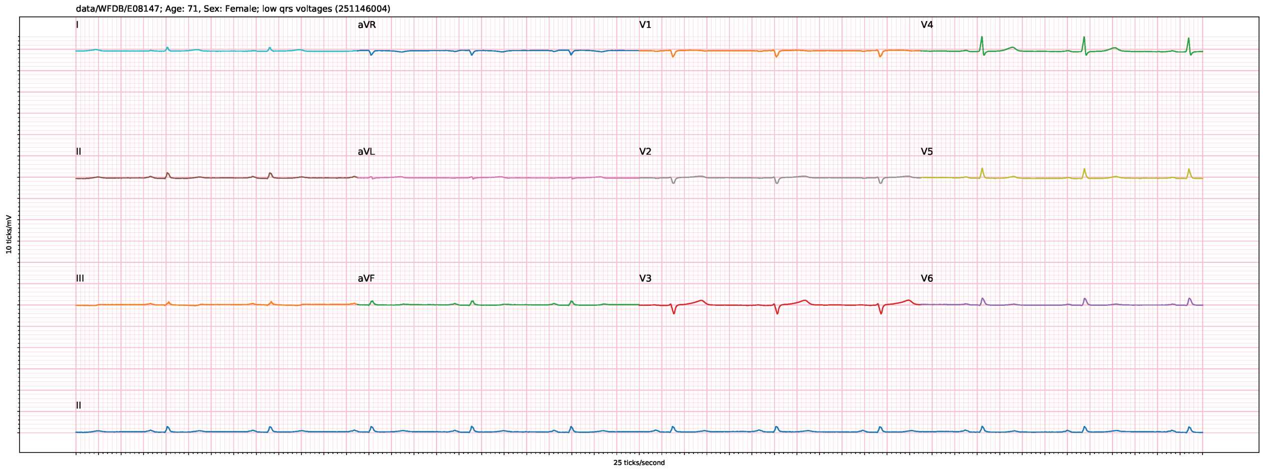

data/WFDB/E08147; Age: 71, Sex: Female; low qrs voltages (251146004)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.17: Instance of 12-lead ECG with low QRS voltages. All limb leads have QRS complex amplitudes of less than 0.5mV and all precordial leads have QRS complex amplitudes of less than 1.0mV.

low QRS voltages (LQRSV) An ECG record with low QRS voltages contains QRS complex amplitudes of < 0.5mV in all limb derived leads (I, II, III, aVR, aVL, aVF) or contains QRS complex amplitudes of < 1mV in all precordial leads (V1, V2, V3, V4, V5, V6) [32]. See Figure 2.17 for an example.

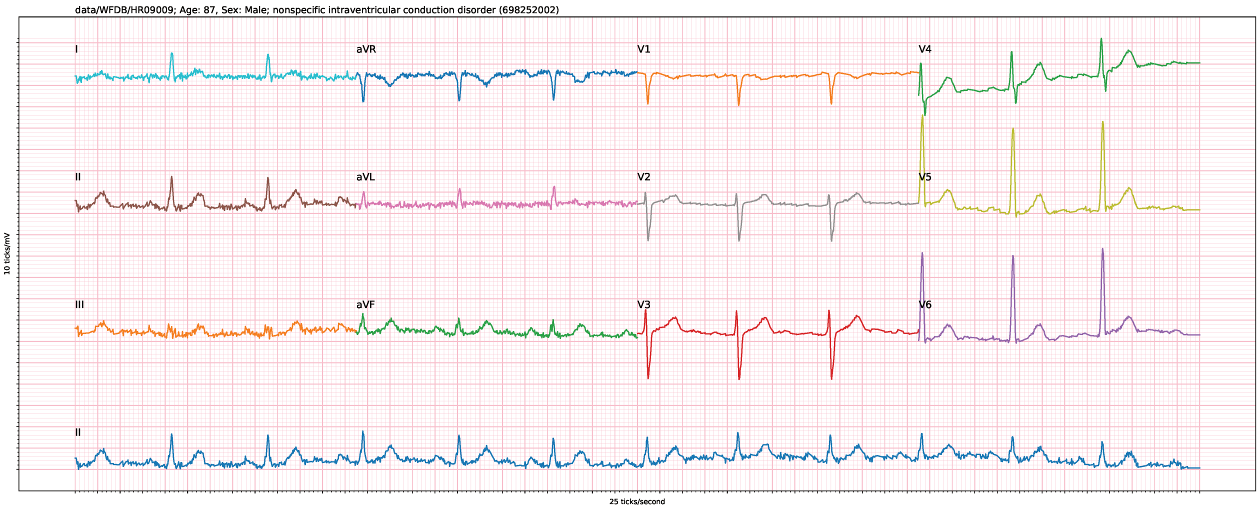

data/WFDB/HR09009; Age: 87, Sex: Male; nonspecific intraventricular conduction disorder (698252002)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.18: Instance of 12-lead ECG with nonspecific intraventricular conduction disorder. QRS complexes exceed 100ms (2.5 squares) but do not exhibit a posterior/anterior skew characteristic on the tail half of the QRS complexes.

nonspecific intraventricular conduction (NSIVCB) An ECG record with nonspecific intraventricular conduction blocks exhibits QRS durations over 100ms but do not contain a posterior/anterior skew characteristic on the second half of the QRS complex [57]. An example can be found in Figure 2.18.

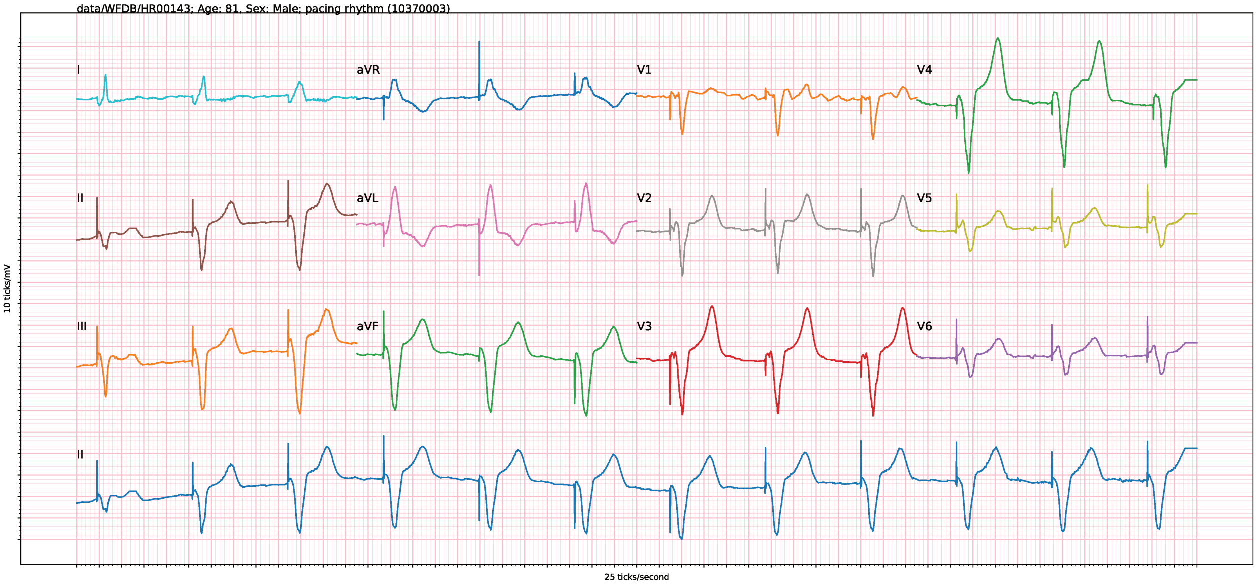

data/WFDB/HR00143; Age: 81, Sex: Male; pacing rhythm (10370003)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.19: Instance of 12-lead ECG with pacing rhythm. Characteristic vertical lines appearing at the P wave onset suggests atrial pacing with normal conduction to the ventricles.

pacing rhythm (PR) This diagnosis indicates the presence of a pacemaker. One common artifact is a vertical line appearing in the ECG prior to the QRS complex [27]. An example ECG record can be found in Figure 2.19.

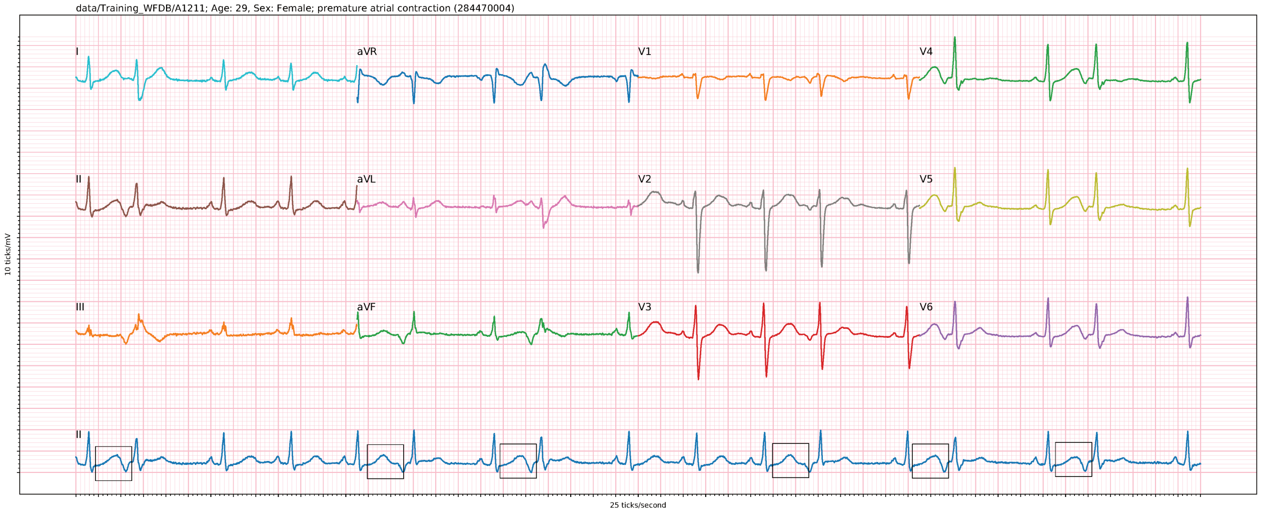

data/Training_WFDB/A1211; Age: 29, Sex: Female; premature atrial contraction (284470004)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.20: Instance of 12-lead ECG with premature atrial contraction. Instances of P waves occurring within preceding beat T wave in lead II are emphasized with black squares.

premature atrial contraction (PAC) These events may occur as single or repetitive events along the ECG record and are characterized by a P wave occurring within the T wave of the preceding beat [57]. An example can be found in Figure 2.20.

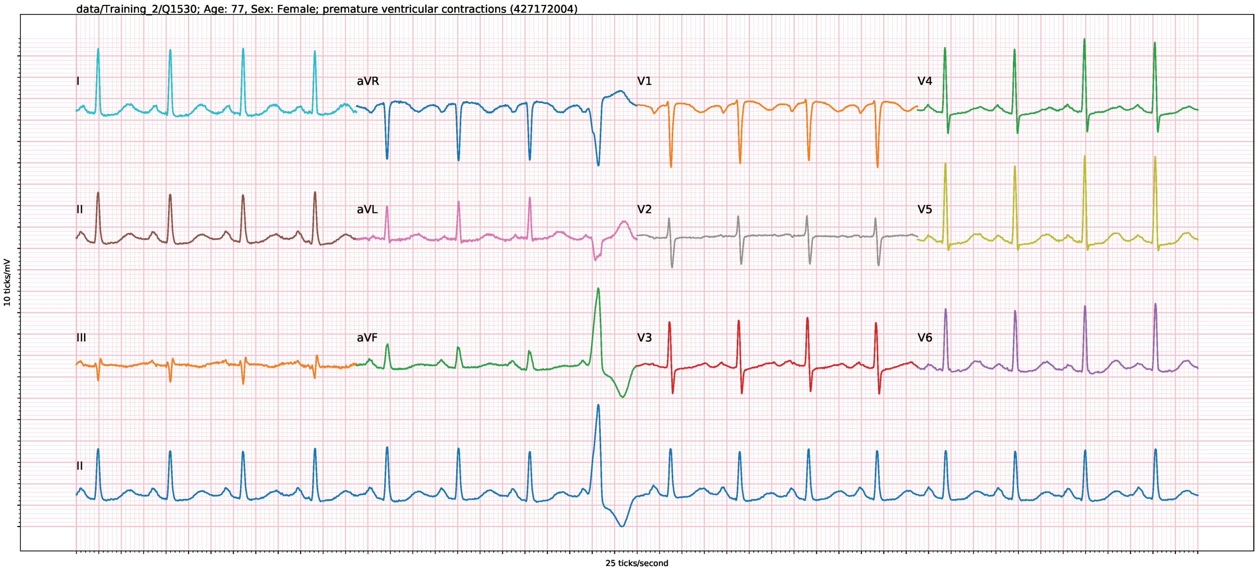

data/Training_2/Q1530; Age: 77, Sex: Female; premature ventricular contractions (427172004)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.21: Instance of 12-lead ECG with premature ventricular contraction. Premature ventricular contraction occurs at the 4.5 second mark.

premature ventricular contractions (PVC) This diagnosis indicates that a ventricular contraction has occurred before the next expected sinoatrial node action potential [47]. It may occur as singular or repetitive events along the ECG record. Refer to Figure 2.21 for an example ECG record containing this disorder.

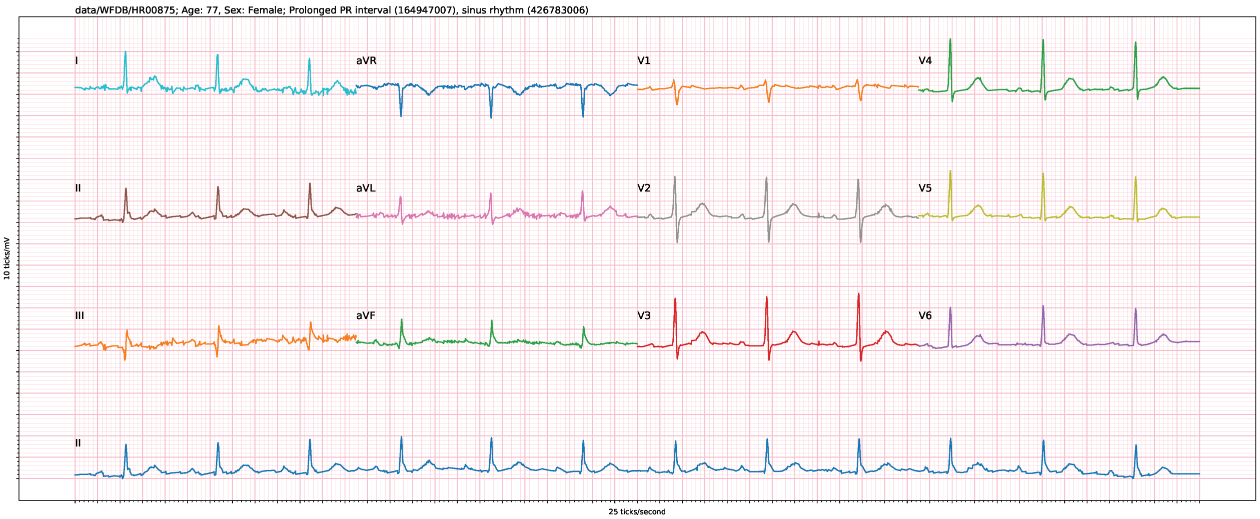

data/WFDB/HR00875; Age: 77, Sex: Female; Prolonged PR interval (164947007), sinus rhythm (426783006)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.22: Instance of 12-lead ECG with prolonged PR interval. Patient record contains PR intervals exceeding 200ms (5 squares) but is also diagnosed as having normal sinus rhythm.

prolonged PR interval (LPR) This disorder indicates a delayed conduction through the atrioventricular node and is characterized by a PR interval exceeding 200ms [29]. This condition has overlap with IAVB but suggests reduced severity. An example record can be found in Figure 2.22.

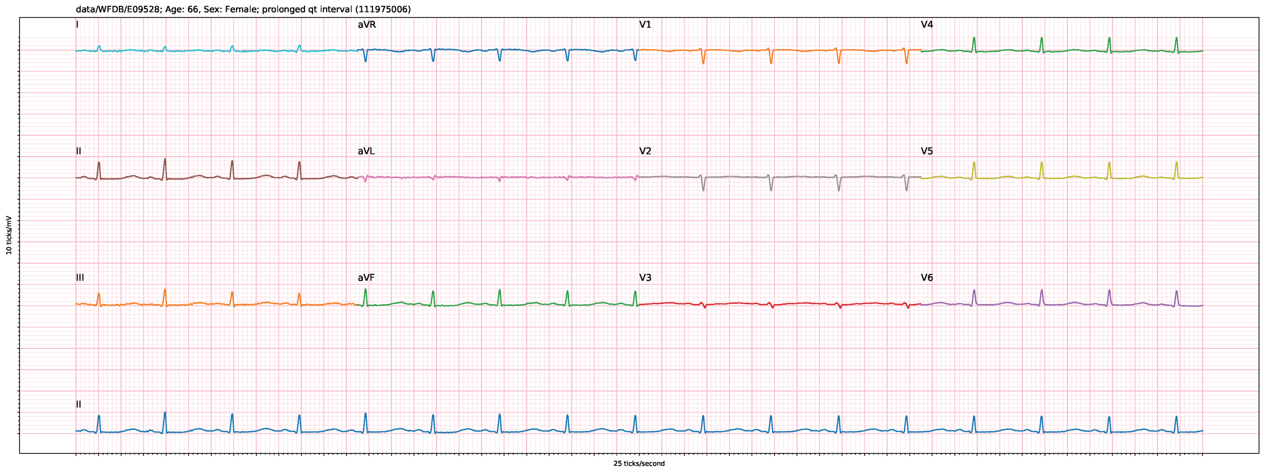

data/WFDB/E09528; Age: 66, Sex: Female; prolonged qt interval (111975006)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.23: Instance of 12-lead ECG with prolonged QT interval. Patient T waves end past the halfway point of the preceding and upcoming R peaks.

prolonged QT interval (LQT) This disorder indicates that the heart is inadequately recovering after each beat and can be quickly identified by the presence of T waves ending beyond the midway point of an RR interval [36]. See Figure 2.23 for an example record with this disorder.

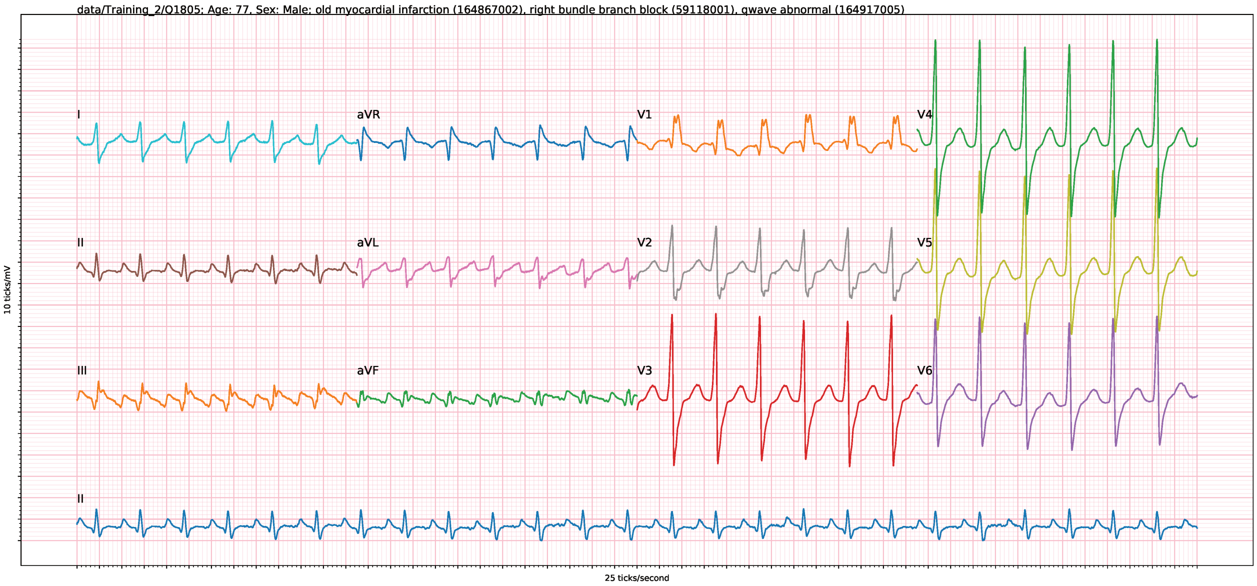

data/Training_2/Q1805; Age: 77, Sex: Male; old myocardial infarction (164867002), right bundle branch block (59118001), qwave abnormal (164917005)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.24: Instance of 12-lead ECG with abnormal Q wave. Patient Q waves exceed 40ms wave duration in leads I, III, aVR, aVL, V2, V3, V4, V5, V6 and have Q wave amplitudes exceeding 25% of the QRS complex amplitude in precordial leads V2-V6.

Q wave abnormal (QAb) Pathologic Q waves indicate the presence of an old myocardial infarction and are diagnosed as having Q wave duration exceeding 40 ms or Q wave amplitude exceeding 25% of the amplitude of the QRS complex [10]. Figure 2.24 showcases an example record containing abnormal Q wave.

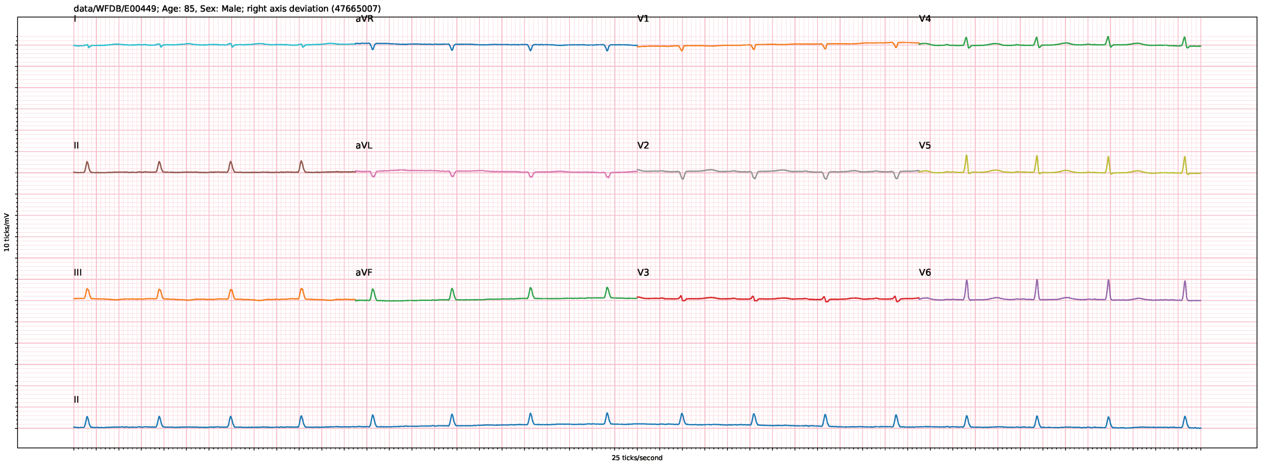

data/WFDB/E00449; Age: 85, Sex: Male; right axis deviation (47665007)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.25: Instance of 12-lead ECG with right axis deviation. A negative lead I with positive leads II and III suggest right axis deviation. Using the hexaxial system, leads II, III, and aVF all have the largest QRS amplitudes (each approximately 0.3mV), resulting in an cardiac axis estimate between 60° and 120°. This overlaps with our right axis deviation criteria where the axis exists between 90° and 180°.

right axis deviation (RAD) This diagnosis indicates the cardiac axis existing between 90° and 180°. See Figure 2.25 for an example ECG record interpreted using the lead I, II, III approach as well as the hexaxial reference system.

right bundle branch block (RBBB) This diagnosis is identical to CRBBB.

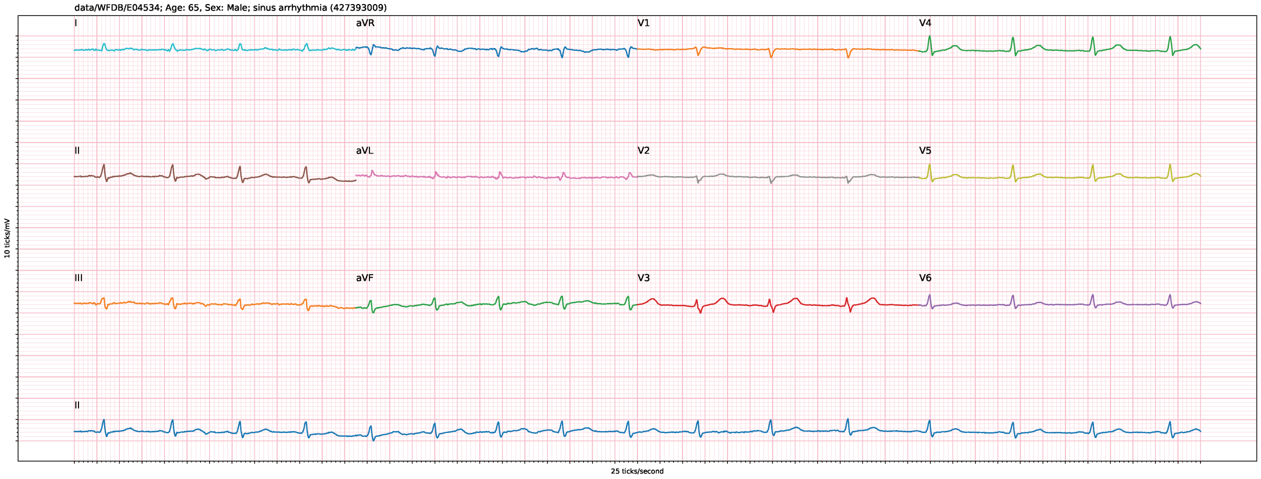

data/WFDB/E04534; Age: 65, Sex: Male; sinus arrhythmia (427393009)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.26: Instance of 12-lead ECG with sinus arrhythmia. Recorded heart rate decreases over time, or distance between R peaks increases over time.

sinus arrhythmia (SA) This diagnosis reflects a change in the beat-to-beat variation over time, resulting in an irregular heart rate [47]. This condition is typically not serious. An example of an ECG record with SA can be found in Figure 2.26.

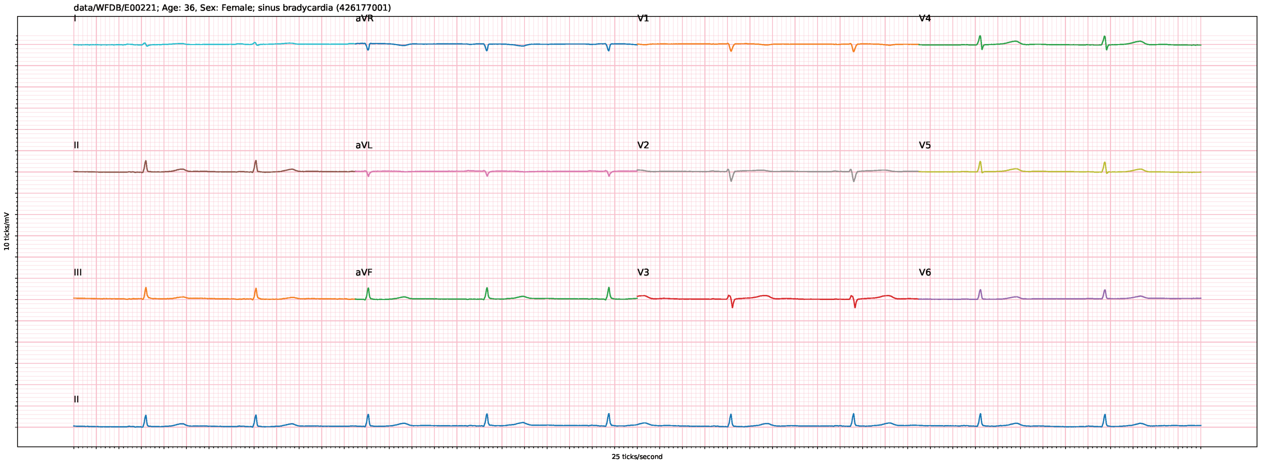

data/WFDB/E00221; Age: 36, Sex: Female; sinus bradycardia (426177001)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.27: Instance of 12-lead ECG with sinus bradycardia. Upright P wave in lead II exists, with a mean heart rate of 59 beats per minute.

sinus bradycardia (SB) This specific type of slow heartbeat is caused by the sinoatrial node firing less than 60 times per minute. It is characterized by an upright P wave in lead II preceding every QRS complex and a heart rate of less than 60 beats per minute in adults. An example record is shown in Figure 2.27.

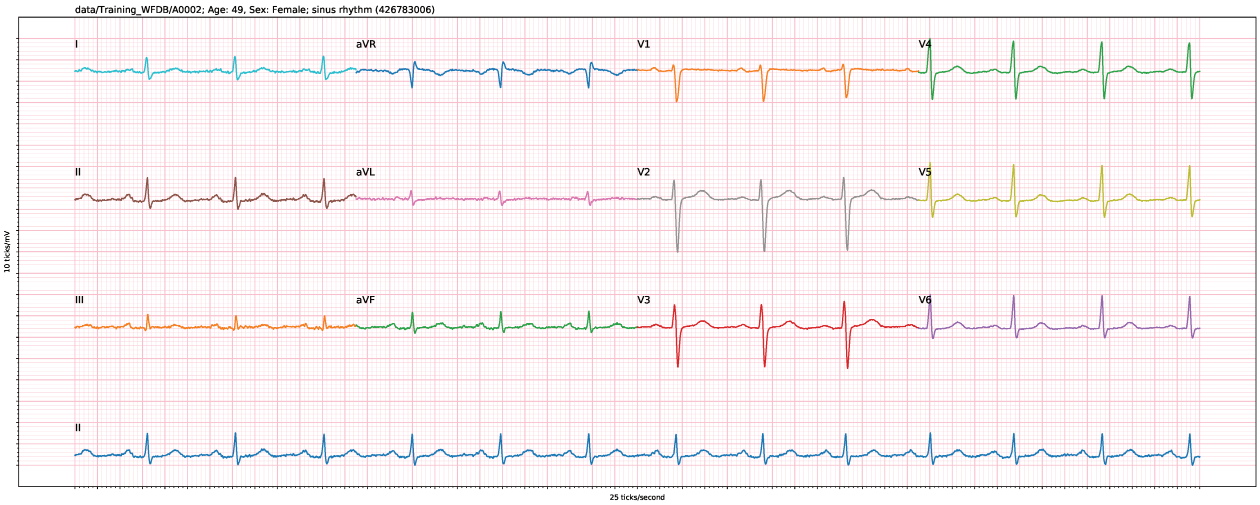

data/Training_WFDB/A0002; Age: 49, Sex: Female; sinus rhythm (426783006)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.28: Instance of 12-lead ECG with normal sinus rhythm. No visible defects or abnormalities indicated.

sinus rhythm (SNR) The normal, healthy case for an ECG record must have an upright P wave in leads I and II, with a QRS complex following each P wave [35]. In an adult, the resting heart rate should be between 60-99 beats per minute. A reference ECG can be found in Figure 2.28

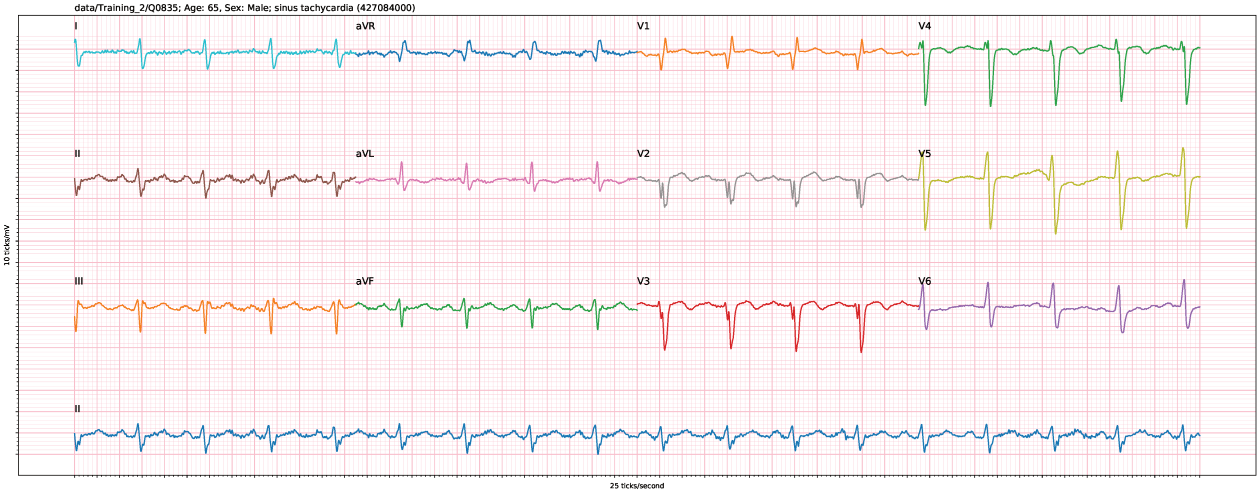

data/Training_2/Q0835; Age: 65, Sex: Male; sinus tachycardia (427084000)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.29: Instance of 12-lead ECG with sinus tachycardia. The average heart rate is 101 beats per minute.

sinus tachycardia (STach) This is when the resting heart rate exceeds 100 beats per minute in adults, or above the normal range relative to the patient age. An example ECG record is shown in Figure 2.29

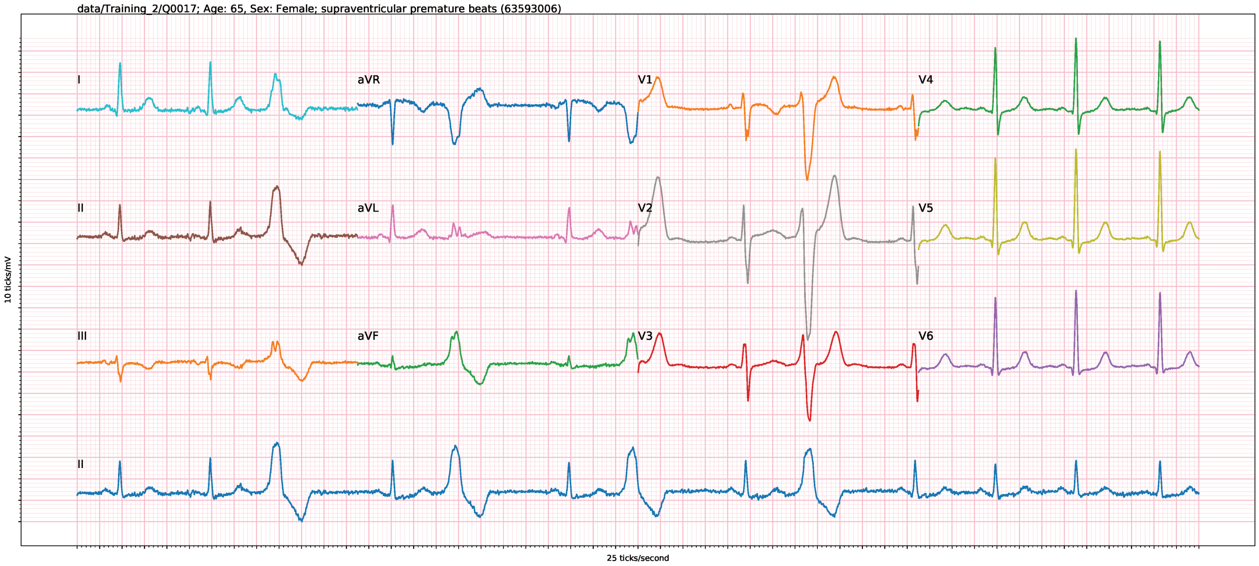

data/Training_2/Q0017; Age: 65, Sex: Female; supraventricular premature beats (63593006)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.30: Instance of 12-lead ECG with supraventricular premature beats. Four instances of premature ventricular beats are observed in this record.

supraventricular premature beats (SVPB) This occurs when the atrial contractions are triggered through invalid conduction such as an origin other than the sinoatrial node [22]. See Figure 2.30 for a reference ECG with this disorder. For the PhysioNet challenge, this diagnosis is treated identically to PAC, despite different ECG characteristics.

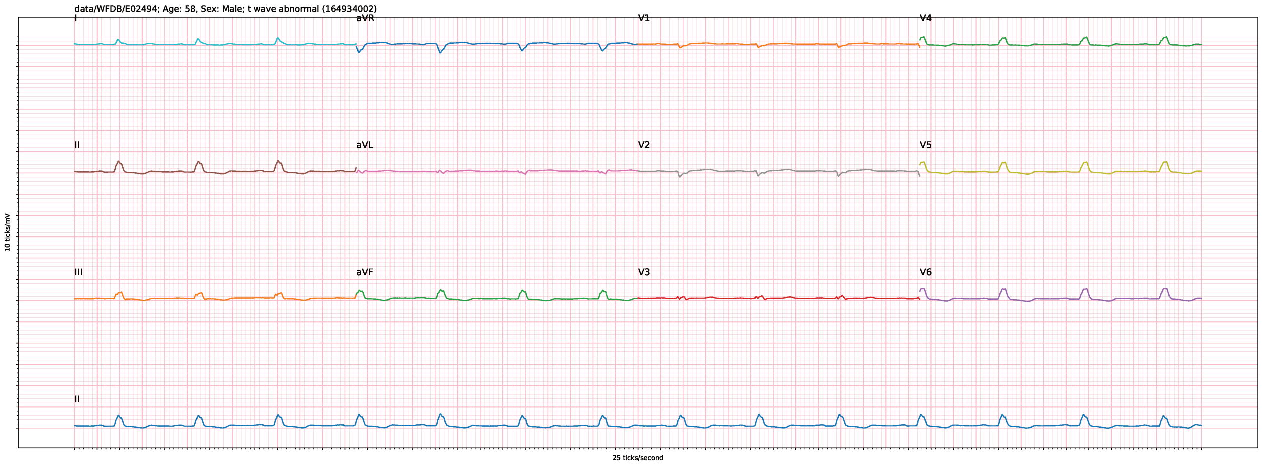

data/WFDB/E02494; Age: 58, Sex: Male; t wave abnormal (164934002)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.31: Instance of 12-lead ECG with abnormal T wave.

T wave abnormal (TAb) An abnormality in the T wave ranges from low T-wave amplitudes to complete inversion of the T wave. Normal T waves typically align with the QRS complex direction. They should be upright in leads I, II, V3-6, and inverted in lead aVR. Normal T waves are typically asymmetric, with a faster second half compared to the first half [57]. See Figure 2.31 for a reference ECG with abnormal T waves.

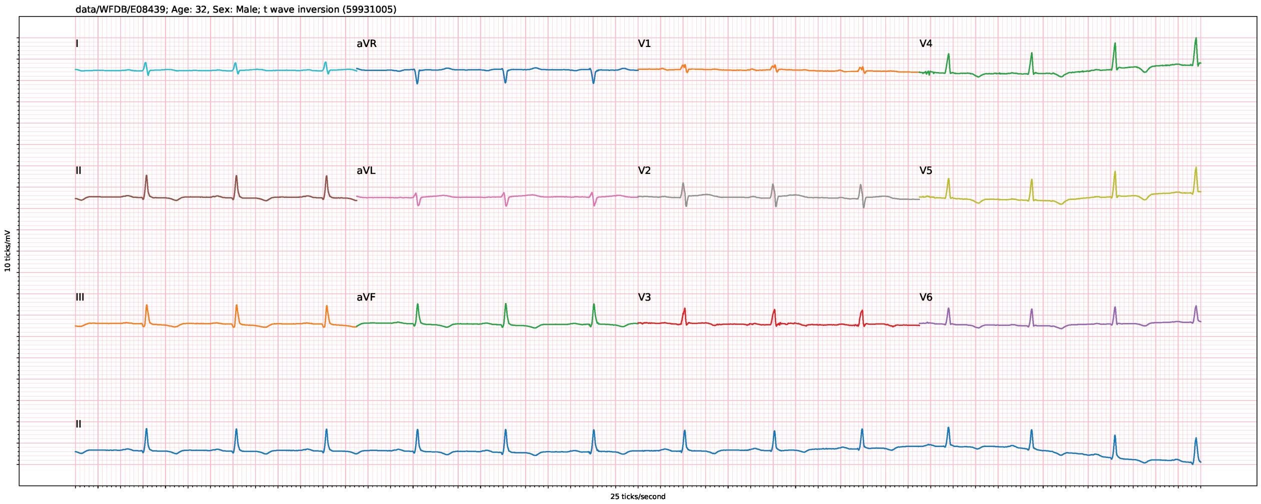

data/WFDB/E08439; Age: 32, Sex: Male; t wave inversion (59931005)
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6
II
10 ticks/mV
25 ticks/second

Figure 2.32: Instance of 12-lead ECG with T wave inversion. Leads with T waves of incorrect direction are observed in leads II, III, aVR (should be inverted), aVF, and V3-6.

T wave inversion (TInv) This condition is semantically a more specific form of TAb, but only evaluating the condition of T wave inversion. Figure 2.32 contains a reference ECG with inverted T waves.

ventricular premature beats (VPB) This diagnosis is identical to PVC.

2.2.3 Scoring Function

The *PhysioNet/CinC 2020 Challenge* [40] rewards partial credit for incorrect classification predictions using a weighted scoring function. The intention of this scoring function is to weigh predictions that result in similar outcomes or treatments as the ground truth labels more favorably. The similarity of two classes aim to minimize potential harm of applied treatment, as treatment response may be similar in both cases.

We define a set of diagnoses $C = \{c_i\}_{i=1}^{m}, \quad i = 1 \dots n$ for m distinct labels in a corpus of n recordings. We compute a confusion matrix $A = [a_{ij}]$ such that each cell contains the sum of all n records that are classified as diagnosis c_i and have the ground truth diagnosis of c_j, see Equation 2.8 and Equation 2.9.

$$a_{ij} = \sum_{k=1}^{n} a_{ijk} \tag{2.8}$$

Due to the multi-label nature of this task, a single ECG record may have multiple diagnoses. The contribution of a single record is normalized by dividing over $|\{x_k \cup y_k\}|$, or the number of classes with a ground truth positive label or classifier output.

$$a_{ijk} = \begin{cases} \dfrac{1}{|\{x_k \cup y_k\}|} & \text{if } c_i \in x_k \text{ and } c_j \in y_k \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

A matrix of weights $W = [w_{ij}]$ is provided by the challenge organizers to specify the reward for a classifier output of class c_i with a ground truth positive label of c_j. See Figure 2.33 for a visualization of the provided weights and relationship to the classification labels.

$$s = \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} a_{ij} \tag{2.10}$$

The objective of the challenge is to maximize the scoring metric s as defined in Equation 2.10. The challenge emphasizes that the reward matrix used W is opinionated and should be tailored to the preferences of the researchers and institutions in future experiments [40].

2.3 Related Work

I summarize the *PhysioNet/CinC 2020 Challenge* [40] successful entrants, emphasizing noteworthy approaches and categorizing the general distribution of the competition strategies. All of the ranked challenge participants have publicly available papers [1]. There are also publications that do not have a official phase challenge entry, due to various failures in the remote evaluation and

---

[1] PhysioNet 2020 successful entrants and papers: `https://github.com/physionetchallenges/physionetchallenges.github.io/tree/master/2020/papers`

training of the submitted entry. All in all, there are 41 successful submissions and 62 publications addressing the 12-lead ECG classification challenge.

The top ranked entry proposed by Natarajan *et al* combined handcrafted ECG features with a deep learning convolutional and transformer neural network model for learned discriminative feature representations [38]. They initially generated over 300 ECG features extracted from lead II, but pruned to 20 derived features using a random forest classifier. The remaining leads were not used in tabular feature engineering, but were applied to training the transformer neural network. They achieved a final official phase validation score of 0.587 with a test score of 0.533. Impressively, despite this approach using deep learning techniques and winning the competition, no external datasets were used to pretrain or otherwise enhance the model.

The second ranked entry submitted by Zhao *et al* used an adapted ResNet for the classification of ECGs and attained a official phase validation score of 0.672 and test score of 0.520 [58]. The age and sex meta features were encoded into their final classification layer using one-hot encoding, relying otherwise on learned features from their deep convolutional network. Additionally, their model was trained on additional data not provided by the official challenge sources. The third ranked entry proposed by Zhu *et al* also used an adapted ResNet with squeeze-and-excitation blocks and scored a challenge validation score of 0.682 with a test score of 0.514 [59]. No metadata was used in their classifier, relying only on signal morphology to generate their predictions.

Of all 62 accepted papers, the overwhelming majority of approaches used deep learning as part of their model. Convolutions were used in 54 (87.1%) of all challenge papers. Recurrent neural networks were used in 32 (51.6%) of the papers, with 17 (27.4%) of the papers using the long short term memory (LSTM) network variant. Three (4.8%) papers used the transformer neural network architecture. Only 7 (11.3%) papers opted not to use deep learning at all, relying on manual feature extraction and traditional machine learning approaches like decision trees and support vector machines (our challenge submission being one of these 7 papers).

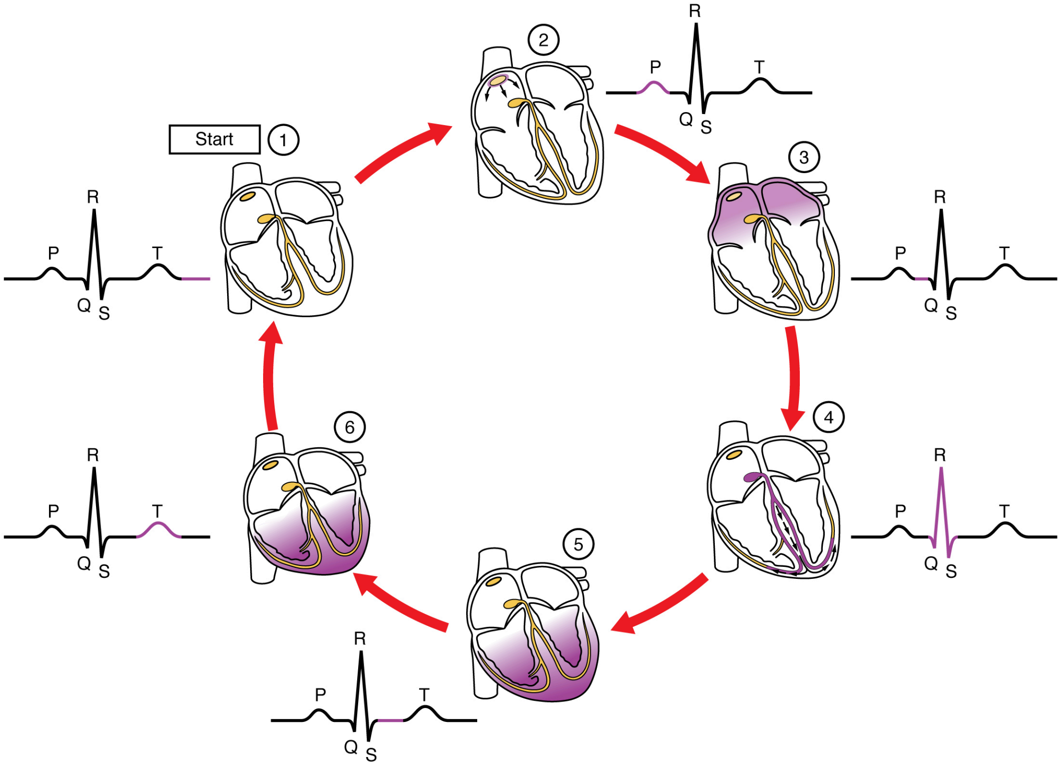


Figure 2.4: Cardiac conduction correlated to ECG tracing waves [9]. 1. The conduction system of the heart is currently at rest, with the ventricles repolarized. 2. The sinoatrial node begins an action potential which permeates across the atria, causing the ECG P wave formation. 3. A 100ms delay in the impulse occurs which allows the atria to complete pumping blood, showing as the PR segment in the ECG. 4. The impulse proceeds through the atrioventricular bundle and bundle branches to the Purkinje fibers, appearing as the QRS complex in the ECG. 5. The contractile fibers of the ventricles are stimulated by the impulse, causing the ventricles to contract and appears as the ST-segment in the ECG. 6. The impulse dissipates and the ventricular muscles relax, causing the ECG T wave formation. Image licensed `CC BY 4.0` from Betts *et al* [9] on the OpenStax platform, source: `https://openstax.org/books/anatomy-and-physiology/pages/19-2-cardiac-muscle-and-electrical-activity#fig-ch20_02_08`.

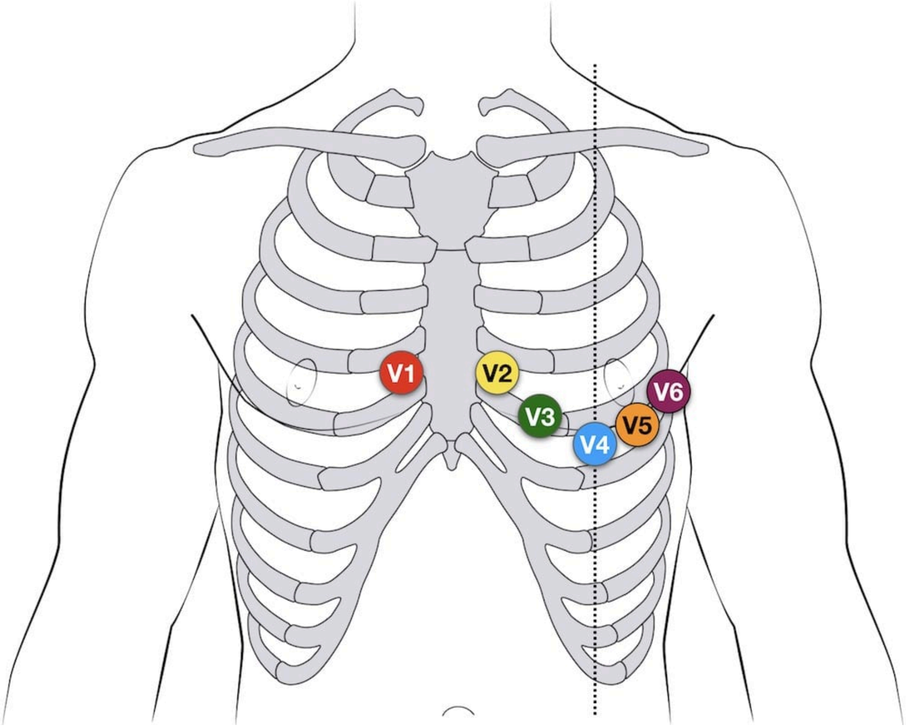


Figure 2.5: Placement of 12-lead ECG for precordial electrodes V1-V6 [16]. V1 is located in the 4th intercostal space (ICS) on the right margin of the sternum. V2 is placed at the 4th ICS on the left margin of the sternum. V3 is placed midway between V2 and V4. V4 is placed on the 5th ICS mid-clavicular line. V5 is placed on the anterior axillary line at the same level as V4 (5th ICS). V6 is placed on the mid-axillary line at the same level as V4 (5th ICS). Image licensed `CC BY-NC-SA 4.0` from Mike Cadogan [16] on the Life In The Fast Lane platform, source: `https://litfl.com/ecg-lead-positioning/`.

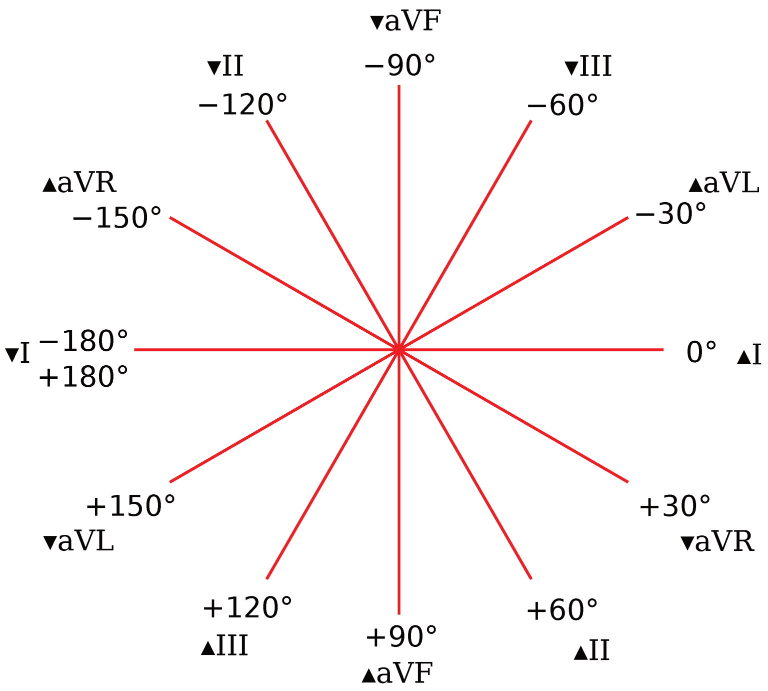


Figure 2.6: Hexaxial (Cabrera) system for determining cardiac axis [49]. Six frontal planes are viewed in the sequence aVL, I, aVR, II, aVF, and III. Within an ECG record, the lead with the maximum QRS peak amplitude change, positive or negative, is used as the estimate for the cardiac electrical axis. Image belonging to `public domain` from Wikipedia users MoodyGroove and Mysid, source: `https://commons.wikimedia.org/w/index.php?curid=2635587`.

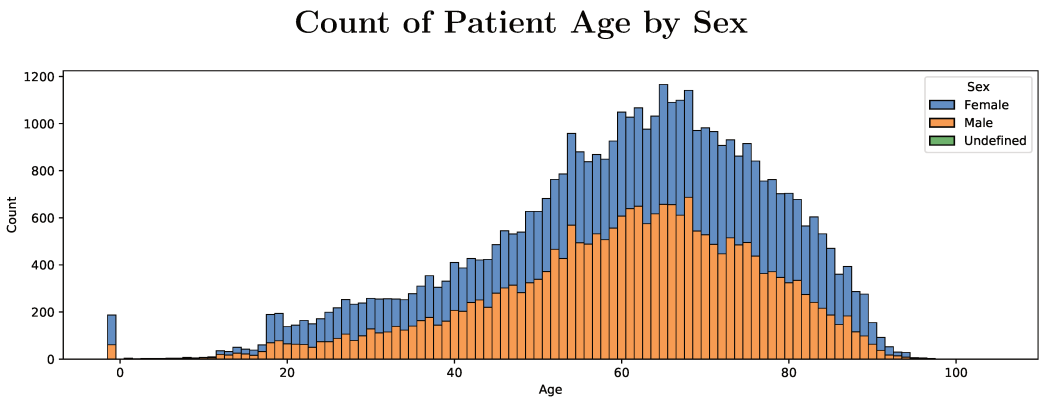
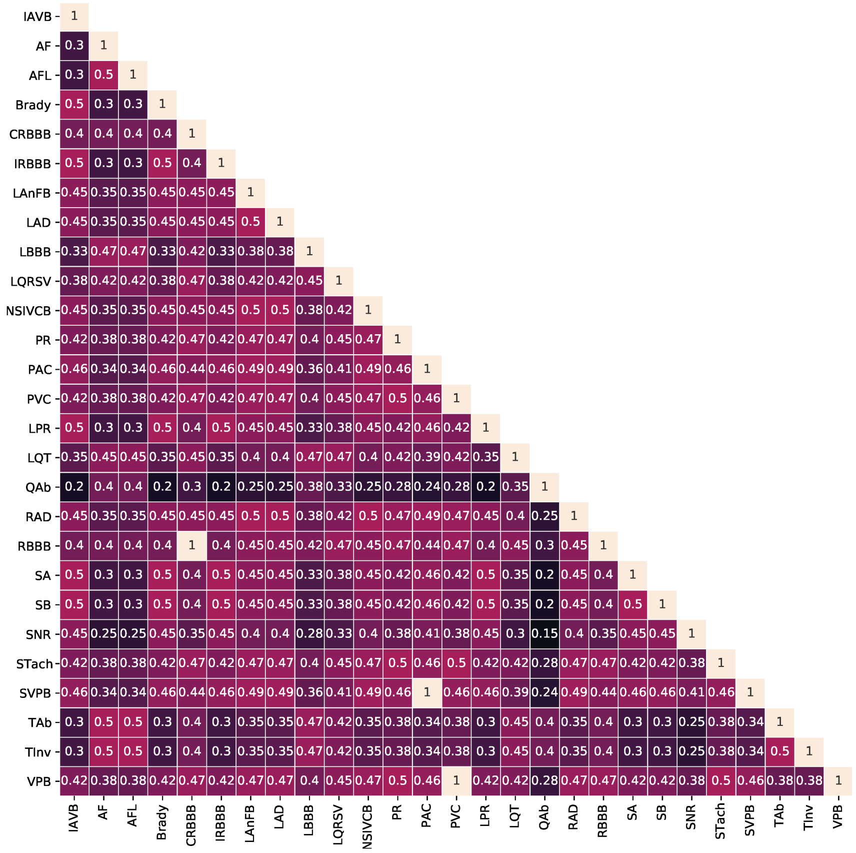


Figure 2.7: Age and sex distribution of the PhysioNet/CinC 2020 public electrocardiogram dataset. Records with age < 0 indicate that age is not provided for the given record. The only record that is missing the biological sex metadata is also missing the age metadata.



	IAVB	AF	AFL	Brady	CRBBB	IRBBB	LAnFB	LAD	LBBB	LQRSV	NSIVCB	PR	PAC	PVC	LPR	LQT	QAb	RAD	RBBB	SA	SB	SNR	STach	SVPB	TAb	TInv	VPB
IAVB	1																										
AF	0.3	1																									
AFL	0.3	0.5	1																								
Brady	0.5	0.3	0.3	1																							
CRBBB	0.4	0.4	0.4	0.4	1																						
IRBBB	0.5	0.3	0.3	0.5	0.4	1																					
LAnFB	0.45	0.35	0.35	0.45	0.45	0.45	1																				
LAD	0.45	0.35	0.35	0.45	0.45	0.45	0.5	1																			
LBBB	0.33	0.47	0.47	0.33	0.42	0.33	0.38	0.38	1																		
LQRSV	0.38	0.42	0.42	0.38	0.47	0.38	0.42	0.42	0.45	1																	
NSIVCB	0.45	0.35	0.35	0.45	0.45	0.45	0.5	0.5	0.38	0.42	1																
PR	0.42	0.38	0.38	0.42	0.47	0.42	0.47	0.47	0.4	0.45	0.47	1															
PAC	0.46	0.34	0.34	0.46	0.44	0.46	0.49	0.49	0.36	0.41	0.49	0.46	1														
PVC	0.42	0.38	0.38	0.42	0.47	0.42	0.47	0.47	0.4	0.45	0.47	0.5	0.46	1													
LPR	0.5	0.3	0.3	0.5	0.4	0.5	0.45	0.45	0.33	0.38	0.45	0.42	0.46	0.42	1												
LQT	0.35	0.45	0.45	0.35	0.45	0.35	0.4	0.4	0.47	0.47	0.4	0.42	0.39	0.42	0.35	1											
QAb	0.2	0.4	0.4	0.2	0.3	0.2	0.25	0.25	0.38	0.33	0.25	0.28	0.24	0.28	0.2	0.35	1										
RAD	0.45	0.35	0.35	0.45	0.45	0.45	0.5	0.5	0.38	0.42	0.5	0.47	0.49	0.47	0.45	0.4	0.25	1									
RBBB	0.4	0.4	0.4	0.4	1	0.4	0.45	0.45	0.42	0.47	0.45	0.47	0.44	0.47	0.4	0.45	0.3	0.45	1								
SA	0.5	0.3	0.3	0.5	0.4	0.5	0.45	0.45	0.33	0.38	0.45	0.42	0.46	0.42	0.5	0.35	0.2	0.45	0.4	1							
SB	0.5	0.3	0.3	0.5	0.4	0.5	0.45	0.45	0.33	0.38	0.45	0.42	0.46	0.42	0.5	0.35	0.2	0.45	0.4	0.5	1						
SNR	0.45	0.25	0.25	0.45	0.35	0.45	0.4	0.4	0.28	0.33	0.4	0.38	0.41	0.38	0.45	0.3	0.15	0.4	0.35	0.45	0.45	1					
STach	0.42	0.38	0.38	0.42	0.47	0.42	0.47	0.47	0.4	0.45	0.47	0.5	0.46	0.5	0.42	0.42	0.28	0.47	0.47	0.42	0.42	0.38	1				
SVPB	0.46	0.34	0.34	0.46	0.44	0.46	0.49	0.49	0.36	0.41	0.49	0.46	1	0.46	0.46	0.39	0.24	0.49	0.44	0.46	0.46	0.41	0.46	1			
TAb	0.3	0.5	0.5	0.3	0.4	0.3	0.35	0.35	0.47	0.42	0.35	0.38	0.34	0.38	0.3	0.45	0.4	0.35	0.4	0.3	0.3	0.25	0.38	0.34	1		
TInv	0.3	0.5	0.5	0.3	0.4	0.3	0.35	0.35	0.47	0.42	0.35	0.38	0.34	0.38	0.3	0.45	0.4	0.35	0.4	0.3	0.3	0.25	0.38	0.34	0.5	1	
VPB	0.42	0.38	0.38	0.42	0.47	0.42	0.47	0.47	0.4	0.45	0.47	0.5	0.46	1	0.42	0.42	0.28	0.47	0.47	0.42	0.42	0.38	0.5	0.46	0.38	0.38	1

Figure 2.33: PhysioNet/CinC 2020 Challenge [40] provided reward matrix W for the diagnoses evaluated in the classification task.

Chapter 3

Gradient Boosting Tree Ensemble

In this chapter, I propose an approach for the classification of 12-Lead ECGs using manual feature engineering and an ensemble of gradient boosted trees. The work contained in this chapter is part of the Computing in Cardiology 2020 PhysioNet Challenge [40], where teams propose algorithms for the multi-label, multi-class classification of 12-lead ECG signals. The submitted work, "Multilabel 12-Lead Electrocardiogram Classification Using Gradient Boosting Tree Ensemble", published [55] in Computing in Cardiology (CinC) 2020, has contributions and edits from Dr. Abram Hindle and Dr. Sunil Vasu Kalmady.

We would like to thank Eric Ly and Leiah Luoma of the Canadian VIGOUR Center for their help and guidance during our research journey.

What to take away from this chapter

- a method for taking variable length signals and extracting tabular signal processing derived features for use as inputs to shallow classifiers;
- the provided dataset is an incomplete approximation of real-world ECGs, as shown in the poor results from the official test set challenge score;
- additional work, including deriving multi-dimensional features using many leads simultaneously, may lead to better classification score and accuracy.

Abstract

The 12-lead electrocardiogram (ECG) is a commonly used tool for detecting cardiac abnormalities such as atrial fibrillation, blocks, and irregular complexes. For the PhysioNet/CinC 2020 Challenge, we built an algorithm using gradient boosted tree ensembles fitted on morphology and signal processing features to classify ECG diagnosis.

For each lead, we derive features from heart rate variability, PQRST template shape, and the full signal waveform. We join the features of all 12 leads to fit an ensemble of gradient boosting decision trees to predict probabilities of ECG instances belonging to each class. We train a phase one set of feature importance determining models to isolate the top 1,000 most important features to use in our phase two diagnosis prediction models. We use repeated random sub-sampling by splitting our dataset of 43,101 records into 100 independent runs of 85:15 training/validation splits for our internal evaluation results.

Our methodology generates us an official phase validation set score of 0.476 and test set score of -0.080 under the team name, CVC, placing us 36 out of 41 in the rankings.

3.1 Introduction

The electrocardiogram (ECG), when correctly interpreted, is an effective tool for detecting cardiac diseases. Despite much research in computerized interpretations of ECGs, trained human over-reading and confirmation is required and emphasized in published reports [45, 33]. This work classifies standard 12-lead ECGs to their clinical diagnosis as part of the *PhysioNet/CinC 2020 Challenge* [40]. We develop a multi-label classification algorithm using entropy and signal processing inspired features and a gradient boosting decision tree ensemble.

3.1.1 Dataset & Scoring Criteria

The official phase dataset contains a total of 43,101 ECG records. Each record contains a set of one or more SNOMED CT codes, with only a subset of 27 codes evaluated in the challenge. The challenge objective is to maximize the metric: $\sum_{ij} w_{ij} a_{ij}$. Given a set of diagnoses $C = \{c_i\}$, we compute a confusion matrix $A = [a_{ij}]$ where a_{ij} contains records that are classified as class c_i and belong to class c_j. The weights $W = [w_{ij}]$, are set by the challenge to indicate clinical similarity between classes. Refer to Perez Alday *et al.* [40] for the description of the challenge scoring function weights and ECG dataset.

3.2 Methodology

Our approach is inspired by existing methods which use feature engineering and shallow learning classifiers [24]. Figure 3.1 shows an overview of our learning algorithm pipeline from first cleaning and preprocessing the ECG, to then extracting the full waveform, heartbeat template, and heart rate variability features, finally using these features as input to our binary classifiers.

We rely on the *NeuroKit2* (version `0.0.40`) neurophysiological signal processing library for ECG signal cleaning, PQRST annotation, signal quality calculation, and heart rate variability metrics [34]. We also use the time series feature extraction library *tsfresh* (version `0.16.0`) for analysis of the PQRST

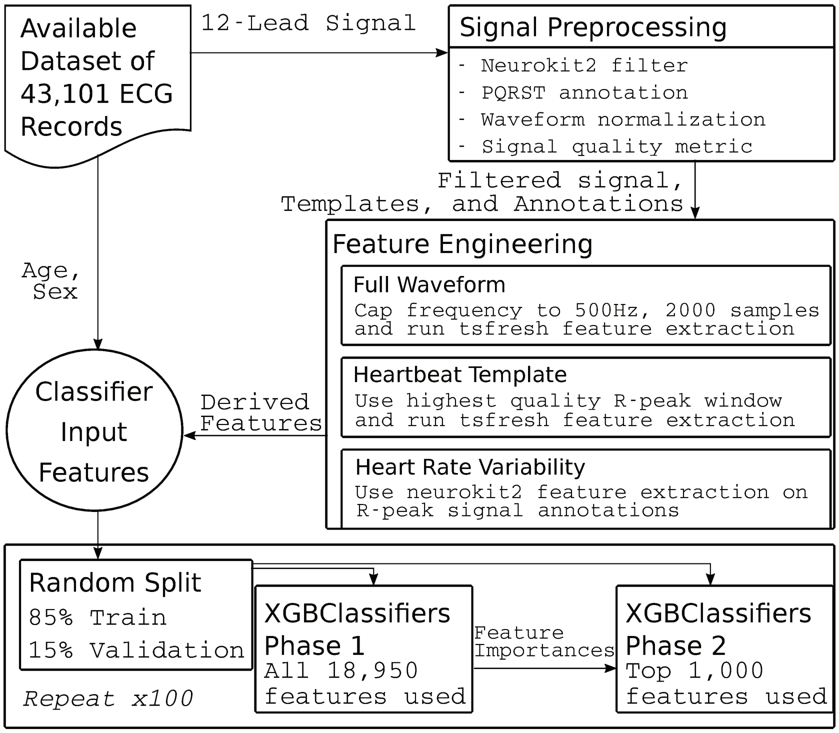


Figure 3.1: Methodology overview. Feature engineering is performed concurrently for each lead then concatenated.

template and the full waveform [20].

3.2.1 Signal Pre-processing

First we perform signal pre-processing to normalize and clean the raw ECG signal. Slow drift and DC offset are removed with a Butterworth highpass filter followed by smoothing using a moving average kernel of 0.02 seconds. Each of the cleaned leads are independently annotated with the PQRST peaks, the PRT onsets, and PRT offsets.

We isolate one candidate heart beat signal for each lead by segmenting heart beat windows as a -0.35 to 0.5 second window around each R-peak, shortening to a -0.25 to 0.4 second window if the mean heart rate exceeds 80 beats per minute. We create an ECG signal quality metric by interpolating the distance of each QRS segment from the average QRS segment in the data. ECG signal quality is therefore relative for each step in the entire length of the signal, where 1 corresponds to beats that are closest to the average QRS and 0 corresponds to beats that are most distant to the average QRS. We use the PQRST beat window with the highest signal quality as our candidate lead heartbeat template.

3.2.2 Feature Engineering

Our engineered features are categorized as one of three categories. Full waveform features are derived using the end-to-end ECG signal. Template features are constructed from the extracted PQRST window during pre-processing. Heart rate variability features rely on the relative distances between each R-peak. Each extraction technique is performed independently per lead and concatenated together prior to classification.

For full waveform and heartbeat template features, we use the cleaned ECG signal and apply the *tsfresh* feature extraction library. For full waveform features, we cap the signal sampling rate to a maximum of 500Hz before limiting the signal to the middle 2,000 samples to remove starting and trailing artifacts. Template features are derived from the isolated heart beat window with highest signal quality. Using the default feature extraction settings, we generate 763 template and 763 full waveform features per lead. The extracted features include autoregressive model coefficients, change quantiles, aggregate linear least-squares regression trends, peak counts, sample/approximation entropy, energy, continuous waveform transform coefficients, fast fourier transform coefficients, and other descriptive statistics of the signal.

Heart rate variability (HRV) features are generated from the cleaned signal and corresponding R-peak annotations using *NeuroKit2*. We use the default feature extraction settings and generate 53 different HRV features per lead. HRV features include: mean, median, standard/absolute deviation, and interquartile range of the RR intervals; standard deviation of the successive differences between RR intervals; proportion of RR intervals greater than 50/20ms over total RR intervals; and geometric indices measuring triangular interpolation of the RR interval distribution.

For each 12-lead record we combine all three categories of engineered features with the age and sex parsed from the ECG record metadata. We arrive at a $12 \cdot (763 + 763 + 53) + 2 = 18,950$ length feature vector per 12-lead record.

3.2.3 Classification

We train a XGBoost binary classifier for each of the 27 clinical diagnoses, using `xgboost@1.1.1` [18]. We sample each training instance with a selection probability proportional to the regularized absolute value of the gradients. Early stopping is set to 20 rounds with binary logistic regression as our objective function.

We use the evaluation scoring weights as instance sample weights, capping positive examples to a 0.5 threshold. For example, when training the 1st degree atrioventricular block (IAVB) classifier we consider instances of bradycardia (Brady), incomplete right bundle branch block (IRBBB), prolonged PR interval (LPR), sinus arrhythmia (SA), and sinus bradycardia (SB) as positive examples with 0.5 weight. Other labels that have scoring function weights below 0.5 are treated as negative examples with a sample weight of 1. To account for the dataset label imbalance, we further scale the positive weight using the number of negative samples over the positive samples in the training set split.

Our classification models are trained in two phases. First, we randomly sub-sample our total dataset, splitting our 43,101 records into an 85:15 training/validation set split. In the phase one, we train using all 18,950 features to estimate the feature importances. Feature importance is defined as the model reported gain in accuracy contributed by the feature over all branches in the decision tree. We average the importances outputted by the 27 binary classifiers to get the mean importance for each feature. We rank all of the features by their mean importance and keep the top 1,000 important features. In phase two, we train new models using the same training and validation split but limiting the classifier input to the top 1,000 most important features. This process is repeated 100 times, exhausting our available dataset.

For the submission component of the competition, we omit training phase one of our classification models due to insufficient computing resources and time constraints. We overcome this limitation by using the phase one models that we trained locally. All 100 phase one model feature importances are averaged together to generate a overall mean feature importance, using our entire available dataset. Our challenge submission's classification model only needs to train the phase two set of classifiers, using the top 1,000 features that we computed as a prior.

3.3 Results

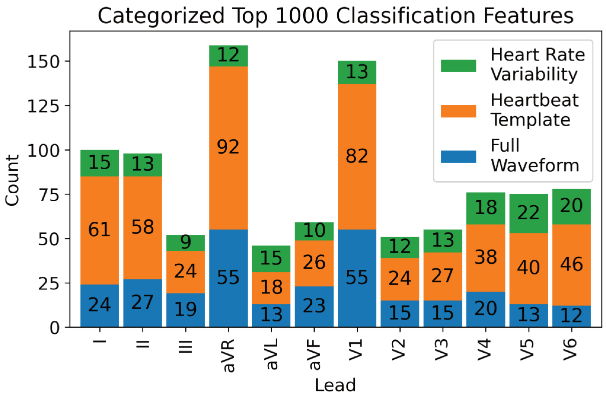


Figure 3.2: Count of lead and feature categories comprising the top 1,000 features. Age, but not sex, is important meta.

A categorical visualization of the top 1,000 features used in the challenge submission model, grouped by lead and feature type, is shown in Figure 3.2. Most of the features are derived from leads `aVR` (159 features) and `V1` (150 features). The `aVL` lead is least represented with only 46 derived features used by the phase 2 classifier. The heartbeat template category containing 536 features is the most numerous feature type. There are 291 full waveform features and 172 heart rate variability features. The age meta feature parsed from the ECG record header is also used.

We present our metrics from the 15% validation splits of the dataset for phases one and two of our classification models in Figure 3.3. Our phase two models have higher mean values for all classification metrics except for accuracy. Using the smaller set of features, our classification metric variances are more closely centered around the mean. Our methodology attains a phase two mean challenge metric score of 0.486. Additionally, we attain phase two mean values for AUROC of 0.891, AUPRC of 0.389, accuracy of 0.254, overall F_1 score of 0.369, F_β of 0.428, and G_β measure of 0.223 using $\beta = 2$.

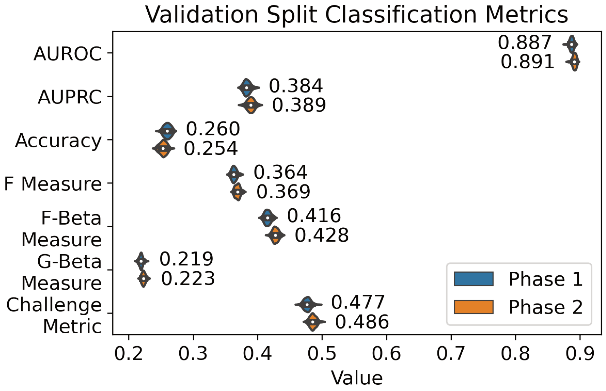


Figure 3.3: Summary of classification metrics over 100 experiments on all labels. Annotations indicate mean value.

Our model's top three best classified labels are normal sinus rhythm (SNR, $\bar{F}_1$: 0.924), left bundle branch block (LBBB, $\bar{F}_1$: 0.840), and sinus tachycardia (STach, $\bar{F}_1$: 0.777). A summary of our phase two F_1 scores on the 100 validation splits for each label is shown in Figure 3.4.

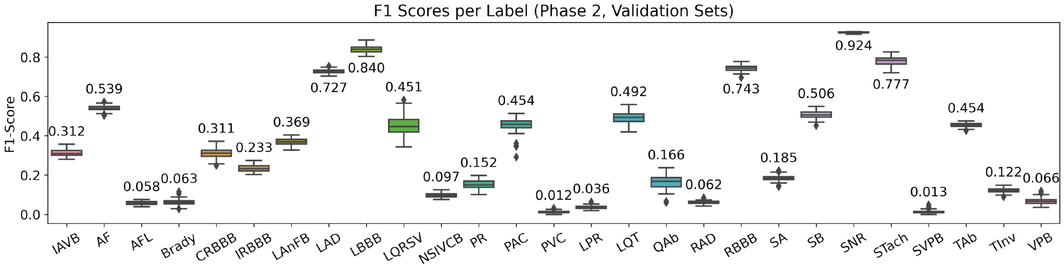


Figure 3.4: Phase two label-wise validation set F_1 scores over 100 independent runs. Annotations indicate mean value.

Furthermore, we run a Pearson correlation coefficient test between the label F_1 means and the label counts within our dataset. The statistical test reveals a Pearson correlation coefficient of 0.602 at a p-value of $9.0 \cdot 10^{-4}$. This result suggests that a positive linear correlation exists between the label occurrence in our dataset and our classification model's F_1 score.

Our methodology achieves a challenge score of 0.476 on the official validation set and -0.080 for the official test set, ranking team CVC at 36 of 41 teams.

3.4 Discussion & Future Work

Despite the label specific scaling of our dataset, the correlation between the label occurrence with the F_1 scores suggest further improvements are necessary to mitigate label imbalance. The label imbalance may be addressed by adding more low occurrence disorders into the existing corpus of ECG records. Synthesizing new records of low occurrence disorders to use as training data may also prove promising. Additionally, exploration of new features to use as classifier inputs may reveal common characteristics of specific heart disorders that are currently missing.

Our approach, although applicable to 12-lead ECGs, perform feature extraction on each lead separately before concatenating the features together for classification. We believe that further improvements can be made utilizing feature extraction approaches capable of handling multi-dimensional time series data.

Our approach does not use additional external datasets, nor do we modify any of the labels provided in the available dataset. We anticipate that further corrections in the ECG diagnosis labels, and including more ECG records, would enable our approach to achieve more competitive competition scores.

We acknowledge that our internal results and corresponding figures report optimistic values for the classification metrics, as our internal split of the dataset does not include a hold-out test set. We rely on the hold out test set from the challenge organizers to fairly evaluate our challenge score. Future work includes replicating our method using a local training, validation, and test set split, reporting label-wise F_1 on the test set.

The requirement of the challenge to train a model on a hold out training set added additional engineering complexity that could not be fully addressed in our final submission. The computation time of training the phase one feature importance models exceeded the allocated time constraints set by the challenge, using their provided cloud virtual machines. Our workaround therefore relies on the feature importances generated locally, using the available, released data. The feature importances used for the challenge submission model may not match the distribution of feature importances of the hold out training set.

3.5 Conclusion

We create an algorithm for the classification of 27 heart conditions using signal processing inspired feature engineering and an XGBoost tree ensemble classifier. We combine a set of 18,950 features from full waveform, heartbeat template, and heart rate variability groups. Using 100 repeated random subsampling of 85:15 train/validation, we train models to get feature importances and distilled out 1,000 most important features. Using this reduced set of 1,000 features, we retrain our models and achieve a mean challenge score of 0.486 on our validation split. For our team, *CVC*, the official phase challenge scores are 0.476 on the validation set and -0.080 on the test set. We attain a rank of 36 of 41 qualifying teams.

Errata

From Section 3.2.2, we set the `ComprehensiveFCParameters` parameter during *tsfresh* feature extraction. For any undefined features, such as the heart rate variability feature set on signals where no PQRST annotations could be generated, `NaN` placeholders are set.

For Section 3.2.3 on the XGB classifier configuration, we use the `dart` booster method with the `gpu_hist` tree method and the `gradient_based` sampling method. All other model initialization parameters are left to their defaults.

Chapter 4

Beat to Sequence Autoencoders

In this chapter, I propose an approach for the 12-lead ECG classification problem using a series of autoencoders to learn a dense embedding representation underlying record. The work contained in this chapter is adapated from "Multilabel 12-Lead Electrocardiogram Classification Using Beat to Sequence Autoencoders" and has been submitted to the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2021. This work has manuscript revisions from Dr. Abram Hindle and Dr. Sunil Vasu Kalmady, as well as the contribution of Figure 4.1 from Amir Salimi.

What to take away from this chapter

- Our beat to sequence autoencoder approach is statistically more sensitive than the XGBoost ensemble method for detecting IRBBB, LAnFB, PR, and RAD;
- However, for overall ECG classification metrics, the beat to sequence autoencoder is a less effective classifier than the prior XGBoost ensemble approach;
- Future work, combining the benefits of the autoencoder embeddings and the shallow tree classifier may lead to a better overall classifier.

Abstract

The 12-lead electrocardiogram (ECG) measures the electrical activity of the heart for physicians to use in diagnosing cardiac disorders. This paper investigates the multi-label, multi-class classification of ECG records into one or more of 27 possible medical diagnoses. Our multi-step approach uses conventional physiological algorithms for segmentation of heartbeats from the baseline signals. We stack a heartbeat autoencoder over heartbeat windows to make embeddings, then we encode this sequence of embeddings to make an ECG embedding which we then classify on. We utilize the public dataset of 43,101 available ECG records provided by the *PhysioNet/CinC 2020 challenge*, performing repeated random subsampling and splitting the available records into 80% training, 10% validation, and 10% test splits, 20 times. We attain a mean test split challenge score of 0.248 with an overall macro F_1 score of 0.260 across the 27 labels.

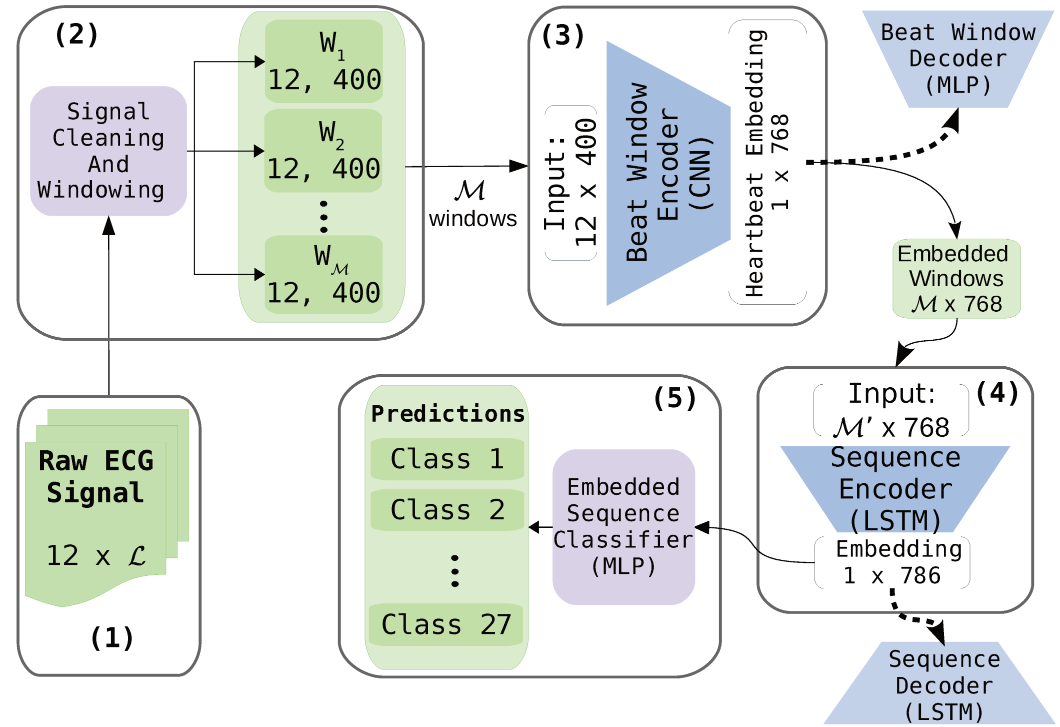


Figure 4.1: Methodology overview. From (1) $\mathcal{L}$-sized ECG signal, we (2) extract $\mathcal{M}$ heartbeat windows prior to (3) heartbeat & (4) sequence autoencoding. Finally we (5) pass our sequence embedding to the classifier to output our 27 predictions.

4.1 Introduction

Although the electrocardiogram (ECG) is an effective tool for detecting cardiac diseases, the analysis of the ECG is a specialized skill requiring training and human over-reading of computerized interpretations. Our work extends the *PhysioNet/CinC 2020 Challenge* [40] involving multi-label classification of ECGs, which are 12 channel 500-1000Hz signals with 27 labels indicating cardiologist diagnoses. Our novel component is the use of sequentially windowed embeddings to classify the ECGs. We train a two component signal autoencoder algorithm, encoding the heartbeat windows, then the sequence of window embeddings, before using the sequence bottleneck embedding for classification. We extend and compare against our prior work by learning autoencoded features rather than using manually engineered features [55].

4.2 Related Work

We are inspired by prior work that uses autoencoders to generate features for signal classification [26, 21]. Recent advancements in machine learning and available data have heralded an influx of multi-lead ECG classification algorithms [43, 19, 54, 6, 39, 55]. We extend our prior work by using neural networks over feature engineering with gradient boosted tree classifiers [55]. Despite the large improvements in automated ECG classification, trained human over-reading and cardiologist confirmation is still mandated during use in the clinical setting [45, 33].

4.2.1 Challenge Dataset and Task Specification

Refer to Perez Alday *et al.* [40] for the ECG sources and competition rules. The challenge provides 43,101 ECG records where each record is labelled as one or more of 111 possible diagnoses. The evaluated 27 label subset is shown in Table 4.1.

We reuse the scoring function in preparation for the 2021 challenge, which extends this task and adds a 2-lead classification variant. We want to maximize the following scoring function: $\sum_{ij} w_{ij} a_{ij}$. Provided predictions $C = \{c_i\}$, we create a confusion matrix $A = [a_{ij}]$ where a_{ij} indicates a record classified as class c_i belongs to class c_j. The weights $W = [w_{ij}]$, shown in Figure 4.2, are challenge defined to provide partial reward for incorrect predictions.

4.3 Methodology

We propose a staged neural network architecture for autoencoding the extracted heartbeats, autoencoding the sequence of heartbeat embeddings, and training a multi layer perceptron classifier. An overview is shown in Figure 4.1. Using 20 repeated random subsampling, we split our 43,101 available ECG records into 80% training, 10% validation, and 10% test splits. No label proportion stratification of the splits occurred.

Table 4.1: Evaluated labels, count and percentage in dataset.

Abbr.	Diagnosis	Count (%)
IAVB	1st degree av block	2394 (5.6%)
AF	atrial fibrillation	3475 (8.0%)
AFL	atrial flutter	314 (0.7%)
Brady	bradycardia	288 (0.7%)
CRBBB	complete right bundle branch block	683 (1.6%)
IRBBB	incomplete right bundle branch block	1611 (3.7%)
LAnFB	left anterior fascicular block	1806 (4.2%)
LAD	left axis deviation	6086 (14.1%)
LBBB	left bundle branch block	1041 (2.4%)
LQRSV	low QRS voltages	556 (1.3%)
NSIVCB	nonspecific intraventricular conduction	997 (2.3%)
PR	pacing rhythm	299 (0.7%)
PAC	premature atrial contraction	1729 (4.0%)
PVC	premature ventricular contractions	188 (0.4%)
LPR	prolonged PR interval	340 (0.7%)
LQT	prolonged QT interval	1513 (3.5%)
QAb	Q wave abnormal	1013 (2.4%)
RAD	right axis deviation	427 (1.0%)
RBBB	right bundle branch block	2402 (5.6%)
SA	sinus arrhythmia	1240 (2.9%)
SB	sinus bradycardia	2359 (5.5%)
SNR	sinus rhythm	20846 (48.4%)
STach	sinus tachycardia	2402 (5.6%)
SVPB	supraventricular premature beats	215 (0.5%)
TAb	T wave abnormal	4673 (10.8%)
TInv	T wave inversion	1112 (2.6%)
VPB	ventricular premature beats	365 (0.8%)

4.3.1 Signal Preprocessing

We use the *NeuroKit2* (`v0.0.40`) neurophysiological signal processing library [34] to annotate our ECG signals and the *SciPy* (`v1.5.2`) family of Python packages for signal filtering and statistical tests [52]. The ECG cleaning approach removes slow drift and DC offset using a Butterworth highpass filter (5Hz, $Q = 0.5$) then smooths the signal using a moving average kernel of 0.02 seconds. The R-peaks, or heartbeat locations, are annotated for each of our 12 signals.

Due to variable quality of sensor placements or noise artifacts caused by

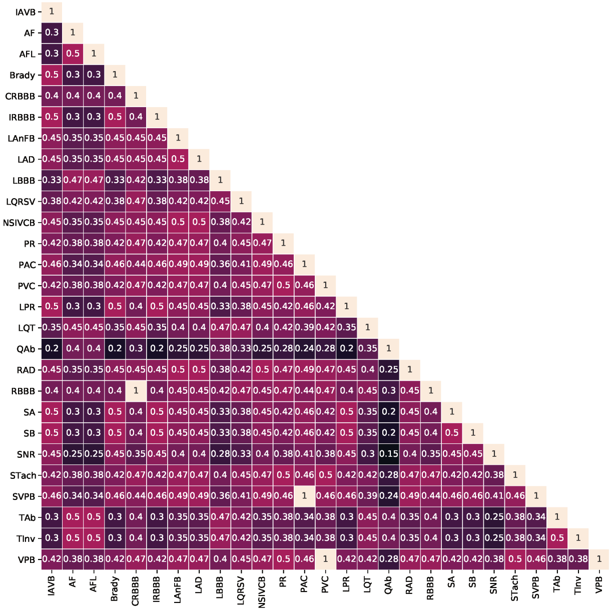


Figure 4.2: Evaluation scoring function weights per label.

patient movements, the independent heartbeat annotations per channel may not be congruent within the ECG record. We address this limitation with a kernel density estimation function fitted to the indices of all the R-peak annotations. The bandwidth is the mean channel-wise heart rate multiplied by a $\frac{1}{4}$ scaling factor. Next, we determine the peaks in the R-peak probability densities by finding all local maxima relative to their neighboring values. We apply a cutoff threshold, dropping peaks that are over two standard deviations away from the mean peak value.

Given the overall R-peak indices for our 12-channel signal, we resample the entire signal such that the mean distance between each R-peak is 400 samples. We slice windows of size 400, positioning the R-peak to occur at one-third of the length, and ignoring windows that do not contain 400 samples. An l_2 normalization step is applied on all remaining windows. We use the Breunig *et al.* local outlier factor algorithm [14] to find the most abnormal heartbeat in the ECG.

Our signal processing steps allow us to extract $\mathcal{M}$ normalized, fixed size heart beat windows from arbitrary length 12-channel ECG records. A full example of the entire signal processing and windowing procedure is provided

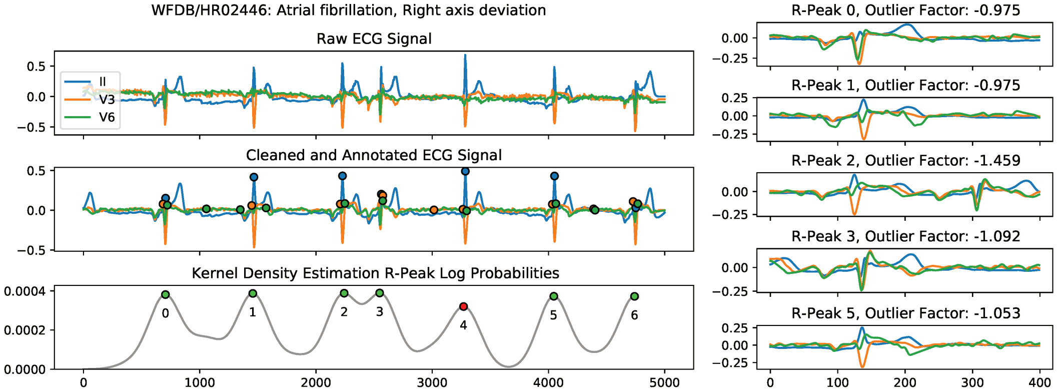


Figure 4.3: Signal processing from the raw signal to the annotated intermediary signal, to output heartbeat windows. Window 4 is dropped due to the cutoff threshold. Window 6 is dropped due to insufficient window size. Window 2 is the most abnormal heartbeat with the minimum outlier factor of the given windows. Only 3 of the 12 leads (`II`, `V3`, and `V6`) are shown for clarity.

in Figure 4.3.

4.3.2 Heartbeat Autoencoder

We rely on the dimensionality reducing properties of autoencoders [25] to encode our heartbeat windows into an embedding, then concatenate our embeddings to get a representation of the overall ECG. Our heartbeat autoencoder converts our 12 × 400 heartbeat windows into embeddings of size 768.

The encoder has 970,420 trainable parameters among three convolutional blocks followed by a linear layer to generate the embedding. Each block contains a convolutional layer followed by batch normalization, a ReLU activation, and a dropout normalization layer. The decoder architecture, with 5,707,456 trainable parameters, contains two linear layers separated by a ReLU activation and a dropout normalization layer with a Tanh nonlinearity applied to the outputs. See Table 4.2 for the heartbeat autoencoder architecture.

We use stochastic gradient descent (SGD) with 0.9 momentum, to optimize our mean square error (MSE) objective. We cyclically oscillate our learning rate between 1.0×10^{-3} and 1.0×10^{-5}. Training stops if the validation loss fails to attain a new minimum value after 3 epochs or after 100 epochs.

Table 4.2: Heartbeat autoencoder neural network architecture.

Block	**Modules**	**Output Shape**
Enc1	Conv1d(12, 16, 164), BatchNorm1d(16), ReLU(), Dropout(p=0.1)	[$\mathcal{M}$, 16, 237]
Enc2	Conv1d(16, 20, 128), BatchNorm1d(20), ReLU(), Dropout(p=0.1)	[$\mathcal{M}$, 20, 110]
Enc3	Conv1d(20, 24, 64), BatchNorm1d(24), ReLU(), Dropout(p=0.1)	[$\mathcal{M}$, 24, 47]
Enc4	Flatten(), Linear(1128, 768)	[$\mathcal{M}$, 768]
Dec1	Linear(768, 1024), ReLU(), Dropout(p=0.1)	[$\mathcal{M}$, 1024]
Dec2	Linear(1024, 4800), View(400, 12), Tanh()	[$\mathcal{M}$, 400, 12]

4.3.3 Embedding Sequence Autoencoder and Classifier

The number of heartbeats extracted from an ECG record varies between 2 beats up to over 3,000 beats. We limit this sequence length, capping the number of heartbeat embeddings used $\mathcal{M}'$ to 20. Beginning with the abnormal heartbeat, we iteratively pick the rest of the candidate heartbeats by prepending the left neighbors and appending the right neighbors. We stop when the neighbors are exhausted or 20 heartbeats are chosen. For records with fewer than 20 beats, empty positions are masked and do not contribute to the loss.

The sequence autoencoder is symmetrical, with identical encoder and decoder architectures. A two layer LSTM module with input and hidden sizes

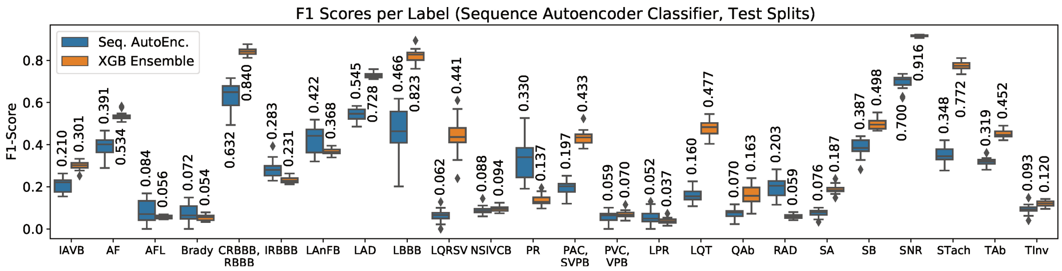


Figure 4.4: Test set F_1 scores of all 27 labels compared with our prior XGBoost ensemble method [55]. Mean values annotated.

Table 4.3: Sequence autoencoder and classifier architecture.

Block	**Modules**	**Output Shape**
Encoder	LSTM(768, 768, num_layers=2, dropout=0.1)	Hidden [768]
Decoder	LSTM(768, 768, num_layers=2, dropout=0.1)	Sequence [$\mathcal{M}'$, 768]

Classifier	Linear(768, 256), ReLU()
	Dropout(p=0.1), Linear(256, 27)
		Predictions
		[27]

set to 768 and dropout of 0.1 is used, containing 9,449,472 parameters. It encodes our $\mathcal{M}' \times 768$ heartbeat embeddings into a bottleneck of size 768. A multilayer perceptron consisting of two linear layers, separated by a ReLU and dropout layer ($p = 0.1$) takes the sequence embedding of 768, computes a hidden representation of 256, and outputs 27 label probabilities. Our classifier has 203,803 parameters. See Table 4.3 for architecture design.

To mitigate internal validity risk the training, validation, and test splits are reused from the heartbeat autoencoder experiment. Our pre-trained heartbeat encoder is frozen and does not update during the training of the sequence autoencoder. We train the sequence autoencoder and classifier simultaneously using SGD and a cyclic learning rate. Our overall loss function is the autoencoder MSE loss added to the binary cross entropy (BCE) classifier loss. We scale the BCE weights to the count of negative samples over the positive samples in the training set split.

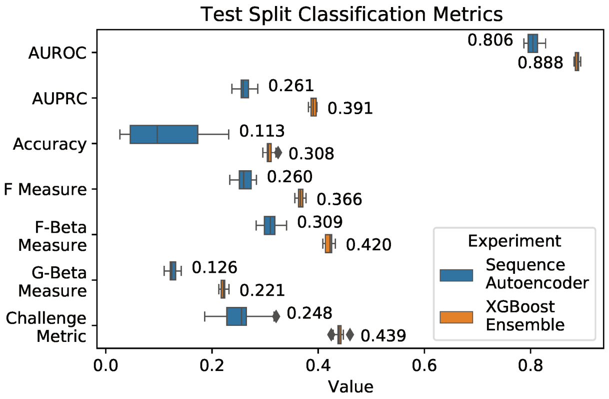


Figure 4.5: Classification metric summary of 20 experiments compared to XGBoost ensembles [55]. Mean values annotated.

Training stops if the validation set challenge score fails to improve after 30 epochs or 200 epochs pass. We use the highest validation set scoring model from all epochs and calculate the challenge metrics on the test set split, setting thresholds to maximize the training data receiver operating characteristic.

4.4 Results and Discussion

We compare our results with our prior XGBoost ensemble classifier [55]. Label-wise test F_1 scores can be found in Figure 4.4. Using the Wilcoxon signed rank test, our autoencoder F_1 means statistically outperform our prior work in detecting incomplete right bundle branch block (IRBBB, $p = 1.9 \times 10^{-6}$), left anterior fascicular block (LAnFB, $p = 9.4 \times 10^{-3}$), pacing rhythm (PR, $p = 1.9 \times 10^{-6}$), and right axis deviation (RAD, $p = 1.9 \times 10^{-6}$).

Overall test classification metrics is shown in Figure 4.5. Our methodology achieves a test split mean *PhysioNet/CinC 2020 Challenge* score of 0.248, AUROC of 0.806, AUPRC of 0.261, accuracy of 0.113, macro F_1 score of 0.260, F_β of 0.309, and G_β of 0.126 using $\beta = 2$. Our autoencoder is worse than our shallow classifier on all summary metrics. Our results cannot be compared with official rankings because the challenge evaluates the algorithms on secret hold-out test sets.

Our methodology trains a neural network using the general shapes of heartbeat windows and indirectly models the overall signal by looking at consecutive heartbeats. Consequentially, due to the variable distances of the R-peaks within an ECG record, portions of the ECG signal not bounded between a heartbeat window are dropped. Because we resample the overall signal to ensure heartbeat windows are 400 samples long, we drop heart rate information. Currently we do not capture features like average heart rate or changes in heart rate over time. Additionally, because of the l_2 normalization of the heartbeat windows, we also do not capture the original signal amplitudes and voltage changes. Future work should expand on our findings to incorporate heart rate velocities, raw amplitudes, and continuous full signal characteristics.

4.5 Conclusion

Using a signal processing and heartbeat window extraction preprocessing step, we train heartbeat autoencoders to be fed into ECG sequence autoencoders before training a multi-label perceptron to classify 27 heart conditions. We run 20 independent experiments, randomly sampling our available dataset into 80% training, 10% validation, and 10% test set splits. Our methodology achieves a mean unofficial test challenge score of 0.248 with an overall macro F_1 score of 0.260.

Errata

Figure 4.1 incorrectly states the embedding size in stage (4) to be `[1 × 786]`. Corrected, this value is `[1 × 768]`.

Extending Section 4.4, all sequence autoencoders early stopped, running for an average of 70.75 (standard deviation of 22.6) epochs.

Chapter 5

Autoencoder Embeddings with Improved Tree Ensemble

In Chapter 4 I discovered that the autoencoder classification models provided worse overall challenge metrics compared to the gradient boosted tree model proposed in Chapter 3, but are more sensitive in detecting IRBBB, LAnFB, and RAD.

This chapter explores the effectiveness of combining the autoencoder learned embeddings with the manually engineered features to train a new set of gradient boosted tree models. I extend the approaches used in Chapters 3 and 4 with the following research predictions:

RQ1 What is the effect of selecting top features with respect to the labelwise classifiers, compared to averaging the feature importances over all classifiers? I predict that labelwise selection of important features will result in an improved classifier, providing a statistically significant higher challenge metric with a mean difference of over 0.01.

RQ2 How important is feature selection when evaluating the classifier challenge metric? I predict that classifiers that do not perform any feature selection will have a statistically significant lower challenge metric than classifiers that perform feature selection, but there will be no significant difference between aggressive pruning (top 100 features) and moderate pruning (top 1000 features).

RQ3 Will incorporating the sequence embeddings from our deep learning autoencoder improve classification challenge metric? Aligning with my thesis statement, I predict that adding in the deep learning autoencoder will statistically significantly increase the overall challenge metric of the classifiers.

5.1 Methodology

I combine the techniques applied in Section 3.2.2 and Section 4.3.3 to convert the variable length features into fixed length input vectors for all ECG records.

This chapter focuses on the training of `xgboost` [18] binary classifiers for each of the 27 labels selected by the PhysioNet/CinC challenge. I explore ten different configurations of the tabular inputs:

1. **All Features with Embeddings**: I combine the heartbeat features, heart rate variability features, and overall waveform features input vector used in Section 3.2.2 (size 18,950) with the autoencoder sequence embedding vector (size 768) to create a combined input vector of size 19,718 and train an XGBoost classifier for each of the 27 diagnosed labels. These label-wise classifiers provide feature importances for use in configurations 3, 4, 5, and 6.

2. **All Features**: I only use the heartbeat features, heart rate variability features, and overall waveform features to create an input vector of size 18,950 as the input to our label-wise XGBoost classifier ensembles. One XGBoost classifier is trained per label. This configuration is identical to the Phase 1 methodology described in Section 3.2.3. The trained classifiers are subsequently used to provide feature importances for configurations 7, 8, 9, and 10.

3. **Averaged Top 1000 Features with Embeddings**: Using the trained models from Configuration 1, all label-wise classifier feature importances provided by the XGBoost classifiers are averaged together before selecting the top 1,000 features. Using this averaged overall set of 1,000 most important features, one XGBoost classifier is trained per label.

4. **Top 1000 Features with Embeddings**: Starting from Configuration 1, I select the top 1,000 features for each XGBoost diagnosis classifier and retrain a new set of 27 classifiers using the reduced feature set. One XGBoost classifier is trained per label only using the importances of the parent model sharing the same label.

5. **Averaged Top 100 Features with Embeddings**: Using the trained models from Configuration 1, all label-wise XGBoost classifier feature importances are averaged together before selecting the top 100 features. One XGBoost classifier is trained per label using the averaged 100 most important features of all parent XGBoost label-to-classifier pairs.

6. **Top 100 Features with Embeddings**: Starting from Configuration 1, I select the 100 most important features for each XGBoost diagnosis classifier and retrain a new set of 27 classifiers. A separate XGBoost classifier is trained per label using the 100 most important features of the parent classifier sharing the same label.

7. **Averaged Top 1000 Features**: Using Configuration 2, the XGBoost classifier feature importances for all labels are averaged together to select the top 1,000 features. These 1,000 features are used to train a new model for each label. This configuration is identical to the Phase 2 methodology described in Section 3.2.3.

8. **Top 1000 Features**: Starting from Configuration 2, a new set of 27 XGBoost classifiers are trained using the reduced top 1,000 most important features per label classifier. One model is trained per label using only the top 1,000 features of the parent model sharing the same label.

9. **Averaged Top 100 Features**: Using Configuration 2, the XGBoost classifier feature importances for all labels are averaged together to select the top 100 features. Using the averaged 100 most important features of all classifiers, a model is trained for each label.

10. **Top 100 Features**: Starting from Configuration 2, a new set of 27 classifiers are trained using the reduced top 100 most important features per XGBoost binary label classifier. The reduced set of 100 features are specific to the label and are not averaged together between other classifiers.

The significant differences distinguishing these approaches from the top 1000 features approach used in Section 3.2.3 are that the importances of features for each of the labels are now evaluated independently, where the prior experiment used the same reduced set of features for all classifiers.

The "gain" feature importance is used. This is the relative contribution of the provided feature to the overall model, defined as the sum of each feature's contribution for every tree in the XGBoost classifier.

I further revise the inadequate dataset partitioning to use Monte Carlo cross-validation 20 times, randomly partitioning the available corpus of public data into 80% training, 10% validation, and 10% test splits.

For each experiment configuration run, I train an XGBoost binary classifier for each of the 27 diagnosed labels. I use the dropout augmented regression tree booster proposed by Vinayak and Gilad-Bachrach [50] and sample the training instances using probabilities proportional to the training gradients. I use the scoring function reward matrix weights from Figure 2.33 as instance sample weights, capping positive examples to a threshold of 0.5. I further scale the positive samples using the ratio of negative samples over positive samples for the given label and dataset split. During training, if the evaluation set binary logistic regression loss fails to improve after 20 epochs of training, we early stop to mitigate overfitting on the training set.

5.2 Results

Refer to Figure 5.1 and Table 5.1 for the test set split classification metric summaries for all experiment configurations. The experiment configuration with the largest mean challenge score is Configuration 3 "Averaged Top 1000 Features with Embeddings", or the top 1000 features from both the engineered

Table 5.1: Test split classification metrics mean ($\bar{x}$) and standard deviations (σ) for all experiment configurations. Bolded value indicates largest mean for metric category.

#	Experiment		AUROC	AUPRC	Accuracy	F Measure	F-Beta	G-Beta	Challenge Metric
1	All Feats w/ Embd	$\bar{x}$	0.8821	0.3813	0.3137	0.3587	0.4047	0.2160	0.4207
		σ	5.3E−3	4.9E−3	6.8E−3	5.9E−3	8.1E−3	3.9E−3	7.7E−3
2	All Feats	$\bar{x}$	0.8815	0.3809	**0.3139**	0.3582	0.4036	0.2150	0.4206
		σ	5.5E−3	5.0E−3	7.9E−3	6.0E−3	8.6E−3	5.2E−3	1.0E−2
3	Avg Top 1000 Feats w/ Embd	$\bar{x}$	**0.8876**	**0.3900**	0.3068	0.3637	0.4165	**0.2194**	**0.4366**
		σ	4.6E−3	5.5E−3	6.5E−3	4.5E−3	6.0E−3	4.3E−3	7.1E−3
4	Top 1000 Feats w/ Embd	$\bar{x}$	0.8836	0.3837	0.3062	0.3613	0.4128	0.2183	0.4316
		σ	4.8E−3	6.4E−3	5.4E−3	5.4E−3	6.7E−3	4.3E−3	7.0E−3
5	Avg Top 100 Feats w/ Embd	$\bar{x}$	0.8740	0.3740	0.2800	0.3482	0.4062	0.2066	0.4215
		σ	5.1E−3	7.8E−3	9.7E−3	6.5E−3	6.8E−3	5.2E−3	9.9E−3
6	Top 100 Feats w/ Embd	$\bar{x}$	0.8836	0.3876	0.2820	0.3588	0.4190	0.2143	0.4348
		σ	4.9E−3	7.0E−3	6.2E−3	5.1E−3	6.8E−3	4.9E−3	7.8E−3
7	Avg Top 1000 Feats	$\bar{x}$	0.8871	0.3890	0.3085	**0.3640**	0.4165	0.2187	0.4358
		σ	5.1E−3	6.7E−3	6.6E−3	6.2E−3	8.1E−3	5.6E−3	8.1E−3
8	Top 1000 Feats	$\bar{x}$	0.8843	0.3836	0.3075	0.3619	0.4126	0.2179	0.4295
		σ	5.2E−3	5.6E−3	6.4E−3	5.2E−3	7.5E−3	4.4E−3	8.1E−3
9	Avg Top 100 Feats	$\bar{x}$	0.8714	0.3748	0.2800	0.3471	0.4035	0.2057	0.4195
		σ	6.7E−3	6.4E−3	5.3E−3	4.6E−3	6.2E−3	5.1E−3	7.8E−3
10	Top 100 Feats	$\bar{x}$	0.8848	0.3857	0.2830	0.3604	**0.4198**	0.2150	0.4335
		σ	4.5E−3	8.5E−3	7.5E−3	4.0E−3	5.3E−3	3.3E−3	8.2E−3

and autoencoder sources. Although Configuration 3 has the highest mean challenge score, it is only statistically different compared to Configurations 1, 2, 5, and 10. See Figure 5.2 for the Wilcoxon signed-rank test evaluated on all configuration pairs.

5.2.1 Averaged vs Labelwise Feature Selection

For RQ1, I investigate if selecting important features by classifier is an improvement to just averaging all of the classifiers together. If my prediction holds true, I expect to see the averaged top feature configurations perform worse than the top feature configurations. With an alpha of 0.001, I analyze the relevant configuration pairs:

- **1000 Features with Embeddings**: Configuration 3 has a higher challenge score than Configuration 4, but it is not statistically significant.

- **100 Features with Embeddings**: Configuration 5 has a lower challenge score than Configuration 6, and *it is statistically significant*, suggesting labelwise selection of features increases the challenge metric.
- **1000 Features**: Configuration 7 has a higher challenge score than Configuration 8, and *it is statistically significant*, suggesting labelwise selection of features decreases the challenge metric.
- **100 Features**: Configuration 9 has a lower challenge score than Configuration 10, and *it is statistically significant*, suggesting labelwise selection of features increases the challenge metric.

RQ1: When the feature pruning is not aggressive, such as taking the top 1000 features, the improvement in the challenge metric score is inconsistent resulting in no clear conclusion. Labelwise selection of features results in a significantly higher challenge metric for the "100 Features with Embeddings" and "100 Features" configurations, suggesting that labelwise selection of features improves classifier performance when the feature pruning is aggressive.

5.2.2 Effectiveness of Feature Selection

In RQ2, I hypothesize that reducing the number of features passed to the classifier from the original unpruned set of features will improve the challenge metric. If my prediction is accurate, I expect to see the models using "All Features" and "All Features with Embeddings" to have lower challenge scores than all other configurations. Using an alpha of 0.001, I compare the appropriate configurations:

- **From All to Top 1000**: Configuration 1 has a *statistically significant* lower challenge score than Configurations 3 and 4, suggesting that pruning from all 19,718 features down to the top 1000 features is effective in improving the classifier's challenge score. In addition, the non-embedding variants Configuration 2 also has a *statistically significant* lower challenge score compared to Configurations 7 and 8, suggesting that a moderate pruning from all 18,950 features down to the top 1000 features is effective in improving the classifier's challenge score.

- **From All to Top 100**: Configuration 1 has a lower challenge score than Configurations 5 and 6, however it is only statistically significant when compared with Configuration 6 that performs labelwise selection of feature importances. When looking at the non-embedding variants, Configuration 2 has a *statistically significant* lower challenge score than Configuration 10, but no significant difference when compared to Configuration 9. These comparisons suggest that when performing aggressive pruning down to a subset of 100 features, the classifier's challenge metric score improves only if labelwise feature selection is applied.
- **From Top 1000 to Top 100**: Configuration 3 has a *statistically significant* higher challenge score compared to Configuration 5. Configuration 4 has no significant difference in challenge score compared to Configuration 6. Configuration 7 has a *statistically significant* higher challenge score compared to Configuration 9. No significant difference in challenge score is found between Configuration 8 and Configuration 10.

RQ2: When pruning from all available features, moderate pruning to reduce the feature space to 1000 inputs is effective in improving the challenge score of the classifiers. When aggressively pruning to reduce the feature space to 100 inputs, the improvement in challenge score is effective only when labelwise importances are used, reaffirming the results of RQ1.

5.2.3 Adding Embeddings vs Without Embeddings

The hypothesis of RQ3 aims to test if incorporating the autoencoder embeddings has any effect on the challenge score of the resulting classifiers. I analyze the relevant configuration pairs, using an alpha of 0.001:

- **Averaged Top 1000 Features**: Configuration 3 containing the embeddings has a higher challenge score than Configuration 7 which does not have embeddings, but it is not statistically significant.

- **Top 1000 Features**: Configuration 4 containing embeddings has a higher challenge score than Configuration 8 which is missing embeddings, but it is not statistically significant.
- **Averaged Top 100 Features**: Configuration 5 that includes the autoencoder embeddings has a higher challenge score than Configuration 9 which lacks embeddings, but it is not statistically significant.
- **Top 100 Features**: Configuration 6 with the embeddings from the autoencoder has a higher challenge score than Configuration 10 which does not contain embeddings, but it is not statistically significant.

RQ3: Adding autoencoder embeddings into the input vector as additional representation of the ECG record does not result in any statistically significant change in the classifier's challenge metric output. These results suggest that my original prediction is incorrect- the proposed approach of combining deep learning autoencoder embeddings with manually extracted features is actually ineffective for improving classification score.

5.3 Discussion

Bengio discusses in an informal academic research panel [8] current and upcoming deep learning challenges, where he dismisses the viability of engineering deep learning models into old-fashioned symbolic machine learning methods, instead proposing learned attention mechanisms as a viable alternative. This is enforced by the experiment results, as no significant improvement in challenge metric can be attributed to the addition of autoencoder embedding representations alone. Additionally, although the engineering of the two different mechanisms into one shallow machine learning classifier is feasible, we obscure the semantics of feature importance for the embedding features. It is no longer clear how to trace the importances assigned to the embedding features back to their original lead sources.

The public dataset provided by the challenge contained notable irregularities that were not corrected when training the models. Example irregularities include:

- ECG records containing voltage over time changes exceeding physiologically possible voltage measurements
- ECG records with miniscule voltage gain, peaks undiscernible from noise;
- ECG records not labeled as bradycardia despite having resting heart rate of below 60 beats per minute;
- LQRSV ECG records incorrectly labeled as AF, TAb, or SNR;
- TAb labeled ECG records undiscernible from noise;
- inconsistent dataset labeling of AF and AFL;

Additional improvements in the quality of the underlying dataset and the discarding of unusable ECG records would likely improve the performance of the classifiers.

For future work, a replication of this study using other shallow classifiers, such as support vector machines, could provide insight into the relative effectiveness of the gradient boosted trees. Using a wide set of tabular features per ECG record appears to be inefficient, as feature pruning was the most useful regularization technique for improving the challenge metric score. Expert domain specific knowledge for feature generation and selection may prove to be a more effective approach than distilling features from a general purpose time series feature extraction library.

Additionally, the challenge provided dataset contained an additional 88 unused classification labels. Future work could consider applying these unused labels as additional features for classification in a gambit to learn diagnosis correlations or interactions.

5.4 Conclusion

This chapter extends and improves upon the gradient boosted tree models from Chapter 3 using the deep learning autoencoder embeddings from Chapter 4. We discover that the label-wise selection of features, in combination with feature pruning, is effective in improving the classifier's challenge metric score, but adding in autoencoder embeddings has no statistically significant effect on the scoring function. Our configuration that attains the highest mean challenge metric is the "Top 1000 Features with Embeddings" setup, which achieves a test split mean challenge metric of 0.4366, with an overall test split average accuracy of 0.3068.

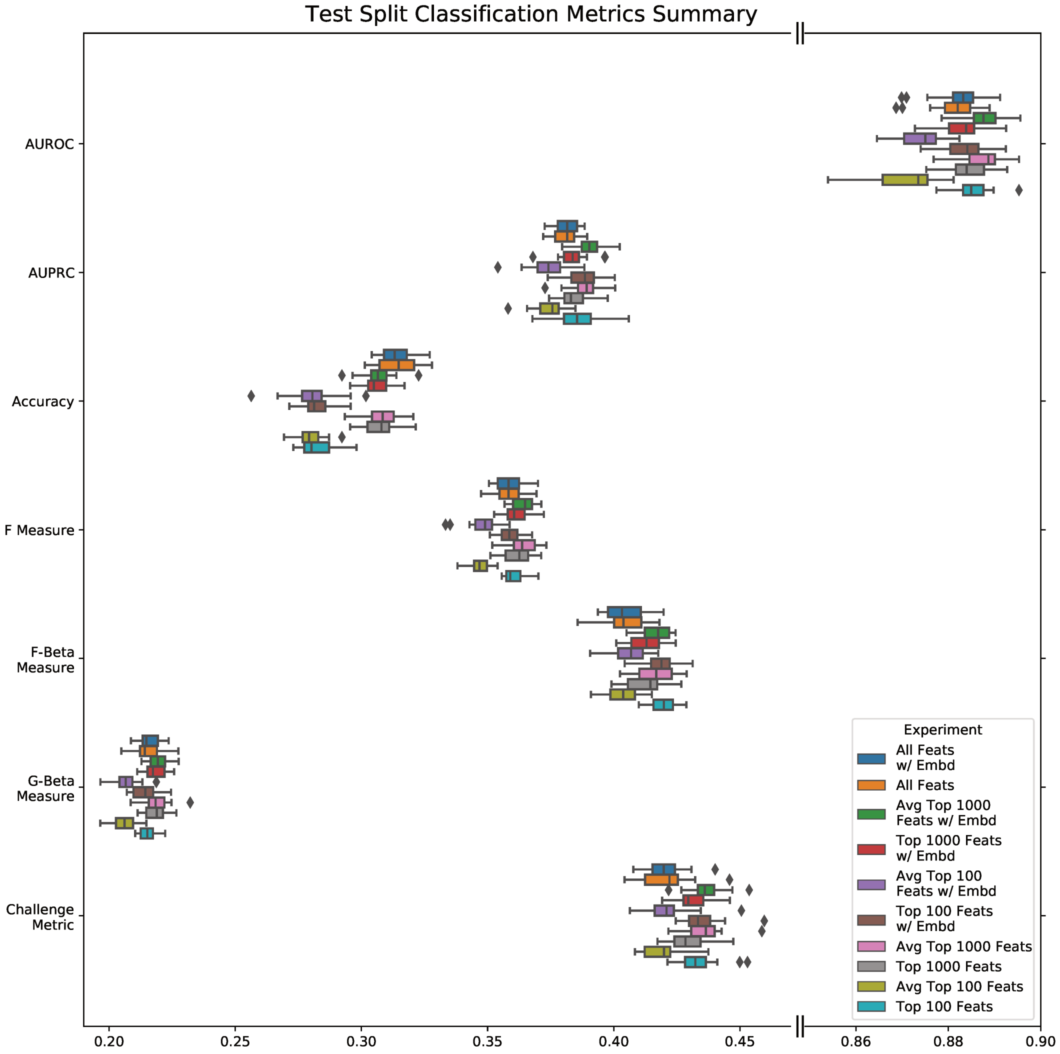
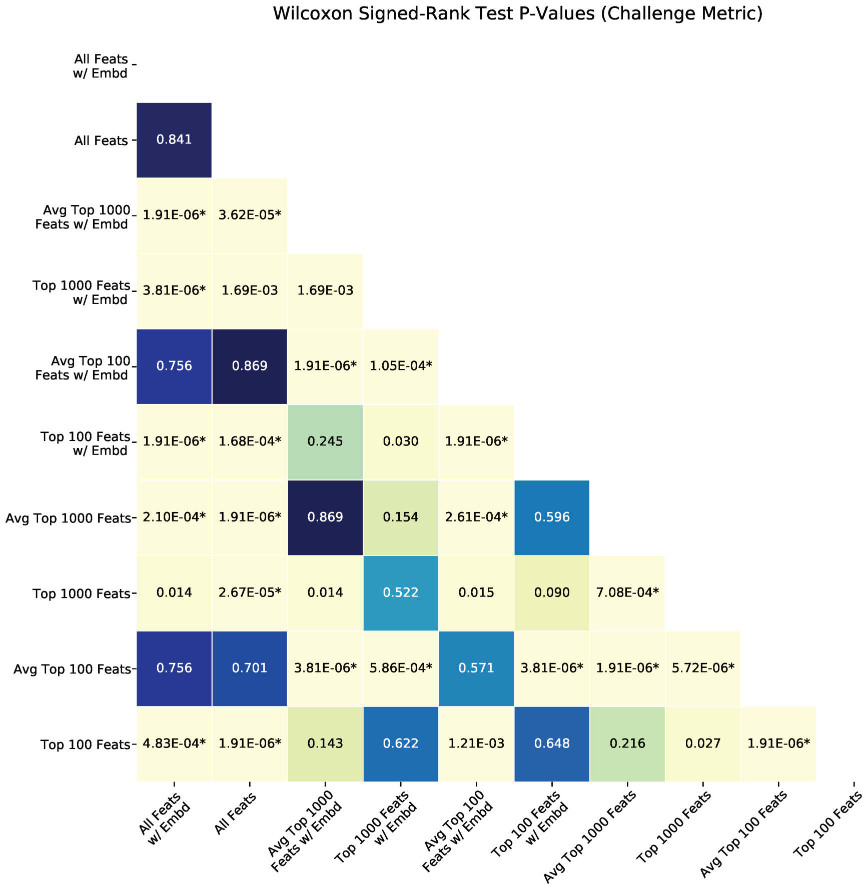


Figure 5.1: Test split classification metrics of XGB ensemble using Configuration 1: All Features with Embeddings; Configuration 2: All Features; Configuration 3: Averaged Top 1000 Features with Embeddings; Configuration 4: Top 1000 Features with Embeddings; Configuration 5: Averaged Top 100 Features with Embeddings; Configuration 6: Top 100 Features with Embeddings; Configuration 7: Averaged Top 1000 Features; Configuration 8: Top 1000 Features; Configuration 9: Averaged Top 100 Features; and Configuration 10: Top 100 Features.



Wilcoxon Signed-Rank Test P-Values (Challenge Metric)

	All Feats w/ Embd	All Feats	Avg Top 1000 Feats w/ Embd	Top 1000 Feats w/ Embd	Avg Top 100 Feats w/ Embd	Top 100 Feats w/ Embd	Avg Top 1000 Feats	Top 1000 Feats	Avg Top 100 Feats	Top 100 Feats
All Feats w/ Embd										
All Feats	0.841									
Avg Top 1000 Feats w/ Embd	1.91E-06*	3.62E-05*								
Top 1000 Feats w/ Embd	3.81E-06*	1.69E-03	1.69E-03							
Avg Top 100 Feats w/ Embd	0.756	0.869	1.91E-06*	1.05E-04*						
Top 100 Feats w/ Embd	1.91E-06*	1.68E-04*	0.245	0.030	1.91E-06*					
Avg Top 1000 Feats	2.10E-04*	1.91E-06*	0.869	0.154	2.61E-04*	0.596				
Top 1000 Feats	0.014	2.67E-05*	0.014	0.522	0.015	0.090	7.08E-04*			
Avg Top 100 Feats	0.756	0.701	3.81E-06*	5.86E-04*	0.571	3.81E-06*	1.91E-06*	5.72E-06*		
Top 100 Feats	4.83E-04*	1.91E-06*	0.143	0.622	1.21E-03	0.648	0.216	0.027	1.91E-06*	

Figure 5.2: Wilcoxon signed-rank test P-values, comparing test split challenge metric distributions of all 6 experiment configurations. Asterisk indicates significantly different challenge metric distribution at an alpha of 0.001.

Chapter 6

Conclusion

In this thesis, I proposed three approaches for the classification of 12-lead ECG records. I demonstrated that ECG records can be classified using traditional signal processing and feature extraction techniques in combination with a shallow gradient boosted tree ensemble algorithm (Chapter 3). I showcased a deep learning ECG record classifier using beat to sequence autoencoders to learn fixed length embeddings from arbitrary length signals (Chapter 4). I proposed a set of experiment configurations, experimenting on the original shallow gradient boosted tree ensemble methodology with labelwise feature pruning and incorporating the autoencoder embedding representations into the classifier inputs (Chapter 5).

To summarize the three predictions provided in the introduction:

- I support my original prediction, showing experimental results indicating that shallow learning boosted decision trees can outperform deep learning autoencoder models on summary classification metrics such as the *PhysioNet/CinC 2020 Challenge* scoring function and overall F-measure.
- I support my original prediction, showing that proper regularization of the input feature space and selection of relevant features for the gradient boosted decision tree classifiers are more effective than concatenating autoencoder embeddings for improving the scoring function output.
- I refute my original prediction, as naively joining deep learning autoencoder embeddings with manually engineered features for decision tree

classifiers does not significantly improve the summary classification metrics in the ECG classification task.

The gradient boosted decision tree approach, stacked autoencoders approach, and combined approaches described in this thesis are capable of predicting multiple cardiac diagnoses from unstructured, 12-lead ECG records. Specific to the *PhysioNet/CinC 2020 Challenge*, our most prominent approach selects the label-wise top 1000 most important features and autoencoder embeddings from the entire input space of features and trains an XGBoost binary classifier for each of our 27 diagnoses. This approach achieves an average test split challenge score of 0.4366, with an overall test split classification accuracy of 0.3068.

6.1 Future Work

The biggest unrealized gain in the classification of ECG records using the *PhysioNet/CinC* provided public data is for a human expert to overread and correct all erroneous labels and discard unusable samples from the available corpus. This can also be addressed by augmenting the available corpus of data with new ECG records that are sourced from known distributions and labelled by trusted cardiologists. Because the *PhysioNet/CinC 2021 Challenge* extends the current challenge and incorporates a 2-lead classification variant, a replication of this study using a subset of 2 leads is also warranted.

For use as a general ECG classifier, the success of the wide and deep transformer architecture proposed by Natarajan *et al* [38] emphasizes the importance of the transformer family of neural networks. When trained on the raw signal and a selection of manually engineered features, transformers may result in superior classifiers than approaches using convolutions, recurrent neural networks, and shallow classifiers alone.

The classification of ECG records in this thesis were limited to the 27 labels defined by the PhysioNet challenge organizers. Future work should tailor the multi-label multi-class classification task to incorporate a more broad scope of cardiac diagnoses, or explore matching diagnoses to high dimensional embedding space as an alternative to discrete binary classifiers per label.

References

[1] General Electric Healthcare - Diagnostic ECG. https://www.gehealthcare.com/products/diagnostic-ecg (Accessed November 4, 2020.

[2] Koninklijke Philips - Diagnostic ECG. https://www.usa.philips.com/healthcare/solutions/diagnostic-ecg/diagnostic-ecg (Accessed November 4, 2020).

[3] AliveCor. AliveCor KardiaMobile & KardiaMobile 6L. https://www.alivecor.com/ (Accessed November 4, 2020).

[4] Robert H. Anderson, Diane E. Spicer, Anthony M. Hlavacek, Andrew C. Cook, and Carl L. Backer. *Wilcox's Surgical Anatomy of the Heart*. Cambridge University Press, July 2013. Google-Books-ID: oAej0kaxZ8cC.

[5] Apple. Apple Watch. https://www.apple.com/ca/watch/ (Accessed November 4, 2020).

[6] Ulas Baran Baloglu, Muhammed Talo, Ozal Yildirim, Ru San Tan, and U Rajendra Acharya. Classification of myocardial infarction with multi-lead ecg signals and deep cnn. *Pattern Recognition Letters*, 122:23 – 30, 2019.

[7] Antoni Bayés de Luna, Adrian Baranchuk, Luis Alberto Escobar Robledo, Albert Massó van Roessel, and Manuel Martínez-Sellés. Diagnosis of interatrial block. *Journal of Geriatric Cardiology : JGC*, 14(3):161–165, March 2017.

[8] Yoshua Bengio. Deep learning challenges. https://cscan-infocan.ca/feature-on-homepage/watch-deep-learning-challenges-with-yoshua-bengio/ (Accessed Nov 3, 2020).

[9] J. Gordon Betts, Kelly A. Young, James A. Wise, Eddie Johnson, Brandon Poe, Dean H. Kruse, Oksana Korol, Jody E. Johnson, Mark Womble, and Peter DeSaix. *Anatomy and Physiology*. OpenStax, Huston, Texas, 2013.

[10] Yochai Birnbaum, Samuel Sclarovsky, Bruria Zlotikamien, Izhak Herz, Angela Chetrit, Liraz Olmer, and Gabriel I. Barbash. Abnormal q waves on the admission electrocardiogram of patients with first acute myocardial infarction: Prognostic implications. *Clinical Cardiology*, 20(5):477–481, 1997.

[11] Robert O Bonow, Douglas L Mann, Douglas P Zipes, and Peter Libby. *Braunwald's heart disease e-book: A textbook of cardiovascular medicine*. Elsevier Health Sciences, 2011.

[12] R. Bousseljot, D. Kreiseler, and A. Schnabel. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. *Biomedical Engineering / Biomedizinische Technik*, 40(s1):317 – 318, 01 Jan. 1995.

[13] C.J. Breen, G.P. Kelly, and W.G. Kernohan. Ecg interpretation skill acquisition: A review of learning, teaching and assessment. *Journal of Electrocardiology*, 2019.

[14] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying Density-Based Local Outliers. *SIGMOD Rec.*, 29(2):93–104, May 2000.

[15] Ed Burns. Left Axis Deviation. https://litfl.com/left-axis-deviation-lad-ecg-library/ (Accessed November 13, 2020).

[16] Mike Cadogan. ECG Lead positioning. https://litfl.com/ecg-lead-positioning/ (Accessed November 2, 2020).

[17] Patrice Carroz, Dominique Delay, and Grégoire Girod. Pseudo-pacemaker syndrome in a young woman with first-degree atrio-ventricular block. *Europace: European Pacing, Arrhythmias, and Cardiac Electrophysiology: Journal of the Working Groups on Cardiac Pacing, Arrhythmias, and Cardiac Cellular Electrophysiology of the European Society of Cardiology*, 12(4):594–596, April 2010.

[18] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[19] Tsai-Min Chen, Chih-Han Huang, Edward S.C. Shih, Yu-Feng Hu, and Ming-Jing Hwang. Detection and classification of cardiac arrhythmias by a challenge-best deep learning neural network model. *iScience*, 23(3):100886, 2020.

[20] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing*, 307:72 – 77, 2018.

[21] A. Gogna, A. Majumdar, and R. Ward. Semi-supervised stacked label consistent autoencoder for reconstruction and analysis of biomedical signals. *IEEE Transactions on Biomedical Engineering*, 64(9):2196–2205, 2017.

[22] Ary L. Goldberger, Zachary D. Goldberger, and Alexei Shvilkin. Chapter 14 - supraventricular arrhythmias, part i: Premature beats and paroxysmal supraventricular tachycardias. In Ary L. Goldberger, Zachary D. Goldberger, and Alexei Shvilkin, editors, *Goldberger's Clinical Electrocardiography (Ninth Edition)*, pages 130 – 143. Elsevier, ninth edition edition, 2018.

[23] Emanuel Goldberger. A simple, indifferent, electrocardiographic electrode of zero potential and a technique of obtaining augmented, unipolar, extremity leads. *American Heart Journal*, 23(4):483–492, April 1942.

[24] S. D. Goodfellow, A. Goodwin, R. Greer, P. C. Laussen, M. Mazwi, and D. Eytan. Classification of atrial fibrillation using multidisciplinary features and gradient boosting. In *2017 Computing in Cardiology (CinC)*, pages 1–4, September 2017.

[25] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[26] B. Hou, J. Yang, P. Wang, and R. Yan. Lstm-based auto-encoder model for ecg arrhythmias classification. *IEEE Transactions on Instrumentation and Measurement*, 69(4):1232–1240, 2020.

[27] Malcolm Kirk. Basic Principles of Pacing. In *Implantable Cardiac Pacemakers and Defibrillators*, pages 1–28. John Wiley & Sons, Ltd, 2007. Section: 1 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470750537.ch1.

[28] Kligfield Paul, Gettes Leonard S., Bailey James J., Childers Rory, Deal Barbara J., Hancock E. William, van Herpen Gerard, Kors Jan A., Macfarlane Peter, Mirvis David M., Pahlm Olle, Rautaharju Pentti, and Wagner Galen S. Recommendations for the Standardization and Interpretation of the Electrocardiogram. *Journal of the American College of Cardiology*, 49(10):1109–1127, March 2007. Publisher: American College of Cardiology Foundation.

[29] Chun Shing Kwok, Muhammad Rashid, Rhys Beynon, Diane Barker, Ashish Patwala, Adrian Morley-Davies, Duwarakan Satchithananda, James Nolan, Phyo K Myint, Iain Buchan, Yoon K Loke, and Mamas A Mamas. Prolonged pr interval, first-degree heart block and adverse cardiovascular outcomes: a systematic review and meta-analysis. *Heart*, 102(9):672–680, 2016.

[30] Anny Lam, Galen S. Wagner, and Olle Pahlm. The classical versus the Cabrera presentation system for resting electrocardiography: Impact on recognition and understanding of clinically important electrocardiographic changes. *Journal of Electrocardiology*, 48(4):476 – 482, 2015.

[31] 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(7):1368–1373, September 2018.

[32] John E. Madias. Low QRS voltage and its causes. *Journal of Electrocardiology*, 41(6):498–500, November 2008.

[33] John E. Madias. Computerized interpretation of electrocardiograms: Taking stock and implementing new knowledge. *Journal of Electrocardiology*, 51(3):413 – 415, 2018.

[34] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and Annabel S H Chen. NeuroKit2: A python toolbox for neurophysiological signal processing.

[35] Steve Meek and Francis Morris. Introduction. I—Leads, rate, rhythm, and cardiac axis. *BMJ : British Medical Journal*, 324(7334):415–418, February 2002.

[36] Arthur J. Moss. Prolonged QT-Interval Syndromes. *JAMA*, 256(21):2985–2987, 12 1986.

[37] Rachel Nall. What do EKG results look like for A-fib? https://www.medicalnewstoday.com/articles/316662 (Accessed November 2, 2020).

[38] 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 *2020 Computing in Cardiology (CinC) PhysioNet Challenge*, pages 1–4, 2020.

[39] J. Niu, Y. Tang, Z. Sun, and W. Zhang. Inter-patient ecg classification with symbolic representations and multi-perspective convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1321–1332, 2020.

[40] 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. *Physiological Measurement*, 2020. In Press.

[41] Philip J Podrid and Peter R Kowey. *Cardiac arrhythmia: mechanisms, diagnosis, and management.* Lippincott Williams & Wilkins, 2001.

[42] QuardioMD. QardioMD: Wireless ECG Monitoring with QardioCore. https://www.getqardio.com/en/qardiomd-ecg/ (Accessed November 4, 2020).

[43] Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M. M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P. S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, Wagner Meira Jr, Thomas B. Schön, and Antonio Luiz P. Ribeiro. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1):1760, April 2020. Number: 1 Publisher: Nature Publishing Group.

[44] N. Saoudi, F. Cosío, A. Waldo, S. A. Chen, Y. Iesaka, M. Lesh, S. Saksena, J. Salerno, W. Schoels, and Working Group of Arrhythmias of the European of Cardiology and the North American Society of Pacing and Electrophysiology. A classification of atrial flutter and regular atrial tachycardia according to electrophysiological mechanisms and anatomical bases; a Statement from a Joint Expert Group from The Working Group of Arrhythmias of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. *European Heart Journal*, 22(14):1162–1182, July 2001.

[45] Stephen W. Smith, Brooks Walsh, Ken Grauer, Kyuhyun Wang, Jeremy Rapin, Jia Li, William Fennell, and Pierre Taboulet. A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation. *Journal of Electrocardiology*, 52:88 – 95, 2019.

[46] Harold Smulyan. The computerized ecg: Friend and foe. *The American Journal of Medicine*, 132(2):153 – 160, 2019.

[47] Surawicz Borys, Childers Rory, Deal Barbara J., and Gettes Leonard S. AHA/ACCF/HRS Recommendations for the Standardization and Interpretation of the Electrocardiogram. *Circulation*, 119(10):e235–e240, March 2009. Publisher: American Heart Association.

[48] Vikto Tihonenko, A Khaustov, S Ivanov, A Rivin, and E Yakushenko. St petersburg incart 12-lead arrhythmia database. *PhysioBank, PhysioToolkit, and PhysioNet*, 2008.

[49] ME Valentinuni, LA Geddes, and HE Hoff. Properties of the 30" hexaxial (einthoven-goldberger) system of vectorcardiography. *Cardiovascular Research Center Bulletin*, 9(2), 1970.

[50] Rashmi Korlakai Vinayak and Ran Gilad-Bachrach. DART: Dropouts meet Multiple Additive Regression Trees. In *Artificial Intelligence and Statistics*, pages 489–497. PMLR, February 2015. ISSN: 1938-7228.

[51] Salim S. Virani, Alvaro Alonso, Emelia J. Benjamin, Marcio S. Bittencourt, Clifton W. Callaway, April P. Carson, Alanna M. Chamberlain, Alexander R. Chang, Susan Cheng, Francesca N. Delling, Luc Djousse, Mitchell S.V. Elkind, Jane F. Ferguson, Myriam Fornage, Sadiya S. Khan, Brett M. Kissela, Kristen L. Knutson, Tak W. Kwan, Daniel T. Lackland, Tené T. Lewis, Judith H. Lichtman, Chris T. Longenecker, Matthew Shane Loop, Pamela L. Lutsey, Seth S. Martin, Kunihiro Matsushita, Andrew E. Moran, Michael E. Mussolino, Amanda Marma Perak, Wayne D. Rosamond, Gregory A. Roth, Uchechukwu K.A. Sampson, Gary M. Satou, Emily B. Schroeder, Svati H. Shah, Christina M. Shay, Nicole L. Spartano, Andrew Stokes, David L. Tirschwell, Lisa B. VanWagner, and Connie W. Tsao. Heart disease and stroke statistics update: A report from the american heart association. *Circulation*, 141(9):e139–e596, 2020.

[52] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[53] 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(1):154, May 2020. Number: 1 Publisher: Nature Publishing Group.

[54] M. Wasimuddin, K. Elleithy, A. S. Abuzneid, M. Faezipour, and O. Abuzaghleh. Stages-based ecg signal analysis from traditional signal processing to machine learning approaches: A survey. *IEEE Access*, 8:177782–177803, 2020.

[55] Alexander W Wong, Weijie Sun, Sunil V Kalmady, Padma Kaul, and Abram Hindle. Multilabel 12-Lead Electrocardiogram Classification Using Gradient Boosting Tree Ensemble. In *2020 Computing in Cardiology (CinC) PhysioNet Challenge*, pages 1–4, 2020.

[56] Alexander William Wong, Amir Salimi, Abram Hindle, Sunil Vasu Kalmady, and Padma Kaul. Multilabel 12-lead electrocardiogram classification using beat to sequence autoencoders. pages 1–5, 2021. Submitted to the International Conference of Acoustics, Speech, and Signal Processing (ICASSP) 2021.

[57] Frank G. Yanowitz. Introduction to ECG Interpretation. https://ecg.utah.edu/ (Accessed November 2, 2020).

[58] 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 forClassifying 12-lead ECGs. In *2020 Computing in Cardiology (CinC) PhysioNet Challenge*, pages 1–4, 2020.

[59] Zhaowei Zhu1, Han Wang2, 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 SE-ResNet. In *2020 Computing in Cardiology (CinC) PhysioNet Challenge*, pages 1–4, 2020.

Appendix A

XGBoost Classifiers with Autoencoder Embeddings

A.1 Label-wise Feature Importances

In an extension of Section 5.2, I provide a categorical breakdown of the feature utilization from Configuration 4 in Figure A.1, as well as Configuration 6 displayed in Figure A.2.

Feature utilization is a value ranging between $[0, 1]$ that represents how many features from the given category is used in the pruned classifier. For example, consider the meta variable *Age*. Within an ECG record, there is only 1 tabular feature representing patient age. A utilization of 1 means that age is always determined to be an important feature to be kept. Consider another category *Heart Rate*. Referring to Section 3.2.2, we know that any given lead may contribute at most 53 heart rate variability features. A utilization of 0.4 means that only 21 of the available 53 features generated from the lead were deemed important and kept in the pruned classifier input space.

A.2 Label-wise F_1 Scores

A plot showcasing all of the labelwise F1 scores outputted by the 10 classifier configurations of Section 5.1 can be found in Figure A.3.

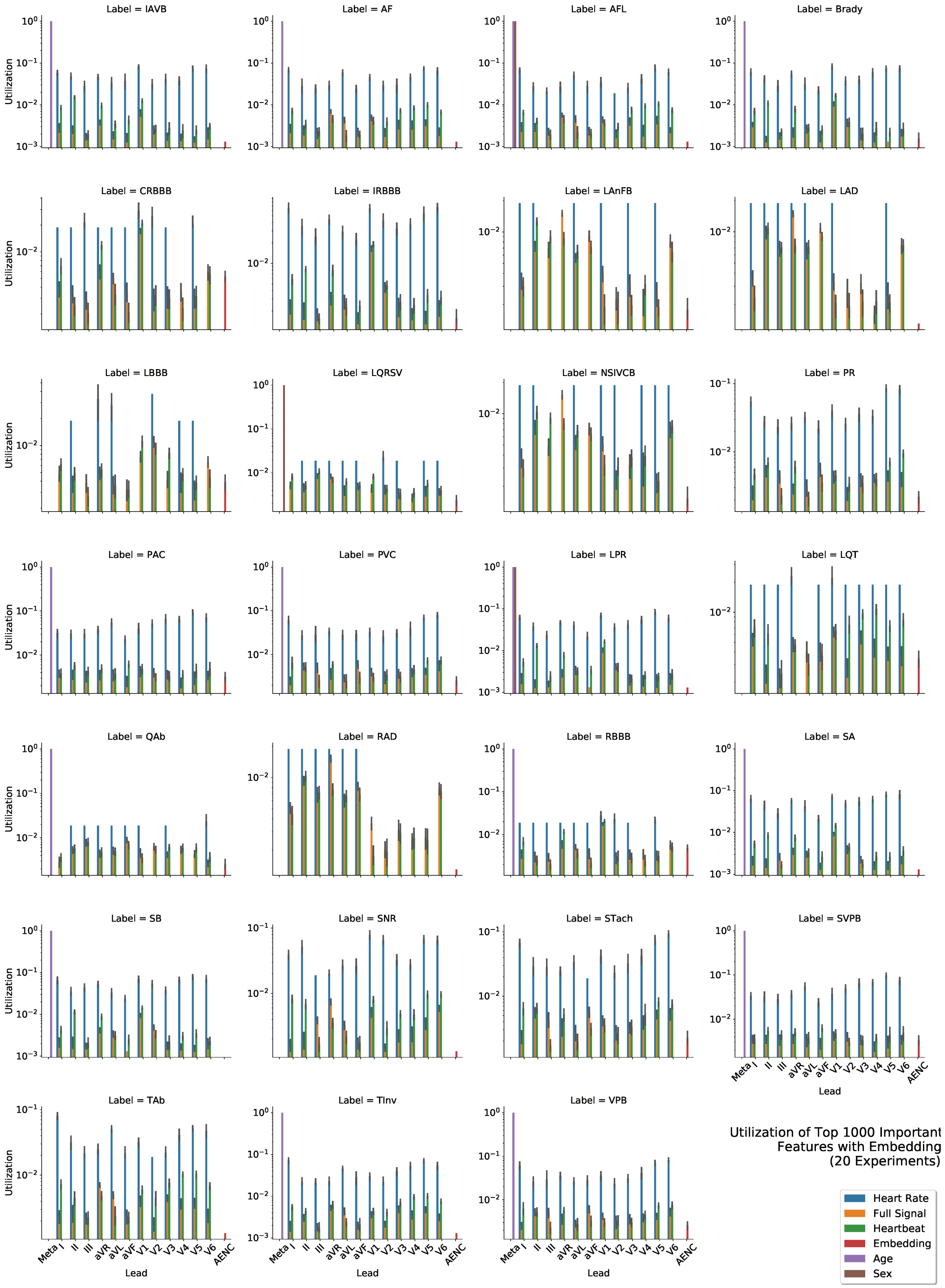


Figure A.1: Feature importances of Configuration 4: "Top 1000 Features with Embeddings", utilization over 20 experiments. The 27 diagnosed labels are displayed separately, showcasing feature derived lead (alternatively Meta or Autoencoder), and category.

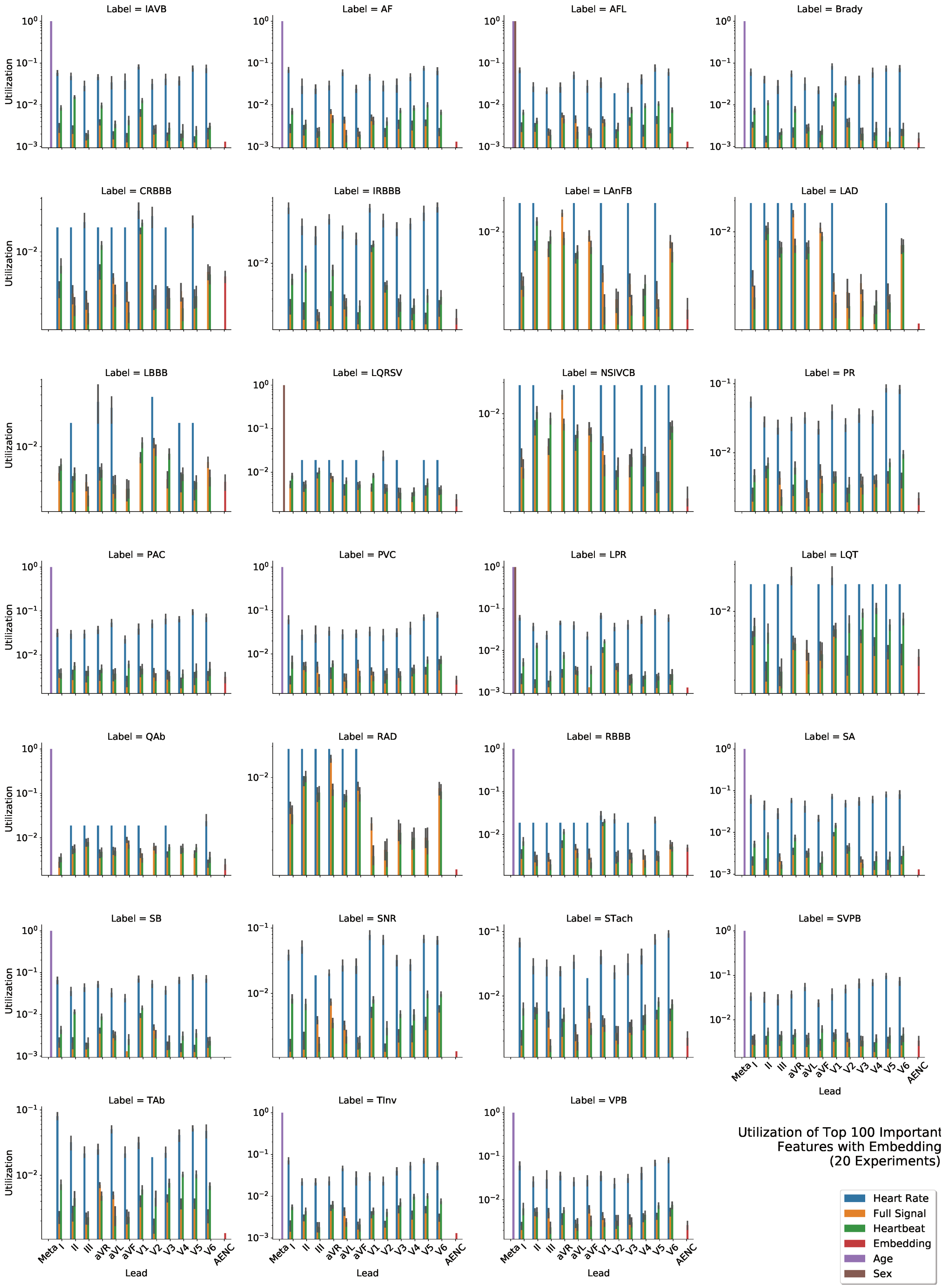


Figure A.2: Feature importances of Configuration 6: "Top 100 Features with Embeddings", utilization over 20 experiments. The 27 labels are displayed separately, showcasing feature derived lead (alternatively Meta or Autoencoder), and category.

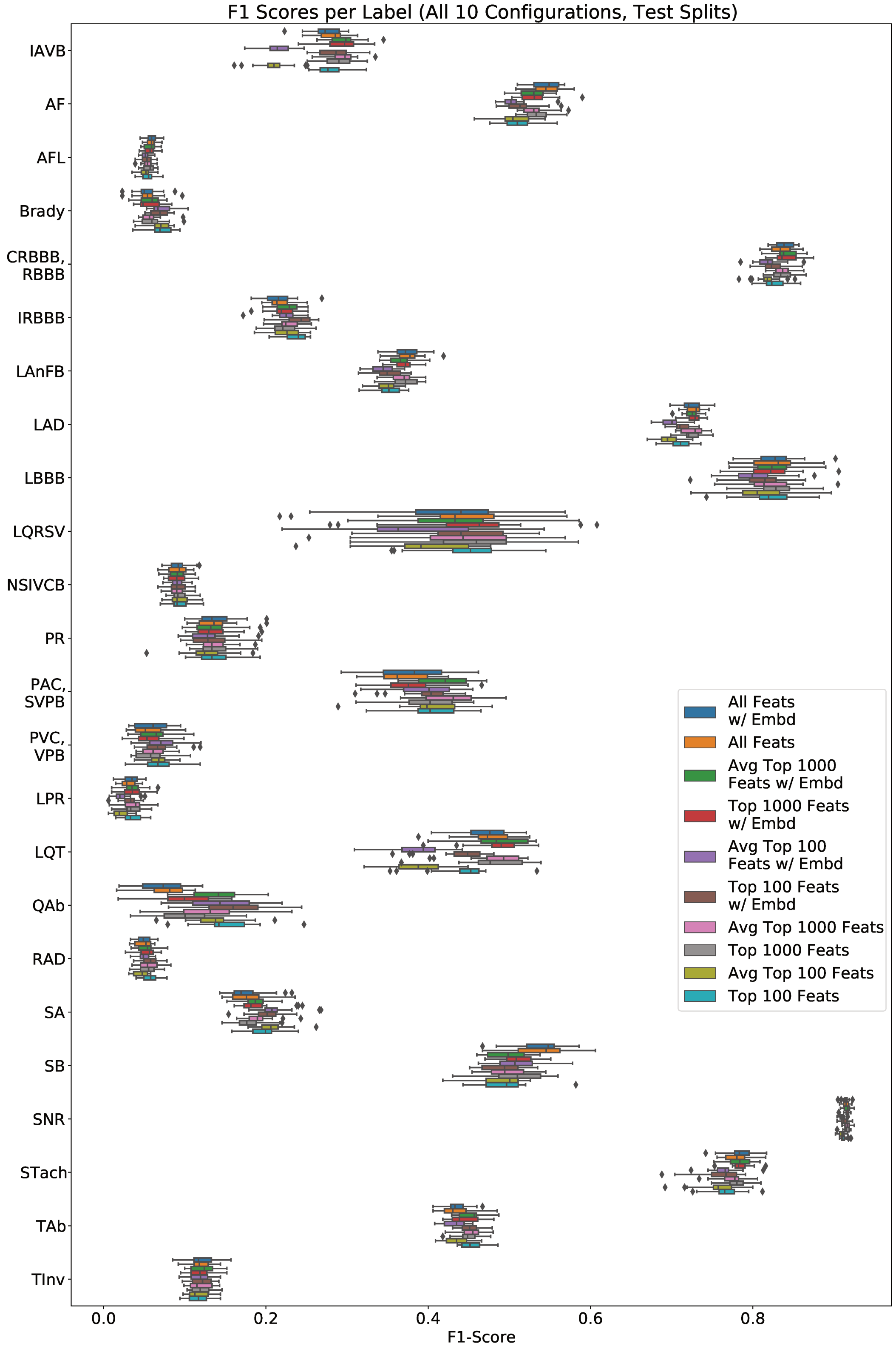


Figure A.3: Test split label-wise F_1 scores for each of the 10 configurations over the 20 independent repeated random subsampling iterations.