Learning Models for Diagnosis and Prognosis from Electrocardiogram Data

by

Weijie Sun

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

Master of Science

Department of Computing Science
University of Alberta

Abstract

The electrocardiogram (ECG) records the electrical activity of a patient's heart movement. It is one of the standard routine healthcare tests as it is non-invasive and easy to apply. In this thesis, we analyze 2 million ECGs and over 260,000 patients' health records from the Alberta Health Service, and propose frameworks for learning diagnostic and prognostic models based on supervised learning methods, including ones for survival prediction. First, we learned many models that each use a patient's ECG to determine if s/he has a specific disease, corresponding to an ICD-10 diagnosis code. Our results show that these diagnosis models can accurately predict numerous health conditions, beyond cardiovascular conditions. Second, we develop ECG diagnosis models for COVID-19 and then use transfer learning to produce models with superior performance. Finally, motivated by the evidence from earlier tasks, we develop binary classification ECG models for predicting all-cause (fixed time) mortality for hospitalized (resp., emergency) patients, and also survival models that produce meaningful survival predictions for each patient. We demonstrate state-of-the-art performance for predicting the time-until-death by using machine learning techniques that first re-express each ECG in latent representations.

Preface

My research work carried out during this thesis period has contributed either directly or partially towards several publications. These are directly related publications:

Chapter 4 is based on "ECG for high-throughput screening of multiple diseases: Proof-of-concept using multi-diagnosis deep learning from population-based datasets", published in the 2021 "Medical Imaging Meets NeurIPS" workshop.

Chapter 4.3 is based on "Improving ECG-based COVID-19 diagnosis and mortality predictions using pre-pandemic medical records at population-scale", published in the 2022 "Learning from Time Series for Health" NeurIPS workshop.

Chapter 5 is based on the manuscript: "Machine learning models using electrocardiograms and laboratory data to predict short and long-term mortality outcomes in a population-level cohort of patients", revised version is being submitted to npj Digital Medicine after peer-review.

and partially related publications:

Wong, A.W., Sun, W., Kalmady, S.V., Kaul, P. and Hindle, A., 2020, September. Multilabel 12-lead electrocardiogram classification using gradient boosting tree ensemble. In 2020 Computing in Cardiology (pp. 1-4). IEEE.

Kumar, N., Qi, S.A., Kuan, L.H., Sun, W., Zhang, J. and Greiner, R., 2022. Learning accurate personalized survival models for predicting hospital discharge and mortality of COVID-19 patients. Scientific reports, 12(1), pp.1-11.

Kalmady, S., Sun, W., Ezekowitz, J., Fine, N., Howlett, J., Savu, A., Greiner, R. and Kaul, P., 2021, May. Improving the calibration of long term predictions of heart failure rehospitalizations using medical concept embedding. In Survival Prediction-

Algorithms, Challenges and Applications (pp. 70-82). PMLR.

Acknowledgments

I am grateful to my supervisors, Professor Dr. Russell Greiner and Professor Dr. Padma Kaul, for giving me many opportunities and suggestions. Thank you for my mentor Dr. Sunil V. Kalmady. Without his patient instruction and professional guidance, I could not finish my thesis and research.

Thanks to my parents (Qinhong Sun and Yan Xia) and friends for their support and understanding. During the pandemic, they support financially and spiritually.

Table of Contents

List of Tables

List of Figures

Abbreviations & Acronyms

AHS: Alberta Health Services.

AP: Average Precision.

AUPRC: Area Under the Precision-recall Curve.

AUROC: Area Under the Receiver Operating characteristic Curve.

ECG: ElectroCardioGraphy.

ED: Emergency Department.

EHR: Electronic Health Record.

IBS: Integral Brier Score.

ICD-10: the 10th revision of the International Statistical Classification of Diseases and Related Health Problems.

IPCW: Inverse Probability of Censoring Weight.

ISD: Individual Survival Distribution.

MTLR: Multi-Task Logistic Regression.

N-MTLR: Neural Multi-Task Logistic Regression:.

XGB: gradient boosted tree ensembles.

Chapter 1

Introduction

Cardiovascular disease is the leading cause of mortality in human beings. Cardiologists often check general cardiac conditions and measure heart function with ElectroCardioGraphy (ECG) tests, as these tests are non-invasive and can assess general cardiac conditions by detecting heart movement. In industrialized countries [1], ECG tests are readily available and performed on most patients in outpatient clinics or inpatient hospitals. These ECG test results could help medical doctors monitor a patient's heart condition, which can identify possible heart abnormalities and be used for the early detection of cardiovascular diseases. For example, atrial fibrillation, an irregular and rapid heart rhythm that can sometimes exceed 400 beats per minute can lead to blood clots in the heart. Clinical experts could recognize atrial fibrillation in ECG signals; see Figure 1.1. Moreover, early detection of atrial fibrillation may lead to proper treatment, which may decrease the risk of stroke, heart failure, and other heart-related complications[2].

Despite excellent availability, there are two reasons patients do not fully receive the benefits of ECG interpretations. Firstly, ECG patterns are complex, making them challenging and time-consuming for clinical physicians. Secondly, as people in rural areas may lack the motivation to take ECG tests, cardiologists in those areas have fewer opportunities to interpret ECGs because of the complexity of ECG patterns. Note that the infrastructure in rural health offices and clinics is inferior to major

metropolitan centers. In developing nations, people in rural and poor communities cannot afford ECG tests and lack the motivation to find health care because of poorer socioeconomic status. The average doctor in rural and poor communities does not have an opportunity to develop their skills to interpret ECG reads effectively.

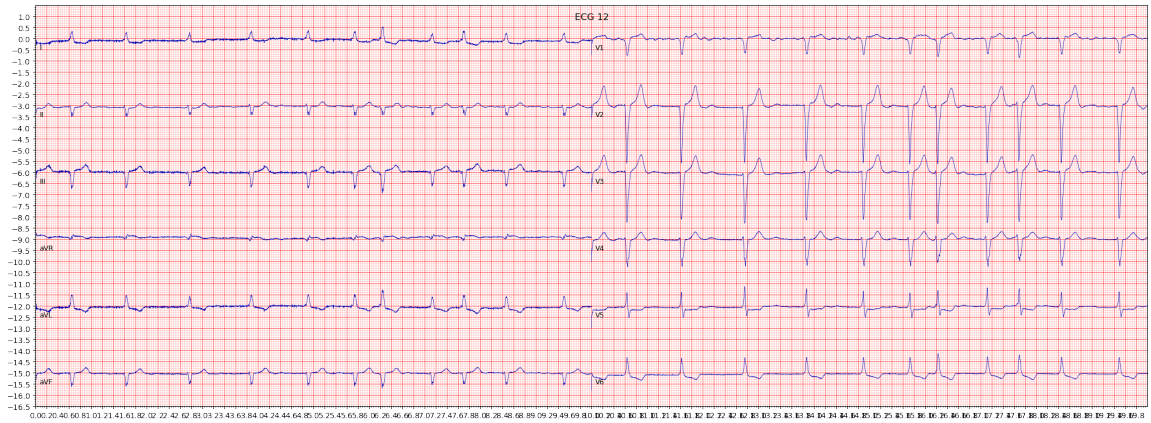

Figure 1.1: An example of a 12-lead ECG of a patient with atrial fibrillation shows low amplitude and nearly hard-to-detect P waves. (Terms defined in Section 2.1 below. Note each row is two different leads.)

ECG interpretation remains a difficult task, even for cardiologists. Over the past decade, Deep Learning (DL) has achieved remarkable success in various medical and biological fields. Clinical experts and computing scientists cooperate to produce algorithms that reach cardiologist-level performance in diagnosing cardiovascular diseases (e.g., arrhythmia) [3]. ECG analysis models can assist clinicians at the point-of-care decision-making by making an accurate prediction and facilitating a learned healthcare system. It is helpful not only to have the model detect current diseases but also, to predict the risk of future cardiovascular problems.

The Alberta Health Services (AHS) has provided us with a large population-level dataset containing 2,015,808 ECGs and records of 3,336,091 emergency visits, 1,071,576 hospitalizations, and over 260,065 patients. AHS uses ICD-10 (the 10th revision of the International Statistical Classification of Diseases and Related Health Problems) [4] codes as the diagnosis codes in emergency and hospitalization records. By indexing the time and the patient's ID, we can identify which ECG scan corre-

sponds to which ICD-10 code. In this thesis, we use ResNet and gradient-boosted tree ensembles (XGB) to produce models that, given a patient's ECG, can identify the ICD-10 diagnosis codes for that patient over a wide range of diseases (both cardiovascular and non-cardiovascular) . In addition, we use the pre-trained ResNet DL model to improve the performance in the diagnosis of COVID-19. Moreover, we implement prognosis methods that, for each patient, predict: (1) binary classifications of the mortality risk model in short-term (30 days) and long-term mortality (1 year and 5 years), and (2) individual survival model, which produces the survival probability for all future time points.

1.1 Contributions

My contributions to this thesis include:

1. We design the episode generation algorithm to combine the data from hospitalization visits, and emergency department encounters with the same clinical episode in Electronic Health Record (EHR) structure data. This allows us to extract episode duration, which can then be used to match other datasets (e.g., ECG data) and combined diagnosis codes from different hospitalization visits and emergency department encounters, all of which are in the same generated episode. See Section 3.1.

2. This study shows that it is possible to learn effective models that use ECG signals to predict a wide range of 1,414 diseases, including cardiovascular diseases and many non-cardiovascular conditions such as mental, neurological, metabolic, and infectious diseases. See Section 4.2.

3. To address the challenge of having relatively few ECG instances of patients with some diseases, we show that the transfer learning strategy with ResNet 1414Dx model could improve the performance in diagnosing such diseases with few instances – e.g., COVID-19 data during the 2020-2021 pandemic duration. See Section 4.3.

4. It is essential to provide the risk scoring system for patients and doctors to help them identify the appropriate individual therapy strategies. Therefore, we provide the ECG prognosis models that can effectively predict the risk score, respectively, for 1-year, short-term (30-days), and long-term (5-years) mortality. We also evaluate the models in random ECG per patient, which we show have good performance in AUROC [5] (also known as C-index).

5. We explore an Individual Survival Distribution (ISD) model, which provides the survival probability for each individual patient for all future time points.

We provide a way to learn this ISD model from the information from a survival dataset that includes patient information and ECG data. This learned model will then predict survival probability at all future time points. Then, we utilized ECG ISD models to produce individual survival curves.

1.2 Thesis Organization

This dissertation contains the following chapters. Chapter 2 provides the relevant background, which describes ECG characteristics and interpretations, previous literature on ECG-based diagnosis and prognosis tasks. Chapter 3 describes characteristics of ECG measurements and ECG leads, pre-processing methods for hospitalization, Emergency Department (ED) visits and in-hospital ECG data, training models (ResNet, XGBoost, and MTLR) and evaluation metrics that we will use in Chapters 4 and 5. Chapter 4 provides a diagnosis model (ResNet 1414Dx) over an ICD-wide range of diseases, based on the patient's ECG. When we tried to learn a model that could predict whether a patient has COVID-19 based on ECG in 2020-2021, there are not enough instances to train a model that performs effectively. This motivated us to use the pre-trained ResNet 1414Dx model to learn a COVID-19 prediction model; this resulted in improved performance. In Chapter 5, we develop machine-learned models that use a patient's ECG data to predict the short- and long- term mortality for patients presenting to the hospital for any reason. Additionally, we learn an MTLR model that can produce an individual survival probability distribution for a patient, which provides the chance that the individual could live until time of event (death). Finally, Chapter 6 provides the summary of the thesis and future plans.

Chapter 2

Literature Review

This chapter reviews the ECG analysis in machine learning tasks, ECG-related diagnosis, and prognosis work. First, Section 2.1 discusses the three different ECG formats: camera-captured ECG images, hand-crafted ECG features, and raw ECG digitized waveform. Then, Section 2.2 summarizes existing learning algorithms that produce machine learned models, which can diagnose cardiovascular diseases from ECG data Most of these experiments are limited to classifying ECG abnormalities and cardiovascular diseases. However, in the clinical background literature, we discovered strong relations between ECG characteristics and numerous diseases seemingly unrelated to cardiovascular conditions. Finally, Section 2.3 compares and reviews the ECG-based prognosis literature. We will show our approaches in binary mortality predictions and ISD predictions.

2.1 ECG analysis

In a fundamental sense, ECG records electrical activity between electrodes on a patient's body to determine cardiac activity [6]. Commonly, medical devices generate multiple lead ECG waveforms by comparing pairs of electrodes placed at different points in the body. See in Figure 2.4. For example, lead I can be recorded from the position of the electrical potential difference from the right-arm electrode to the left-arm electrode.

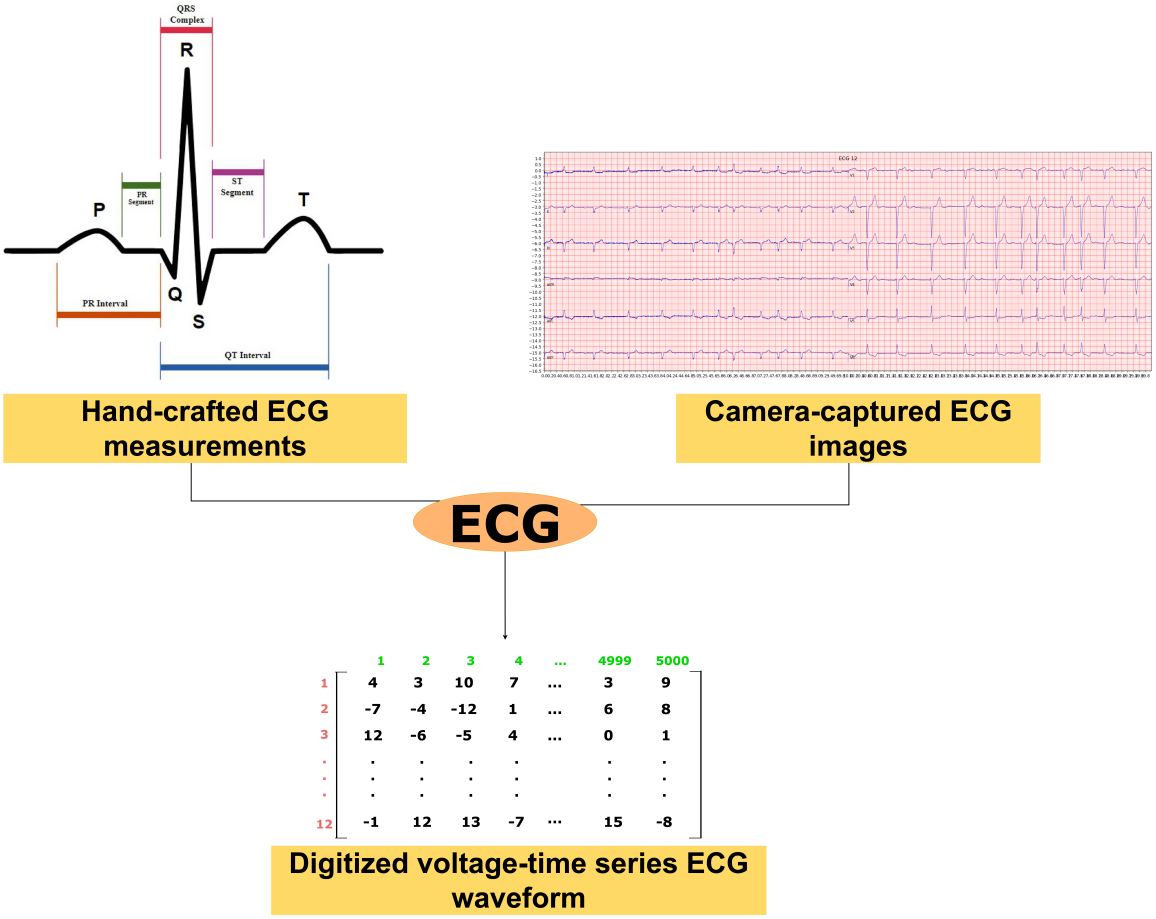

Figure 2.1: Image shows three main formats of ECG data: Hand-crafted ECG measurements, Camera-captured ECG images, and Digitized voltage-time series ECG waveform.

In recent years, machine learning models have reached near-cardiologist performance levels in analyzing and classifying ECG data [3]. ECG analysis methods rely

on the three main formats of ECG data to generate features in Figure 2.1. (1) Camera-captured ECG images: These images are typically used by clinical experts who have received professional training in ECG data interpretation. In addition, these interpretations are generally used as ground-truth labels for diagnostic prediction tasks. (2) Hand-crafted ECG measurements: Traditional ECG plots [7] are presented with the P, Q, R, S, and T waves in Figure 2.3, which summarize heart functioning with the wave amplitude (relevant for diagnosing heart rhythm abnormality) and wave duration (relevant for heart frequency abnormality).

In each heartbeat, the electrical impulse will go through the sinoatrial node, spread across the atrium area, pass the atrioventricular node, and get into the ventricular septum of the heart [8]; see Figure 2.2. ECG signals consist of several waves in each electrical impulse in Figure 2.3. The following list summarizes the hand-crafted feature description [9].

1. P wave: the first little bump in Figure 2.3. When the heart is in a normal state, this wave is ≤ 0.3 millivolts (mV). and has a width of ≤ 0.12 seconds (s).

2. QRS-Complex: the big spike represents ventricular depolarization in Figure 2.3 after the P-wave area. In the standard scenario, this QRS wave has a width of 0.06–0.12 s, and the spike height depends on the lead.

3. T-wave: after this QRS-spike, there is a "bump" shortly after the complex representing repolarization of the ventricles, called the T-wave. In a normal heart, the T wave has a positive value in all leads.

4. PR-interval: it is measured from the front of the P-wave to the beginning of QRS-Complex. This wave is 0.12–0.20 s of width in a normal heart scenario.

5. QT-interval: it connects the QRS complex and the T wave, and represents the duration where the ventricles are depolarized.

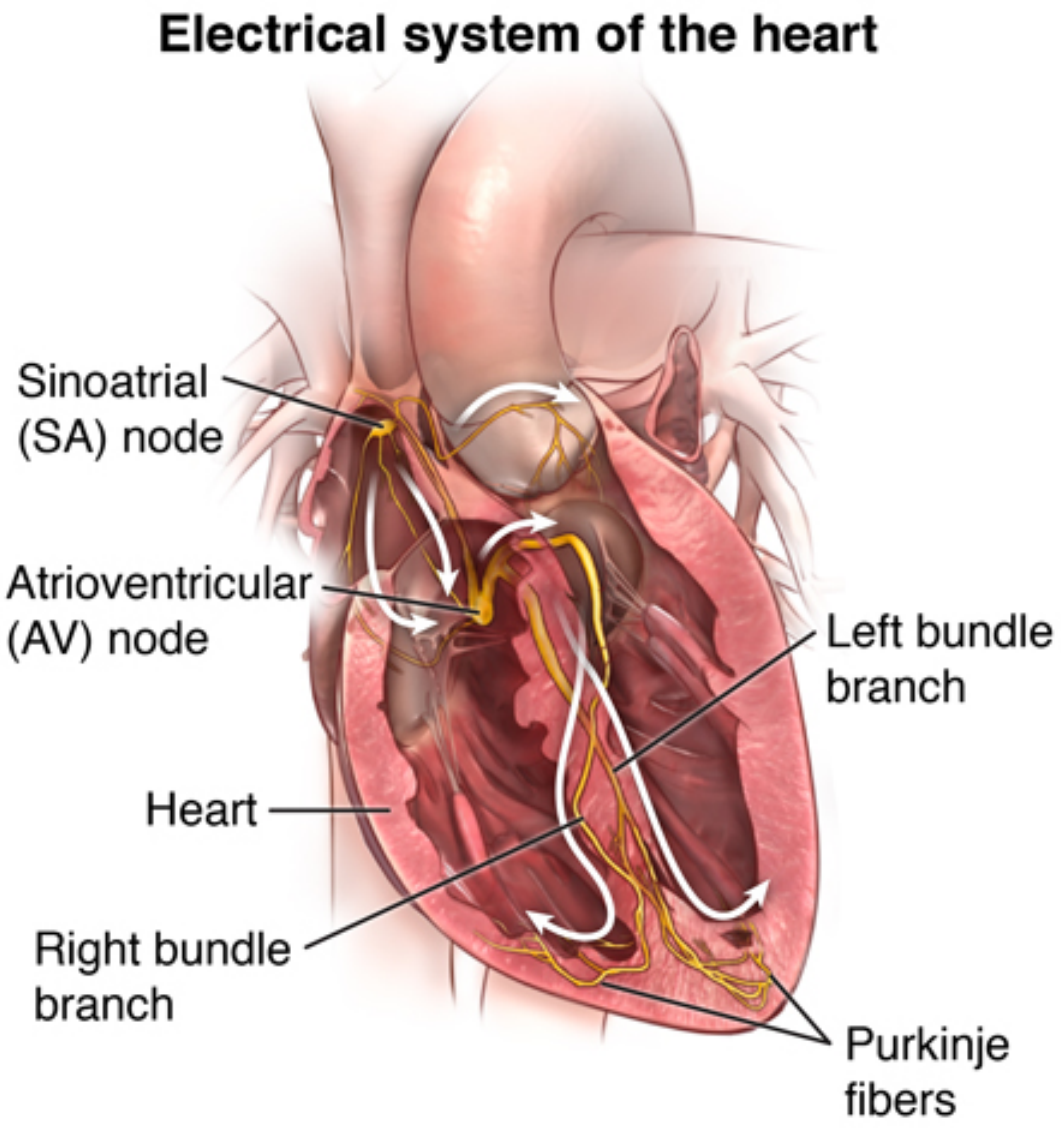

Figure 2.2: Image shows heart's pumping action (arrows) in various chambers of the heart.
Image source from https://www.hopkinsmedicine.org/health/conditions-and-diseases/anatomy-and-function-of-the-hearts-electrical-system

However, extracting hand-crafted ECG measurements manually from the ECG plots can be very time-consuming and resource intensive. Therefore, knowledge-based algorithms can extract traditionally 'hand-crafted' ECG measurements [10, 11].

(3) Digitized voltage-time series ECG waveform: This literature shows machine

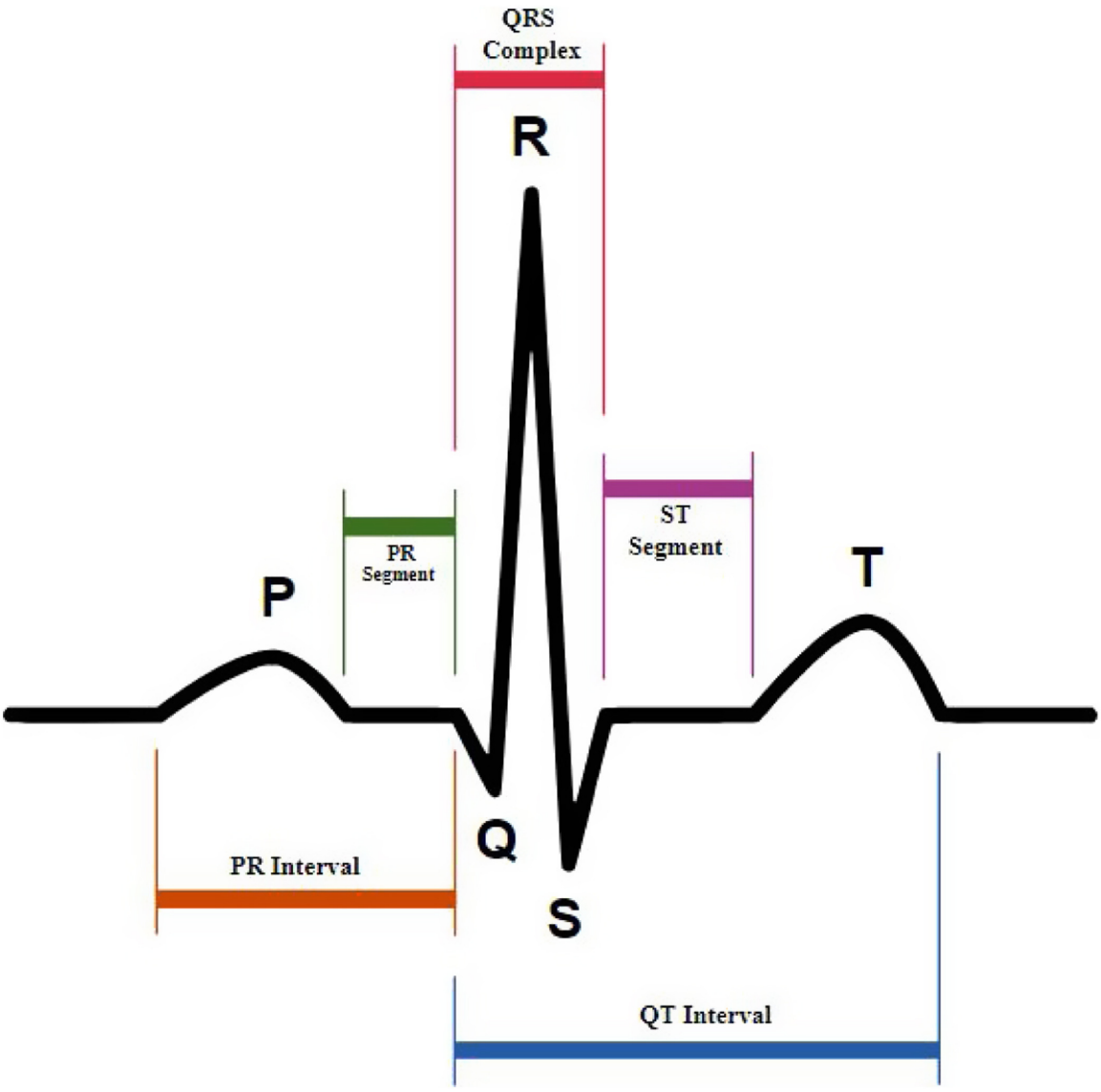

Figure 2.3: Above image shows the normal sinus rhythm ECG trace that include the PQRST peaks, PR Interval, PR Segment, QRS Complex, ST Segment, and QT Interval.

Image source from
https://commons.wikimedia.org/wiki/File:SinusRhythmLabels.svg contributed by
Agateller (Anthony Atkielski)

learned models that perform well even for high-dimensional data. In this case, machine learning experts can train the deep learning model directly from the digitized ECG waveform [12–15].

In the digital 12-lead ECG traces test, clinical doctors place ten electrodes on the skin's surface [16]. Figure 2.4 shows the 10 Ecg Placement on the skin of a human.

There are six precordial electrodes to produce leads V1 to V6. Moreover, the other four limb electrodes are right arm (RA), right leg (RL), left arm (LA), and left leg (LL). Based on values from the four limb electrodes, it can compute the values for the other six leads: I, II, III, aVR, aVL, and aVF. In Table 2.1, we show how 12 leads record from ten electrodes.

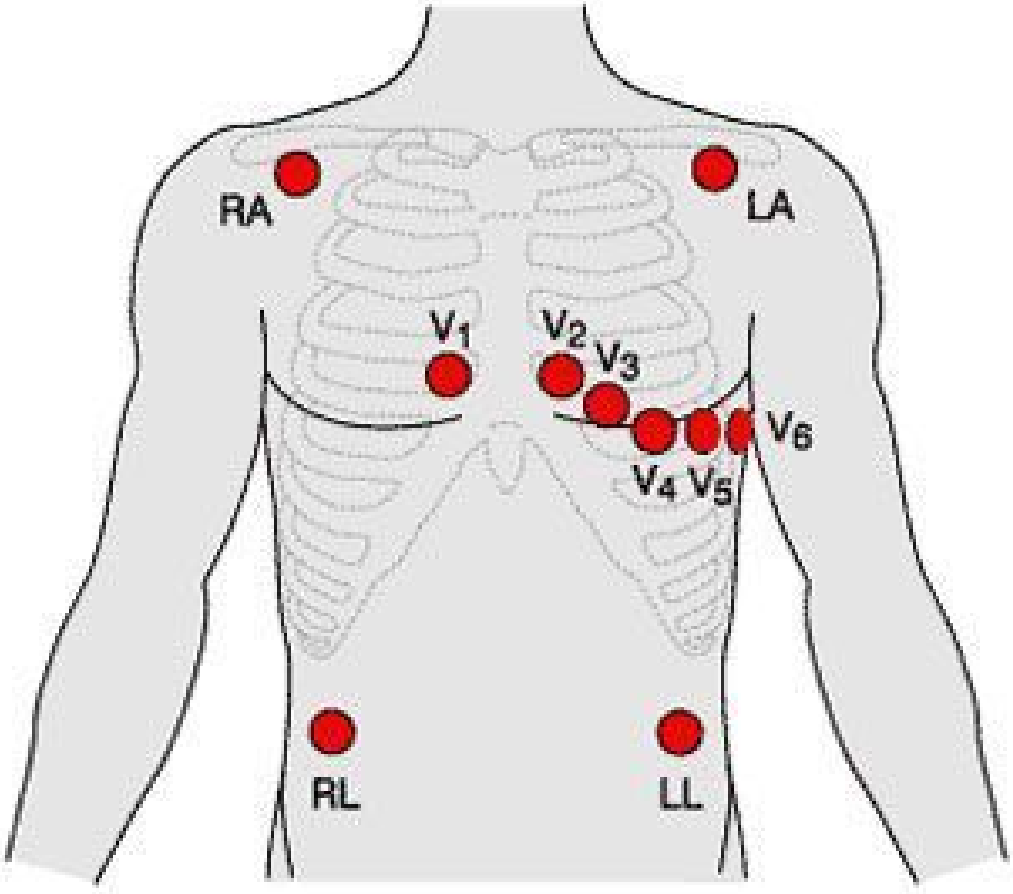

Figure 2.4: Image shows 12 Lead Ecg Placements.
Image source from
https://anatomynote.com/medical-appliance/12-lead-ecg-placement/

The heart's electrical activity has approximately three orthogonal directions: right/left, superior/inferior, and anterior/posterior. See Figure 2.5. The leads I, II, III, aVR, aVL, and aVF describe its motion in the heart at a vertical plane. On the other hand,

Electrode	Placement
V1	4th Intercostal space to the right of the sternum
V2	4th Intercostal space to the left of the sternum
V3	Midway between V2 and V4
V4	5th Intercostal space at the midclavicular line
V5	Anterior axillary line at the same level as V4
V6	Midaxillary line at the same level as V4 and V5
RL	Anywhere above the ankle and below the torso
RA	Anywhere between the shoulder and the elbow
LL	Anywhere above the ankle and below the torso
LA	Anywhere between the shoulder and the elbow

Table 2.1: 10 ECG electrodes' positions and names

the leads V1 - V6, check the heart at a horizontal plane. Table 2.2 describes the 12 leads.

Through the AHS, we were granted access to many ECG images, in multiple ECG data formats, see Figure 2.1. As a result, we trained different models using hand-crafted ECG measurements or digitized ECG waveforms in ECG analysis and classification tasks.

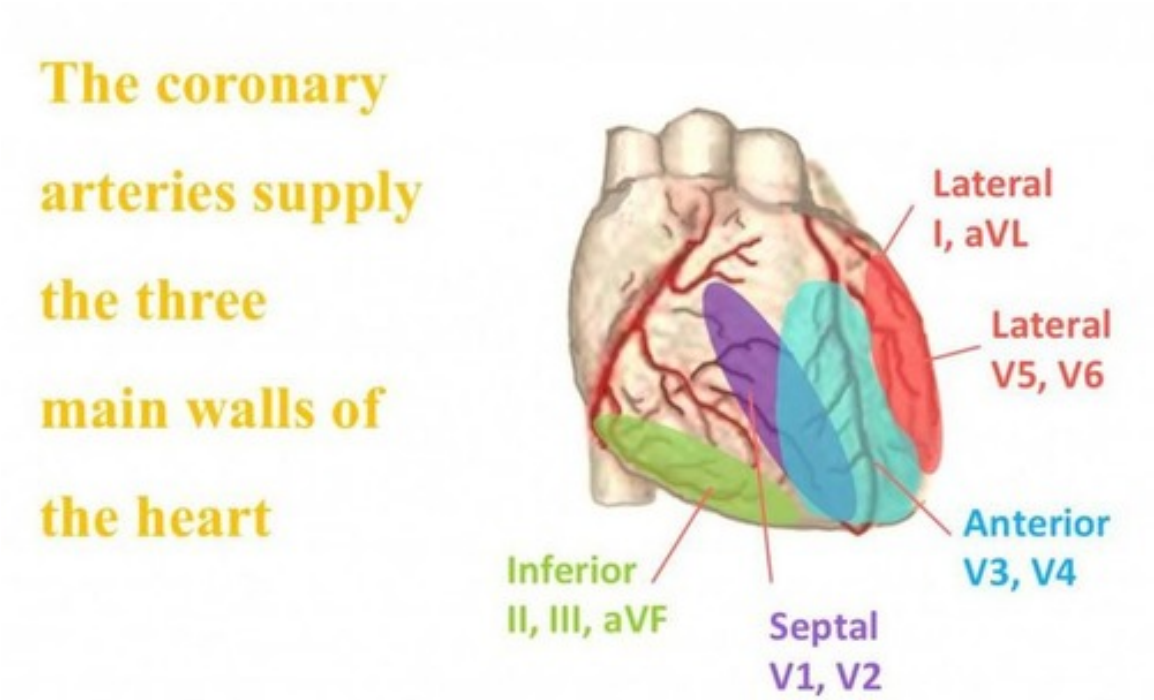

Figure 2.5: Image shows three orthogonal directions: right/left, superior/inferior, and anterior/posterior.

Image source from https://quizlet.com/116751872/12-lead-ekg-1-howell-q6-flash-cards/ contributed by eugene muro

Lead	Negative Electrode	Positive Electrode	heart plane
Lead I	RA	LA	Lateral
Lead II	RA	LL	Inferior
Lead III	LA	LL	Inferior
aVR	LA + LL	RA	None
aVL	RA + LL	LA	Lateral
aVF	RA + LA	LL	Inferior
V1			Septal
V2			Septal
V3			Anterior
V4			Anterior
V5			Lateral
V6			Lateral

Table 2.2: 12 ECG leads description

2.2 ECG based diagnosis

ECG captures the propagation of the electrical signal in the heart and is routinely used to diagnose cardiovascular diseases [17]. However, the complexity of ECG signals is still challenging and time-consuming to interpret, even for trained ECG experts. By developing machine learning models, computing scientists collaborate with medical experts to extract meaningful information from ECGs. The learned models reach near cardiologist-level performance in heart condition classification tasks [3].

Influenced by traditional ECG applications, most ECG diagnoses have been limited to typical ECG abnormalities. For example, in classification experiments, some learned models classified ECG as either normal or abnormal (summarizing multiple cardiac problems using single class) using a CBRNN model, which consists of two sub-networks: convolutional neural network (CNN) and bi-directional recurrent neural network (BRNN) [18]. Other studies have predicted more specific diseases but focus mainly on cardiovascular-related disorders. Avanzato et al. [19] used a convolutional neural network model to address the multi-class: normal, atrial premature beat, or premature ventricular contraction. However, patients could experience more than one of these conditions simultaneously. In other words, the observed ECG could have features arising from multiple conditions. Ribeiro et al. [14] designed a deep neural network multi-label classification model to recognize six types of abnormalities, meaning a single ECG waveform may reveal multiple ECG abnormalities. In general, most machine learning diagnosis studies focus only on typical ECG abnormalities, such as arrhythmias [20], valvulopathy, cardiomyopathy, and ischaemia [21]. We extend the focus to a wide range to both cardiovascular and non-cardiovascular diseases.

Several clinical studies have shown strong associations of ECG abnormalities with numerous diseases beyond cardiovascular conditions, including mental disorders (depression [22], bipolar disorder [23]); infectious conditions (HIV [24], sepsis [25]);

metabolic diseases (diabetes type 2 [26], amyloidosis [27]); drug use (psychotropics [28], cannabis [29]); neurological disorders (alzheimer disease [30], cerebral palsy [31]); respiratory diseases (pneumoconiosis [32], chronic obstructive pulmonary disease [33]); digestive system diseases (liver cirrhosis [34], alcoholic liver disease [35]); miscellaneous conditions (chronic kidney disease [36], preterm labour [37], systemic lupus erythematosus [38]), etc. However, despite well-established clinical associations of ECG changes with multiple diseases, few studies have used the information contained within ECGs to predict non-cardiovascular conditions. A major challenge is the lack of available large training datasets of digitized ECGs labeled with the appropriate diagnostic information related to a wide range of disease types. In this context, standardized administrative health data, routinely generated at each encounter, provide a wonderful opportunity to explore the full spectrum of patient diagnoses. These data include the professional clinical records of diagnosis and any comorbidities the patients may have or develop during the visit.

2.3 ECG based prognosis

Cardiovascular diseases are the main cause of death in Canada and globally. Just in 2019, cardiovascular diseases took around 17.9 million lives [39]. This is why clinical researchers seek models that can accurately estimate the probability of mortality and assist clinicians in prioritizing their medical resources. In a recent publication, Kwon et al. [40] implemented a deep neural network that learned a model that used ECG data to predict patients' 12- and 36- month mortality following acute heart failure (DAHF). Van de Leur et al. learned models that performed well in in-hospital all-cause mortality of COVID-19 patients with pre-trained deep neural network (DNN) using age, sex, and the raw ECG waveforms [41]. However, no one has yet explored the feasibility and value of linking ECG data to longitudinal population-level administrative health data to assist clinicians at point-of-care decision-making to complete the cycle of quality and facilitate a learning healthcare system [42, 43].

Raghunath et al. [15] predicted 1-year all-cause mortality from ECG voltage–time traces with custom-designed DL architecture that utilized convolutional neural networks using five branches to accommodate varying durations of ECG acquisition across the groups of leads at the population level (nearly 2.3 million ECGs). We compare the results in Chapter 3. We develop additional binary mortality classification models for shorter-term (30-days) and longer-term (5-years) mortality binary outcomes learned from population-level ECG data. Moreover, we developed Multi Task Logistic Regression (MTLR) [44], and Neural Multi-Task Logistic Regression (N-MTLR) [45] models to produce personalized survival curves. The personalized survival curve is a single curve that shows the probability of a patient's chance to survive at all future time points. See in Section 3.2.

Chapter 3

Method

Section 3.1 summarizes the data from the patient's administrative records and the ECG data from the Philips IntelliSpace ECG system. Also, that section explains the preprocessing methods, which include the episode generation algorithm and how to identify poor quality ECGs, which will be removed. Next, Section 3.2 proposes the architecture and specifies the hyperparameters of the ResNet DL model for the 12-lead digitized ECG format, the XGBoost model for the Philips ECG measurement format, and the ISD algorithms: MTLR and N-MTLR. Finally, Section 3.3 discusses our evaluation method in the threshold-based binary classification and individual survival distribution models.

3.1 Data

The province of Alberta, Canada, has a single-payer (Ministry of Health: Alberta Health) and single-provider (Alberta Health Services) for its healthcare system. As a result, the 4.4 million residents of the province have universal access to the hospital, ambulatory, laboratory, and physician services.

For this study, we linked ECG data for each patient with the following administrative health databases using that patient's unique health number: (1) the Discharge Abstract Database (DAD) containing data on hospitalizations including admission date, discharge date, most responsible diagnosis, up to 24 other diagnoses, and discharge status (one of transfer, discharge home, died); (2) the National Ambulatory Care Reporting System (NACRS) database of all hospital-based outpatient clinic (including the Emergency Department, aka ED) visits. The NACRS data include the date of admission, primary diagnosis, up to 9 other diagnoses, and the discharge status; the Alberta Health Care Insurance Plan Registry (AHCIP), which provides demographic information (patient's sex, year of birth) and date of death (if relevant); and vital status death registry. In case of conflicting mortality status or dates (1.1% of patients), the vital status registry was prioritized over the DAD, NACRS, and AHCIP records. The study cohort included patients presenting to 84 EDs or hospitals between February 2007 and April 2020 in the northern Alberta Health Zone and contained 2,015,808 ECGs, 3,336,091 ED visits, and 1,071,576 hospitalizations associated with 260,065 patients. Concurrent healthcare encounters for a patient (ED visits and/or hospitalizations) that occurred within a short period were considered transfers (for example, from ED to hospital admission or from community hospital to tertiary hospital) and grouped into episodes. In general, an "episode" is a sequence of one or more ED or INP encounters that are considered a single "hospitalization", based on the following procedure.

We consider consecutive healthcare encounters for a patient who has several ED

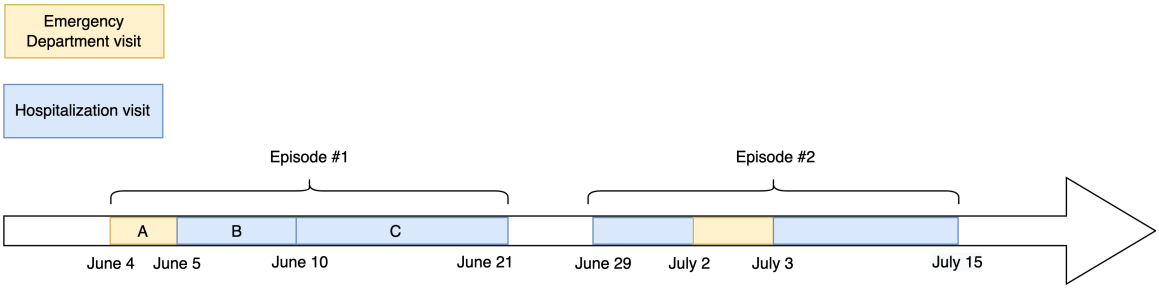

Figure 3.1: Episode generation example from Mr. ABC's health records timeline

visits and/or INP hospitalizations that occurred within a short period of time to be transferred, and so group them into episodes. For example, Figure 3.1 shows that patient Mr. ABC had an ED visit in facility A on June 4. He was diagnosed with the following ICD-10 codes I124, I20 in facility A. Then, he transferred interfacility to inpatient hospitalizations in hospital B, starting from June 5 and staying until June 10 with following ICD-10 diagnosis codes I124, I20, Y30, I50. Then, as he recovered from his diseases, he moved to community hospital C which is close to his family, from June 10 to June 21. He still has some chronic diseases with ICD-10 codes I124, I50, Z20. On June 21, he was discharged from community hospital C. According to the above scenarios, we consider the three ED or hospital visits to be in the same episode (Episode # 1). Thus, his diagnosis codes are shared between episode start date-time (June 4) and episode end date-time (June 21). Then, 8 days after June 21, he re-visited the hospital for another disease; we consider this to be the start of a new episode (Episode # 2). Based on the first 3 visits, the ICD-10 codes for Episode # 1 are I124, I20, Y30, I50, Z20.

We use flowchart in Figure 3.2 used to define an "episode". Each patient can have a series of encounters, including inpatient and outpatient (including ED) visits. We say two consecutive encounters (event 1 and event 2) belong to the same healthcare episode whenever:

1. The patient visits the ED. Then s/he is transferred to a hospital facility within 48 hours, and ED's discharge code is not "home".

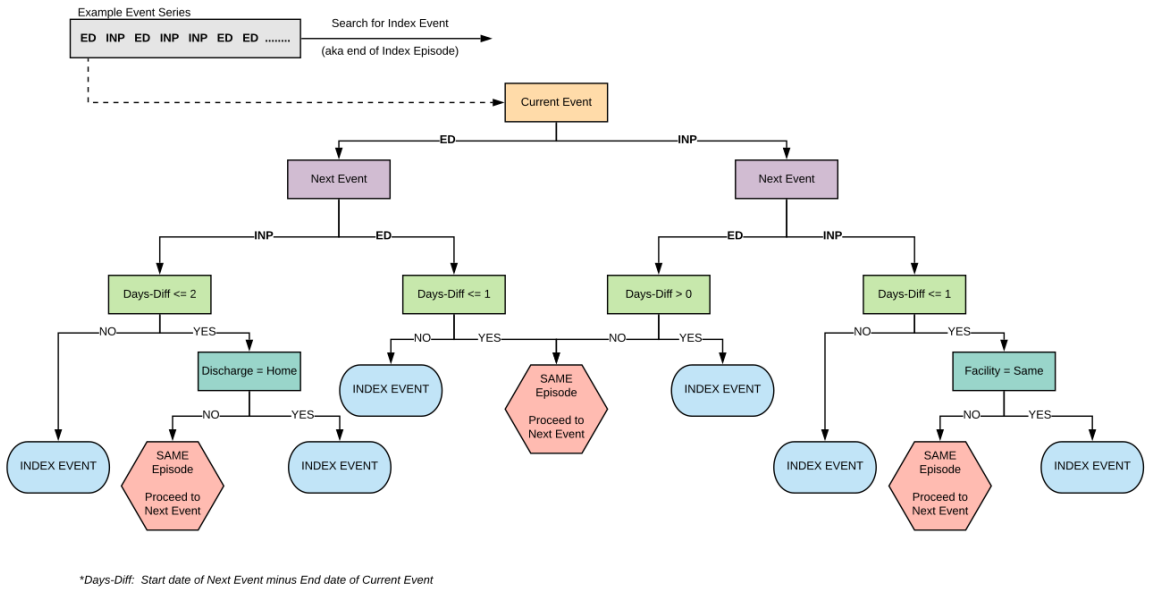

Figure 3.2: Episode generation: In this algorithm, the inputs are two consecutive events for the individual patient. The leave of the decision trees are represent "index event" (two consecutive events are not in the same episode) and "same episode" (two consecutive events are in the same episode)

2. The patient visits the ED. Then s/he is transferred to another ED within 24 hours.

3. The patient is admitted as an inpatient in the hospital. Then s/he is transferred to ED on the same day.

4. The patient is admitted as an inpatient in the hospital. Then s/he is transferred to the different hospital within 24 hours.

Otherwise, the episode ended with event 1 and event 2 started the new episode. Our episode generation algorithm combines continuous healthcare visits and diagnosis codes from these encounters into a single healthcare episode. An ECG record is linked to a healthcare episode if the acquisition date is within the time frame between an episode's admission date and discharge date.

We use standard 12-lead ECG traces and ECG measurements from the Philips IntelliSpace ECG system [11]. For each of the 12 leads, there was a sequence of

ECG voltage sampled at 500 Hz for 10 seconds. We also use the ECG measurements that are automatically generated by the ECG machine manufacturer's built-in algorithm, including atrial rate, P duration, RR interval, Q wave on-set, Fridericia rate-corrected QT interval, heart rate, PR interval, QRS duration, QT interval, Bazett's rate-corrected QT interval, frontal P axis, frontal QRS axis in the initial 40 ms, frontal QRS axis in the terminal 40 ms, frontal QRS axis, frontal ST wave axis (equivalent to ST deviation), frontal T axis, horizontal P axis, horizontal QRS axis in the initial 40 ms, horizontal QRS axis in terminal 40 ms, horizontal QRS axis, horizontal ST wave axis, and horizontal T axis. Table 3.1 lists ECG measurements.

The following reasons could degrade the quality of ECG signals: patient movement, respiration, sweating, muscle tremors, electrical interference, Etc. [11] The ECG machine manufacturer's built-in quality algorithm also identified poor quality ECGs and then displayed warning flags, showing the presence of muscle artifact, AC noise, baseline wander, QRS clipping, and leads-off [11]. Furthermore, the quality of the ECG trace is essential for machine learning models trying to learn the relationship between ECG abnormalities and output labels, as a poor quality ECG trace could be misinterpreted, which degrades performance of the learned ECG model. For that reason, we exclude poor quality ECG by annotation:

1. (1) ECG has artifacts measurement variables whose values are of 'Light', 'Marked', 'Severe'.
2. (2) QRS complexes are clipping.
3. (3) ECG signal is outside the measurement parameters of the instrument.
4. (4) QRS-related parameters cannot be measured in the rhythm group.

Variable	Definition	Unit	Short version
Atrialrate	Atrial rate	beats per minute	Atrial Rate
Pdur	P wave duration	Milliseconds	P duration
RRint	RR interval	Milliseconds	RR Interval
Qonset	Q wave onset	Millivolts	Q onset
QTcf	Fridericia Rate-Corrected QT interval	Milliseconds	Fridericia QTc
Heartrate	Heart Rate	Milliseconds	HR
PRint	PR interval	Milliseconds	PR interval
QRSdur	QRS duration	Milliseconds	QRS duration
QTint	QT interval	Milliseconds	QT interval
QTcb	Bazett's Rate-Corrected QT interval	Milliseconds	Bazett's QTc
Pfrontaxis	Frontal P axis	Degrees	Frontal P
i40frontaxis	Frontal QRS axis in Initial 40 ms	Degrees	Frontal i40msQRS
t40frontaxis	Frontal QRS axis in Terminal 40 ms	Degrees	Frontal t40msQRS
Qrsfrontaxis	Frontal QRS axis	Degrees	Frontal QRS
Stfrontaxis	Frontal ST wave axis	Degrees	Frontal ST
Tfrontaxis	Frontal T axis	Degrees	Frontal T
Phorizaxis	Horizontal P axis	Degrees	Horizontal P
i40horizaxis	Horizontal QRS axis in Initial 40 ms	Degrees	Horizontal i40msQRS
t40horizaxis	Horizontal QRS axis in Terminal 40 ms	Degrees	Horizontal t40msQRS
Qrshorizaxis	Horizontal QRS axis	Degrees	Horizontal QRS
Sthorizaxis	Horizontal ST wave axis	Degrees	Horizontal ST
Thorizaxis	Horizontal T axis	Degrees	Horizontal T
tonset	T wave onset	Millivolts	T onset

Table 3.1: Full forms of ECG measurement names

The Philips DXL ECG algorithm identifies each of the above poor quality criterias [11].

After excluding the ECGs that could not be linked to any episode, ECGs of patients under 18 years of age, and ECGs with poor signal quality, the remaining analysis cohort contained 1,605,268 ECGs from 748,773 episodes of 244,077 patients. See Figure 3.3 for the flowchart of the study design, showing sample sizes for the overall study, experimental splits, and different outcomes.

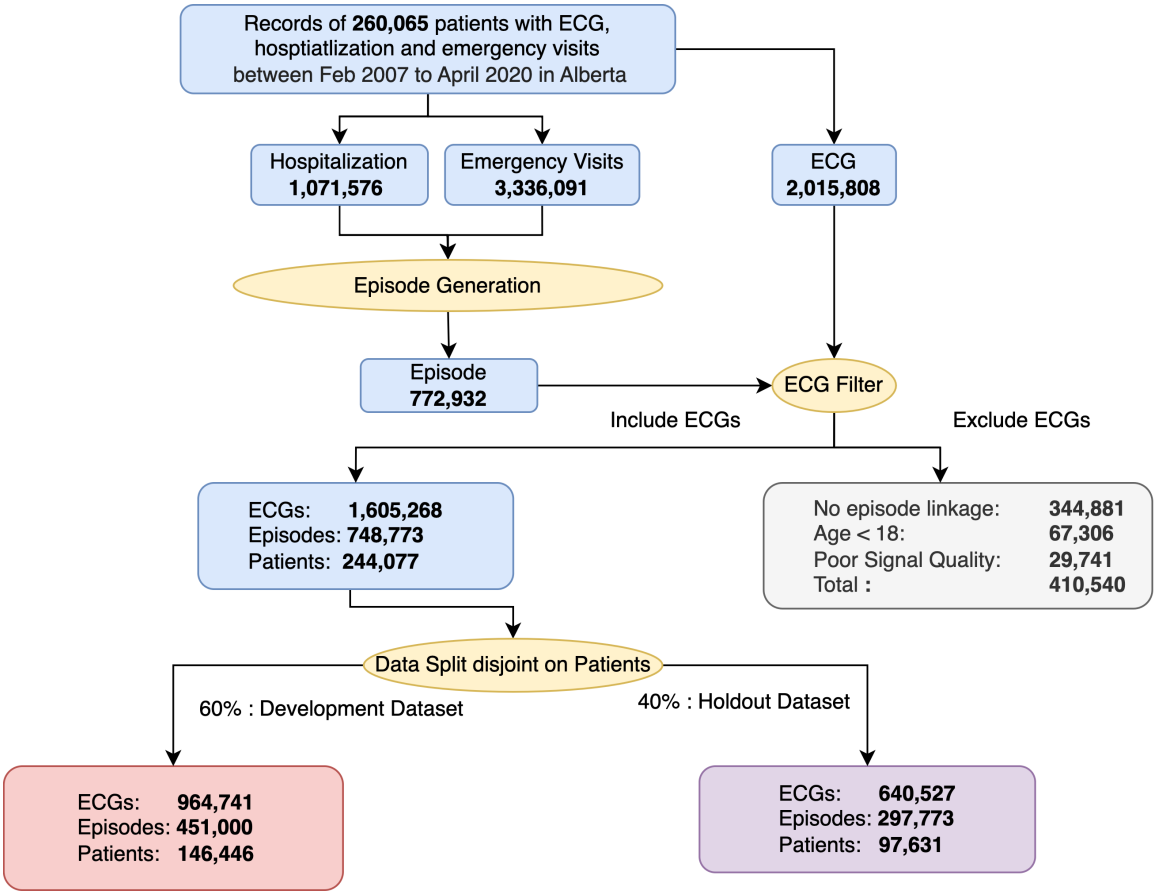

Figure 3.3: Flowchart of the study design showing the sample sizes for different splits and outcomes

This study was approved by the University of Alberta Research Ethics Board (Pro00120852). The ethics panel determined that the research is a retrospective database review for which subject consent for access to personally identifiable health information would not be reasonable, feasible, or practical.

3.2 Learning Algorithm

In this section, we quickly summarize the three main learning algorithms used for learning the diagnostic models used in Chapter 4 and the prognostic models used in Chapter 5. In the diagnosis tasks, we convert the diagnosis problem to a multi-label learning task using ResNet and use XGB to predict each single diagnosis code. In the prognosis task, we provide two different ways of modeling risk for all-cause mortality. (1) Binary mortality classification methods: we provide the ECG prognosis models to predict the calibrated probability of mortality for 1-year, short-term (30-days), and long-term (5-years) mortality. (2) ISD methods: we estimate the time until death and produce a survival probability curve for each individual patient. The following models are implemented in Python 3.8. We train all models with 8 Tesla V100-SXM2 GPUs and 32 GB of RAM per GPU.

3.2.1 Gradient boosted tree ensembles (XGB) model

We train gradient boosted tree ensembles (XGB) [46] models, which are ensembles of multiple decision trees. The XGB model uses binary logistic regression as the objective task and squared loss as the objective function. We tuned the hyperparameters, such as maximum tree depth, min child weight, and scale positive weight, using 5-fold grid-search internal cross-validation within the training sets. The models are learned for a maximum of 200 epochs, and the learning process is stopped if performance loss in the training/tuning set does not reduce for ten consecutive epochs. To learn multi-labels – e.g. in Chapter 4 – we learn an XGB model for each single label (diagnosis code).

3.2.2 Deep learning (DL) model

We use DL models to classify diagnosis codes in Chapter 4 and mortality binary outcomes in Chapter 5. In the DL model, we implement a convolutional neural network (CNN) based on the ResNet, consisting of a convolutional layer, 4 residual

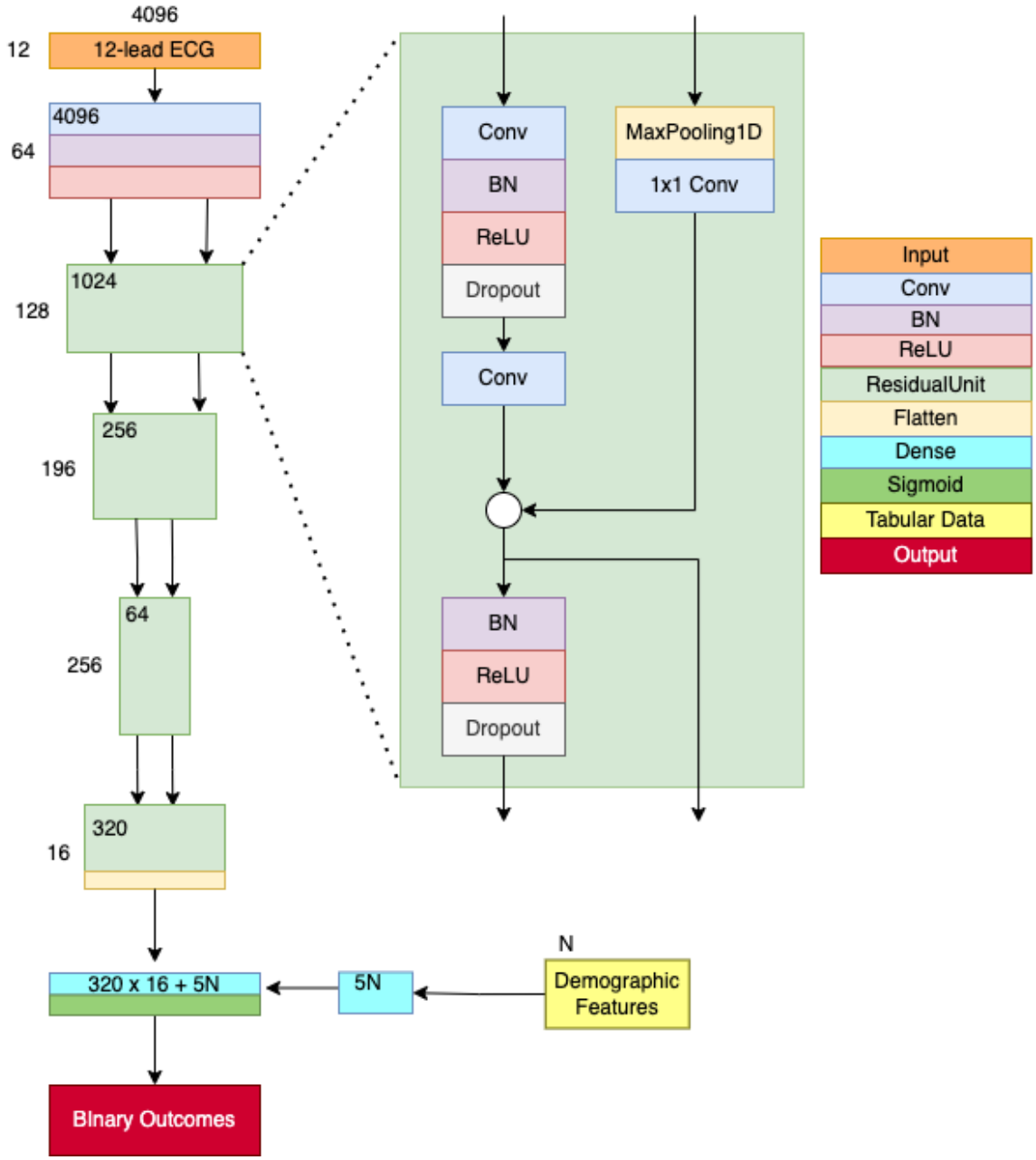

Figure 3.4: Schematic of deep learning model architecture used in the study

blocks with 2 convolutional layers per block, followed by a dense layer; see Figure 3.4. We use batch normalization [47], ReLU, and dropout [48] after each convolutional layer. Our architecture is based on a model trained on a large ECG dataset from Brazil to identify abnormalities in 12-lead ECGs [14] with some modifications to accommodate tabular data input for demographic features and binary output. Each

ECG instance is represented as a 12 x 4096 numeric matrix. If additional features such as age, sex or other tabular features are used, they are input as the binary feature (sex; 1 feature) or continuous values (age; 1 feature), then passed to a 5N fully connected layer (where N is the number of tabular features), then concatenate with the dense layer, and finally pass to a sigmoid function to produce the output. Binary cross-entropy is used as the loss function with the initial learning rate of $1x10^{-3}$, Adam optimizer [49], ReLU activation function, kernel size of 16, batch size of 512, and a dropout rate of 0.2 with other hyperparameters set to default. Models are learned for a maximum of 70 epochs. The learning rate is reduced to $1x10^{-5}$ if there is no improvement in tuning loss for seven consecutive epochs, and the learning process is stopped if the loss in the tuning set does not reduce for nine epochs. Training Each DL model took 30 min per epoch.

3.2.3 Multi-Task Logistic Regression

Survival prediction corresponds to estimating the time until an event of interest will occur for individuals – here, we learn a model that produces a survival distribution for each patient. We obtain our survival model with 2 different approaches: Multi Task Logistic Regression (MTLR) [44] and the Neural Multi-Task Logistic Regression model (N-MTLR) [45].

Notation

Here, we define our notation: $D = \{[\vec{X}_i, T_i, \delta_i]\}_i$: is a survival dataset, where $\vec{X}_i$ is the patient's vector features, T_i is a non-negative value, which is the time until i-th patient is either censored or dead, and $\delta_i \in 0, 1$ indicates if i-th patient is dead (1) at that time T_i, or is censored (0).

$f(t) = p$(die at time t) (simple description without patient's features) is the probability density function for the event happening at time t and $F(t) = \int_{s=0..t} f(s)ds = P$(dead before t) is the cumulative density function for the event that happened be-

fore time t.

$S(t)$ is the survival function:$S(t) = P(T > t) = 1 - F(t)$.

$h(t)$ is the hazard function representing the risk of the event of interest.

$$h(t) = \lim_{dt \to 0} \frac{P[t \leq T < t + dt | T \geq t]}{dt}$$

or

$$h(t) = \frac{f(t)}{S(t)}$$

Multi-Task Logistic Regression (MTLR) model

The MTLR model is essentially a series of logistic regression models, each providing the probability of the event of interest happened within each interval. The MTLR model is built by the following steps:

1. We first divide the analysis timeline into m time bins. $\tau = (t_1, ...t_m)$ - e.g., we could uniformly divide 1 year study time into 10 disjoint bins from the distributions of uncensored patients.

2. We build a logical regression for each time bin that estimates the chance that an event happens more than the i-th time bin.

$$P_{\vec{\theta_i}}(T \geq t_i | \vec{x}) = (1 + \exp(\vec{\theta_i} \cdot \vec{x} + b_i))^{-1}, 1 \leq i \leq m$$

Then we generate the binary vector Y with m time bins, whose i-th entry is 1 if the event occurs in the i-th time bin. For example, if the patient's event occur in the 4-th bin, then$\vec{Y} = [0, 0, 0, 1, 0, 0, 0, 0, 0, 0]^\top$. However, if the patient is censored, as the event occurs after he left the analysis, then we know that the patient is alive until the censored time bin. For example if patient is censored after 7-th bin, then $\vec{Y} = [0, 0, 0, 0, 0, 0, ?, ?, ?, ?]^\top$.

Here, we consider the probability mass function for discrete time bins (probability density function for continuous time points), which gives the probability that the event

occurs in the i-th time bin. Using $\Theta = (\vec{\theta_1}, \ldots, \vec{\theta_m})$ (where each θ_i is a dimension vector representing weights of patient's features, X_i is the patients' features) and $f\Theta(\vec{x}, k) = \sum_{i=k+1}^{m}(\vec{\theta} \cdot \vec{x} + b_i)$ for $0 \leq k \leq m$, the probability mass function:

$$f(a_s, \vec{x}) = P[T \in [t_{s-1}, t_s]|\vec{x}] = \frac{\exp(\sum_{i=s}^{m-1} \vec{x} \cdot \vec{\theta_i} + b_i)}{\sum_{k=0}^{m} \exp(f_\Theta(\vec{x}, k))}$$

3. In N-MTLR [45] study, we use a deep learning framework via a multi-layer perceptron (MLP) by replacing $(\vec{\theta_i} \cdot \vec{x} + \vec{b_i})$ to deep learning feature $\Psi_i(\vec{x})$ in the following formula.

$$P_{\vec{\theta_i}}(T \geq t_i|\vec{x}) = (1 + \exp(\Psi_i(\vec{x})))^{-1}, 1 \leq i \leq m$$

Then the probability mass function in which the event occurs in s time bin could be converted the following formula.

$$f(a_s, \vec{x}) = P[T \in [t_{s-1}, t_s)|\vec{x}] = \frac{\exp(\sum_{i=s}^{m-1} \Psi_i(\vec{x}))}{\sum_{k=0}^{m} \exp(\sum_{j=k+1}^{m} \Psi_j(\vec{x}))}$$

3.3 Evaluation Methods

3.3.1 Binary classification evaluation metrics

In the classification task evaluation methods, our supervised machine learning models, which we also name as threshold-model, produces a numerical prediction score (a number between 0 and 1) for the binary label of the given instance – eg, for Mr ABC, that score may be 0.57. In order to assign predicted binary output to each label, we typically set a threshold Q, then if that score $\geq$ Q, then binary output y' = True, else y' = False. In Figure 3.5, when the predicted output is consistent with the ground truth value, we call it a True Positive (TP) or a True Negative (TN). If the predicted value is negative, but the ground truth value is positive, we call this outcome False Negative (FN). If the predicted value is positive, but the ground truth value is negative, we call this outcome a False Positive (FP). We binarized prediction probabilities into binary classes using optimal cut-points derived from training set with Youden's index [50], and generated the following threshold-based metrics: F1 Score, Specificity, Recall, Precision and Accuracy. Further, we evaluated calibration of our models to see whether predicted probabilities agree with observed proportions using Brier Score [51].

In the threshold based model, we need to pick the optimal cut point, based on Youden's index (Recall + Specificity - 1), to split the binary output. Moreover, we use the commonly used evaluations metrics (threshold-free metrics) in the clinician community. Then we take multiple values from 0 to 1 as thresholds and evaluate with aggregate values from multiple thresholds with threshold-free metrics. We reported the following threshold- free performance metrics on the holdout set - Area Under the Receiver Operating characteristic Curve (AUROC), Area Under the Precision-recall Curve (AUPRC), and Average Precision (AP) [52].

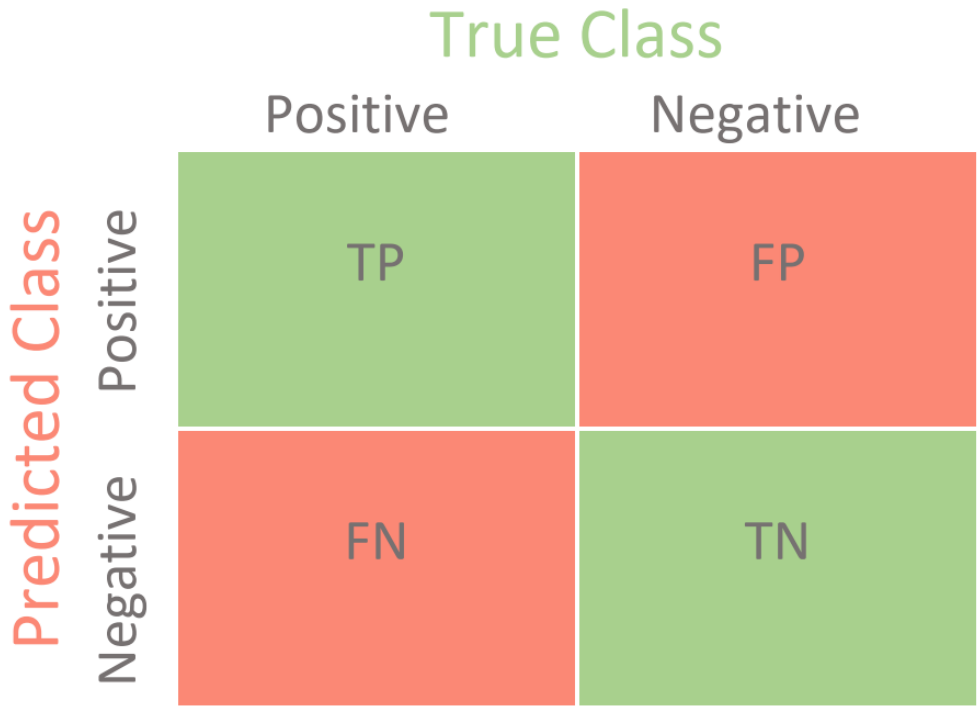

Figure 3.5: confusion matrix example

Accuracy

The accuracy is a popular score which is the number of exact matches in the overall test dataset in Equation 3.1. However, the cons of accuracy are also obvious. The classes of the labels might be unbalanced in deployment scenarios – e.g. in the early stage of the COVID-19 pandemic, there could be less than 100 positive COVID-19 patients in a city with a population of a million. Then, if we design a degenerate COVID-19 test that claims that all citizens of the city are negative for COVID-19, the accuracy would be very high, above 99.9%. However, note that none of the positive COVID-19 patients could be detected by the test, and hence this test would not help.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

Specificity

Specificity, which is also known as true negative rate, refers to the probability that an instance that is negative, is labeled as negative.

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

Recall

Recall, which is also known as true positive rate, refers to the probability that an instance is positive, is labeled as positive.

$Recall = \frac{TP}{TP+FN}$

Precision

Precision refers to the probability that an instance's label is positive, is predicted as positive.

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

AUROC

AUROC is the metric to evaluate the model performance over all thresholds. According to Fawcett's paper [5], we need two functions to draw the AUROC curve: the True Positive Rate (TPR) and False Positive Rate (FPR).

$TPR = \frac{TP}{TP+FN}$

$FPR = \frac{FP}{TN+FP}$

The TPR equals the ratio of the correct positive prediction results in overall positive samples during the test. The FPR equals the ratio of the correct prediction results in overall negative samples during the test.

FPR and TPR draw a ROC curve as x and y axes; when we choose a threshold value τ_i, we could receive TPR and FPR from this specific threshold i classifier. We can consider multiple threshold values, to produce multiple threshold classifiers, each with $(tpr_{\tau_i}, fpr_{\tau_i})$. Then we can connect the dots $(tpr_{\tau_i}, fpr_{\tau_i})$ and calculate the area

under the connected curve.

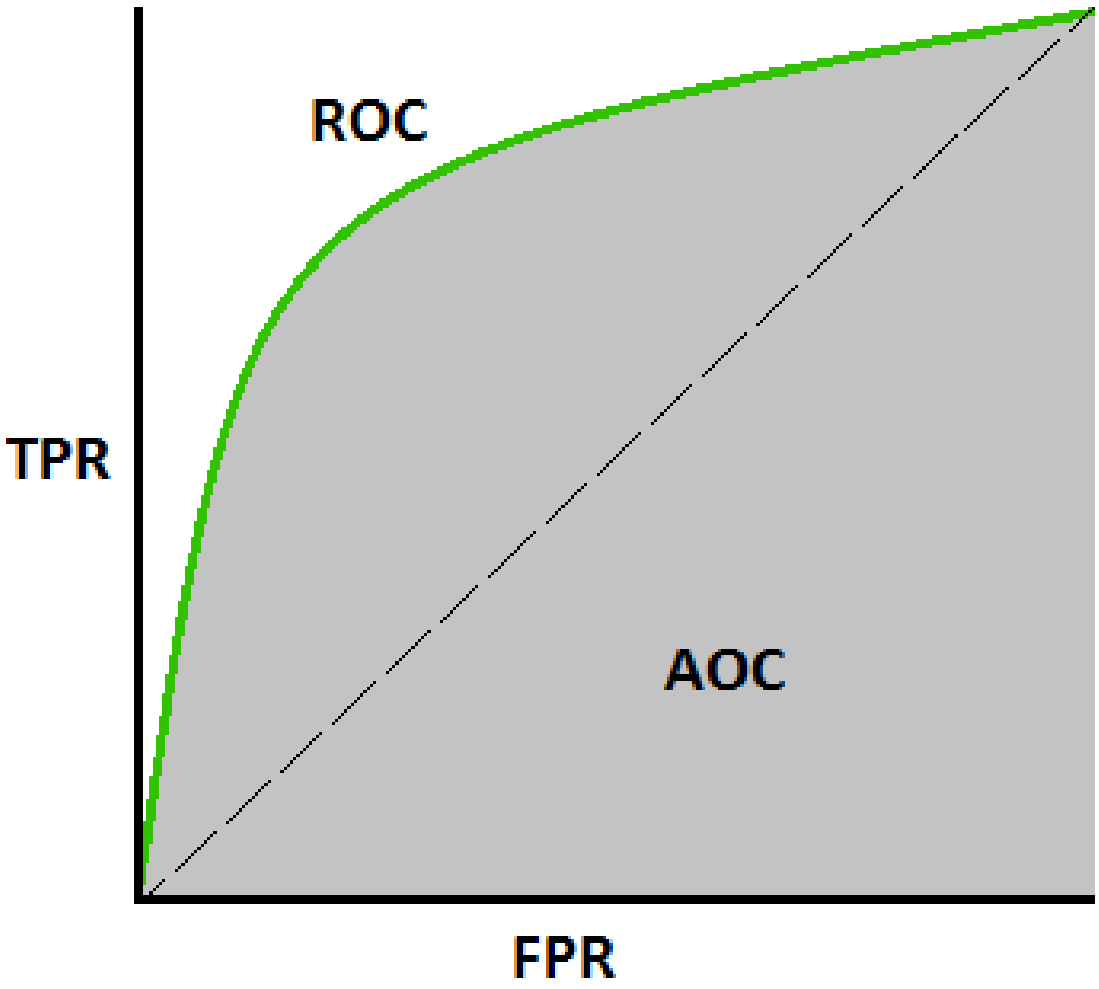

Figure 3.6: AUROC Plots

The baseline of AUROC is 0.5, which means the dummy classifier will predict all instances as constant. The advantage of AUROC is that AUROC is a single metric facilitating comparison to evaluate models with multiple thresholds between range 0 to 1. This is a well accepted performance metric in clinical literature [53] and it is useful when the optimal threshold to be used in the deployment environment is unknown during the time of model evaluation.

In our binary classification task, the data is imbalanced in terms of classes. Moreover, the cost of positive and negative misclassifications is not available to us. There-

fore, we rely on threshold-free methods that use precision (also known as positive predictive value) and recall (also known as sensitivity) metrics that are traditionally used in clinical literature. In addition to AUROC, we also include the AUPRC and AP as threshold-free binary classification performance metrics.

AUPRC and AP

AUPRC is the area under the curve formed by connecting by dots of $(recall_{\tau_i}, precision_{\tau_i})$, which are the recall and precision value associated with the i-th threshold. Then, we could draw a PR curve with $(recall_{\tau_i}, precision_{\tau_i})$. Accordingly, we calculate the area under the connected curve as the AUPRC. AUPRC displays the trade-off between precision (instead of specificity) and recall over all possible threshold values. Fawcett [5] has shown that the AUPRC is preferred over the AUROC for evaluating uncommon or rare diseases.

AP is another way to calculate the weighted average of precisions at each threshold:

$AP = \sum_{\tau_i}(R_{\tau_i} - R_{\tau_i - 1})P_{\tau_i}$ where P_{τ_i} and R_{τ_i} are the precision and recall at the i-th threshold [52].

Brier Score

A model with better calibration is equal to that the individual predicted probability is meaningful after considering the others' predicted probabilities [54]. Brier Score's baseline value is 25%; smaller score indicates better calibration [51].

$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$

F1-score

The F1-score is a binary classification evaluation metric, which is calculated from precision and recall of the evaluation data [55].

$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{\mathrm{tp}}{\mathrm{tp} + \frac{1}{2}(\mathrm{fp} + \mathrm{fn})}$

3.3.2 Survival prediction evaluation metrics

This section introduces the ISD evaluation metrics, which we used to compare the performance of our ISD models, from Haider's study [56]. We consider three metrics: model discriminability (Concordance Index), the average absolute value between actual times to predicted event times (L1 Loss), and goodness of a predicted probability score in an interval of time points (Integral Brier Score). Using parametric hypothesis tests for evaluating the calibration can be unsuitable since high statistical power would flag even minor deviations in calibration as significant. Since we have a population-level testing dataset, the calibration evaluations (e.g., 1-calibration [57] and D-calibration [56]) that use the χ^2 or HL tests, are too sensitive for a large data set.

Concordance index (C-index)

The concordance index (C-index) is one of the most common evaluation metrics for measuring the discriminability of a risk model. The range of the C-index is from 0 to 1, and the baseline of the C-index is 0.5, which means if we randomly assign the probabilities in instances, then the probability of correct order is 0.5. The larger value of the C-index indicates a better model performance. The larger value of the C-index indicates a better model performance. The C-index is also a generalization of AUROC.

The concordance algorithm must first determine the set of all comparable pairs. We define the pairs (i,j) of patients, which patient j is alive when patient i is dead, are comparable pairs (CP (D))

$$CP_{i,j} = I\{t_i < t_j \wedge \delta_i = 1\} + I\{t_i = t_j \wedge \delta_i = 1 \wedge \delta_j = 0\}$$

, where we use δ definition in the Section 3.2.3.

Next, when we calculate the correct ranked comparable pairs, we consider Antolini's study [58], where $r(\vec{x_i})$ indicates the risk score of patient i.

$$CPcorrect_{i,j} = I\{r(\vec{x_i}) < r(\vec{x_j})\} \cdot CP_{i,j}$$

Finally, we estimate C-index by estimating the percentage of correctly ranked comparable pairs in all comparable pairs. In another word, given two randomly patients i and j in a ISD model with 80% C-index, if $r(\vec{x_i}) > r(\vec{x_j})$, then there are 80% probabilities that patient i's event will come before patient j's event.

$$C - index = \frac{\sum_{i=1}^{n}\sum_{j=i}^{n} CPcorrect_{i,j;i\neq j}}{\sum_{i=1}^{n}\sum_{j=i}^{n} CP_{i,j;i\neq j}}$$

L1 loss

According to the definition of ISD model, ISD model produces an individual survival curve, which also predicts median survival time $\hat{t}^{0.5}$ which is 50% chances for patient to survive $\hat{t}^{0.5}$ until event comes. [1]

For an uncensored patient, the L1-loss: the average absolute value of the difference between the median survival time $\hat{t}^{0.5}$ and true event time $t_{event} = d$. However, since the true event time is unknown in censor patients, we consider two approaches: L1-hinge loss and L1-marginal loss.

We know that a patient survived at least until the censor time c. Then, we calculate L1-hinge loss is $max(0, c-\hat{t}^{0.5})$. If $c \leq \hat{t}^{0.5}$, then L1-hinge loss is 0, otherwise, L1-hinge loss is $c - \hat{t}^{0.5}$. Therefore, the L1-hinge loss is:

$$L1_{hinge}(D, \hat{t}^{0.5}) = \frac{1}{|D|}[\sum_{i\in D_{uncensor}} |d_i - \hat{t}_i^{0.5}| + \sum_{k\in D_{censor}} max(0, c_k - \hat{t}_k^{0.5})]$$

In L1-marginal loss, we also assign a "Best-Guess" (BG) value to each censored patient, which is that patient's expected survival time given that s/he already survived until c.

[1] According to Haider's study [56], we will use extrapolation to extend the last point, if user's survival curve does not reach to median until the study ends.

$$BG(c) = c + \frac{\int_c^{\infty} S(t)dt}{S(c)}$$

In our evaluation method, we use Kaplan-Meier [59] (estimator for the survival function) $\hat{S}_{KM}(\cdot)$ from training data set to estimate survival function $S(\cdot)$.

The L1-marginal loss is defined as

$$L1_{margin}(D, \hat{t}^{0.5}) = \frac{1}{|D_{uncensor}| + \sum_{k \in D_{censor}} a_k} [\sum_{i \in D_{uncensor}} |d_i - \hat{t}_i^{0.5}| + \sum_{k \in D_{censor}} a_k |BG(c_k) - \hat{t}_k^{0.5}|]$$

where α_k indicates the weight in each Best-Guess estimation to contribute the L1-marginal loss, since the instances with early censor time give less information to the cases with late censor time, we offer more weight to the late censor time instance in L1-marginal loss by set $a_k = 1 - \hat{S}_{KM}(c_k)$.

[Integrated] Brier Score

The Brier score [60] is a commonly used evaluate metric that (is claimed to) measures both calibration and discrimination [56, 61–63]. The Brier score defines the mean squared error between the true event status and the predicted survival probability at each time t'. The perfect Brier score is 0 whenever the model predicts only 1 at event time, and 0 at other times, but it is unrealistic in survival curve. The baseline of the Brier score evaluation metrics is 0.25, when we have a dummy model predict $\hat{S}(t'|\vec{x}) = 0.5$ for all patients x. Therefore we seek a model that does better. The lower score indicates the better model.

In the uncensored dataset ($D_{uncensor}$), the Brier score at time t' is

$$BS_{t'}(D_{uncensor}, \hat{S}(t'|\vec{x})) = \frac{1}{D_{uncensor}} \sum_{|\vec{x_i}, d_i| \in D_{uncensor}} (I[d_i \leq t'] - \hat{S}(t'|\vec{x_i}))^2$$

.

To extend the Brier score to a series of time points, we use the Integrated Brier Score (IBS) which provides the average value of Brier score across the time interval.

PatientId	Time(t)	censor bit(1)	G^ (t)	Weight 1/G^ (t_i)	KM(t)
	0		1		1
S1	1	uncensored	3/4	1	4/5
S2	2	censored	3/4	0	4/5
S3	3	uncensored	3/8	4/3	8/15
S4	4	censored	3/8	0	8/15
S5	5	uncensored	3/8	8/3	0

Table 3.2: Example with IPCW

$$IBS(\tau, D_{uncensor}, \hat{S}(\cdot|\cdot)) = \frac{1}{\tau}\int_0^{\tau} BS_t(D_{uncensor}, \hat{S}(t|\cdot)dt$$

Note this analysis implicitly assumes we have only uncensored patients ($D_{uncensor}$), we add the subset of censored patients (D_{censor}) into the Brier score calculation. Graf, Erika et al. [64] propose using the Inverse Probability of Censoring Weights (IPCW) method, which the censor instance weight equally to the uncensored instances. In another words, the uncensored patients' weight $\vec{G}(t) = \prod_{j:t_j<t} \frac{n_j - cn_j}{n_j}$, where n_j indicates the number of all patients are possible censor at time t and cn_j indicates the number of all censor patient at time t.

Here is the example:

In Table 3.2, we list 5 patients and show weight in IPCW.

1. S1: S1: We count S1's $\hat{G(t)}$ as 1, since no one died before S1.

2. S2: S2 is censored and three patients who are possibly censored or died after S2's censor time. We split the weight equally into 3 parts.

3. S3: When we calculate the weight of patient S3, we are not only adding weight of 1 for patient S3, but also we add the weight of 1/3 from patient S2.

4. S4: S4 is censored and only S5 who is possibly censored or died after S4's censor time.

5. S5: when we calculate S5's weight, we sum S2's weight of 1/3, S4's weight of 4/3 (which includes the S2's weight of 1/3) and S5's original weight of 1, to get the total weight of 8/3.

We use the following formula to calculate the IBS with IPCW weight.

$$IBS(\tau, D, \hat{S}(\cdot|\cdot)) = \frac{1}{\tau}\int_0^\tau \left(\frac{\left(0 - \hat{S}(t, \vec{x}_i)\right)^2 \cdot \mathbf{1}_{t_i \leq t, \delta_i = 1}}{\hat{G}(t_i)} + \frac{\left(1 - \hat{S}(t, \vec{x}_i)\right)^2 \cdot \mathbf{1}_{t_i > t}}{\hat{G}(t)} \right) dt$$

where **1** is the indicator function.

3.3.3 Bootstrap Model Comparison

For model comparisons, for each evaluation, we use 1000 iterations of random and replace a selection of ECG instances to demonstrate consistency in the model performance. We use all available ECGs in the training set and corresponding labels while training these models. We use the same training and testing splits (including the random selections) for the various modeling scenarios to compare performance directly. The performance scores are compared between models by bootstrapping 100 instances with random replacement sampling from each of 10 iterations of random ECG selection to generate a total of 1000 bootstrap replicates. The difference in the model performances is evaluated based on the overlap of 95% confidence intervals of evaluation scores of the compared models.

Chapter 4

ECG-based diagnosis for multiple diseases

In recent years, DL methods have performed close to current levels of clinical expertise in diagnosing cardiovascular diseases [17] and heart abnormalities. As discussed in Section 2.2, several clinical studies show strong associations between non-cardiovascular diseases and ECG abnormalities. This chapter describes how we first extract diagnosis labels from AHS data, then implement multiple diagnosis models, and test these models to show evaluation performance for both cardiovascular and non-cardiovascular diseases. In addition, we also explore the challenge of implementing this process to diagnose for COVID-19. Despite having a relatively small number of COVID-19 instances, we have been able to identify that transfer-weight DL method can improve the diagnosis of COVID-19.

4.1 Method

4.1.1 Analysis Cohort

We split our ECG dataset into the development set (random 60%: 146,446 patients with 964,741 ECGs, used for training and internal validation) and external holdout set (remaining 40%: 97,631 patients with 640,527 ECGs) while ensuring that ECGs from the same patient are not shared between the sets.

Characteristics of patient cohorts used in the study are described in Table 4.1. At the time of the ECG, the average age of patients in the development and holdout set is 65.85 years. Recall, however, that we select the first ECG per episode from the holdout to use for the final evaluations. The average age here is slightly lower, 64.66 years because older patients have more ECGs than younger ones and the first ECG is in the early stage of the medical care. Similarly, men have more ECGs, so the proportion of men is slightly lower in the first ECG per episode in the evaluation set than in the development of holdout sets (52.72% vs 56.60%).

4.1.2 Prediction Task

We use the population-based data set with various medical conditions in-hospital ECG in this task. Here, we use diagnoses coded using the tenth World Health Organization International Classification of Diseases (ICD-10) [4], which we use for diagnosis labels. Whenever there is more than one ECG in a healthcare episode, we use only the first ECG (with acceptable signal quality) for evaluation, as it would be preferable in actual clinical practice to make a diagnostic prediction at the first point of care in the ED or hospital. Therefore, we train one ResNet and XGB model for each of the full sets of ICD-10 codes. To do this, we first train and evaluate the performance within the internal five cross-validation set and use this to select 275 top-performance ICD-10 codes with the best discrimination performance (AUROC). We then retrain the models (for each of these selected labels) on the entire internal validation set and

	Full Data	Development set	Holdout set	First ECG per episode in holdout set*
ECG Number	1605268	964741	640527	297773
Atrial rate	85.60 ± 46.15	85.56 ± 46.11	85.67 ± 46.20	84.89 ± 42.06
P duration	155.92 ± 116.60	156.00 ± 116.39	155.79 ± 116.91	158.93 ± 113.25
RR interval	790.81 ± 213.11	790.90 ± 212.63	790.68 ± 213.83	782.79 ± 201.37
Q wave onset	508.84 ± 6.51	508.82 ± 6.29	508.87 ± 6.82	508.99 ± 6.21
Fridericia Rate-Corrected QT interval	434.86 ± 38.05	434.96 ± 38.04	434.71 ± 38.07	431.33 ± 35.20
Heart Rate	81.64 ± 23.22	81.61 ± 23.18	81.69 ± 23.28	81.91 ± 22.04
PR interval	169.34 ± 38.46	169.44 ± 37.66	169.18 ± 39.65	167.89 ± 39.12
QRS duration	101.36 ± 24.26	101.40 ± 24.23	101.31 ± 24.30	99.66 ± 23.04
QT interval	399.81 ± 54.83	399.94 ± 54.74	399.63 ± 54.96	395.40 ± 51.06
Bazett's Rate-Corrected QT interval	455.02 ± 40.09	455.10 ± 40.09	454.89 ± 40.09	451.86 ± 37.18
Frontal P axis	44.85 ± 35.52	44.81 ± 35.41	44.91 ± 35.69	45.43 ± 34.22
Frontal QRS axis in Initial 40 ms	27.50 ± 46.30	27.44 ± 46.37	27.59 ± 46.20	28.69 ± 43.58
Frontal QRS axis in Terminal 40 ms	45.36 ± 88.15	45.74 ± 88.28	44.80 ± 87.96	45.41 ± 87.24
Frontal QRS axis	19.98 ± 54.37	20.04 ± 54.38	19.88 ± 54.37	21.07 ± 52.31
Frontal ST wave axis	90.94 ± 88.23	90.98 ± 88.07	90.87 ± 88.48	84.27 ± 85.68
Frontal T axis	55.70 ± 67.76	55.48 ± 67.60	56.03 ± 68.00	50.44 ± 59.22
Horizontal P axis	20.69 ± 47.30	20.63 ± 47.14	20.77 ± 47.52	19.43 ± 44.76
Horizontal QRS axis in Initial 40 ms	27.79 ± 48.39	27.86 ± 48.17	27.69 ± 48.71	29.18 ± 43.86
Horizontal QRS axis in Terminal 40 ms	34.10 ± 129.50	33.94 ± 129.40	34.35 ± 129.66	33.33 ± 128.73
Horizontal QRS axis	-0.91 ± 78.19	-1.11 ± 77.91	-0.61 ± 78.62	-1.44 ± 74.31
Horizontal ST wave axis	97.02 ± 64.99	96.98 ± 65.00	97.09 ± 64.97	93.18 ± 60.92
Horizontal T axis	64.46 ± 58.98	64.27 ± 58.96	64.73 ± 59.01	57.77 ± 51.43

Table 4.1: Characteristics of patient cohorts used in the diagnosis tasks

evaluate the external holdout set based on the set of diseases and categories selected during the internal validation.

In short, the main goal of our study is to identify which diseases (considering both ones that had previously known associations with ECGs and the others) can be accurately diagnosed from the patient's first ECG per episode based on a learned DL model. We show the data details and models in the process flowchart, Figure 4.1.

Based on the process described in Section 3.1, we generate hospital episodes from emergency department (ED) visits or inpatient (INP) hospitalizations encounters. In each ED visit or INP hospitalization encounter, medical doctors recorded the diagnosis codes in ICD-10 format. Then, we assign doctors' diagnosis results with each ECG, based on two assumptions" (1) since most episodes have a short duration, we assume that all ECG tests in a single episode have the same diagnosis labels. (2) We assume that INP doctors report more comprehensive medical tests and provide more accurate diagnoses than ED doctors. Hence, if INP encounters and ED visits are in the same episode, we would only take INP encounters diagnosis.

Our dataset includes 13,179 unique ICD-10 codes/diseases. According to our assumptions #1 and #2, a patient experiences the same diseases throughout a single episode. After data cleaning and exclusions, we extract 11,221 unique ICD-10 codes. Each ICD-10 code is 3 to 7 characters that specify a specific disease, consisting of one English letter, followed by at least two Arabic numerals, which denote the general disease category. E.g., 'I214' refers to 'Non-ST elevation (NSTEMI) myocardial infarction', and 'I21' refers to its broader category, 'Acute myocardial infarction'. We extract diagnostic information from ICD codes and corresponding categories and use them for prediction modeling. To diminish the size of predicted labels, we use only the 1414 ICD codes (full code, exact match) that are each linked to at least 1000 ECGs.

It aims to provide a proof-of-concept for high-throughput screening of an ICD-wide range of diseases based on ECG.

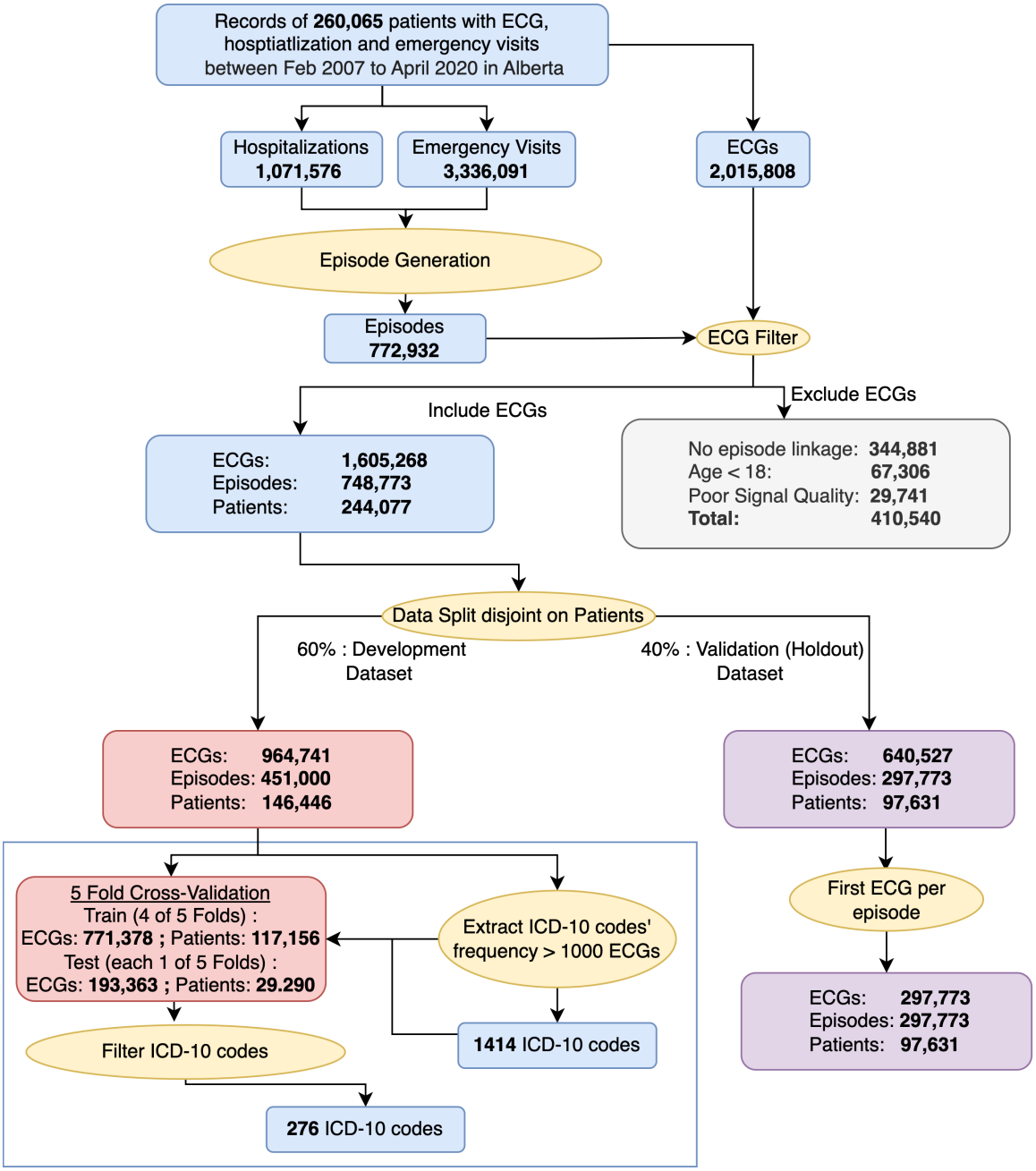

Figure 4.1: Flowchart of the study design showing the sample sizes for different splits and prediction task (ICD-10 codes)

4.2 Result

In our internal five cross-validations, we found a total of 276 ICD-10 codes (275 ICD-10 codes in the ResNet model and 45 ICD-10 codes in the XGB model) out of 1414 ICD-10 codes that have AUROC $\geq$ 80%, and (AUPRC $\geq$ 0.05 or AP/test) positive pro- portion $\geq$ 20. These 276 ICD-10 codes belong to 17 categories from ICD-10 code descriptions. Finally, we examine the replication of these lists in the external validation with ResNet and XGB models. Then, we list the number of ICD-10 diagnosis codes that model the performance of AUROC $\geq$ 80%, and AUROC $\geq$ 90% in Table 4.2. Additionally, the Appendix provides the AUROC bar plots for all 1414 ICD code categories and the diagnosis codes.

# of ICD-10 codes	with ResNet		XGBoost	
ICD-10 category	AUROC>90	AUROC>80	AUROC>90	AUROC>80
Certain infectious and parasitic diseases	0	4	0	3
External causes of morbidity and mortality	3	13	1	13
Injury, poisoning and certain other consequences of external causes	3	19	1	13
Codes for special purposes	0	1	0	1
Diseases of the respiratory system	1	9	1	2
Factors influencing health status and contact with health services	4	15	2	16
Endocrine, nutritional and metabolic diseases	3	11	1	4
Congenital malformations, deformations and chromosomal abnormalities	2	3	1	2
Diseases of the genitourinary system	1	8	1	4
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	3	8	0	4
Diseases of the musculoskeletal system and connective tissue	0	1	0	2
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	0	3	0	1
Neoplasms	0	9	0	4
Diseases of the circulatory system	26	61	8	34
Mental and behavioural disorders	11	24	9	20
Diseases of the digestive system	6	15	0	5
Diseases of the nervous system	1	5	0	3

Table 4.2: list of categories with number of top performing ICD-codes (wrt AUROC) that could be predicted from the patient's first in-hospital ECG using DL.

4.3 COVID-19 Diagnosis

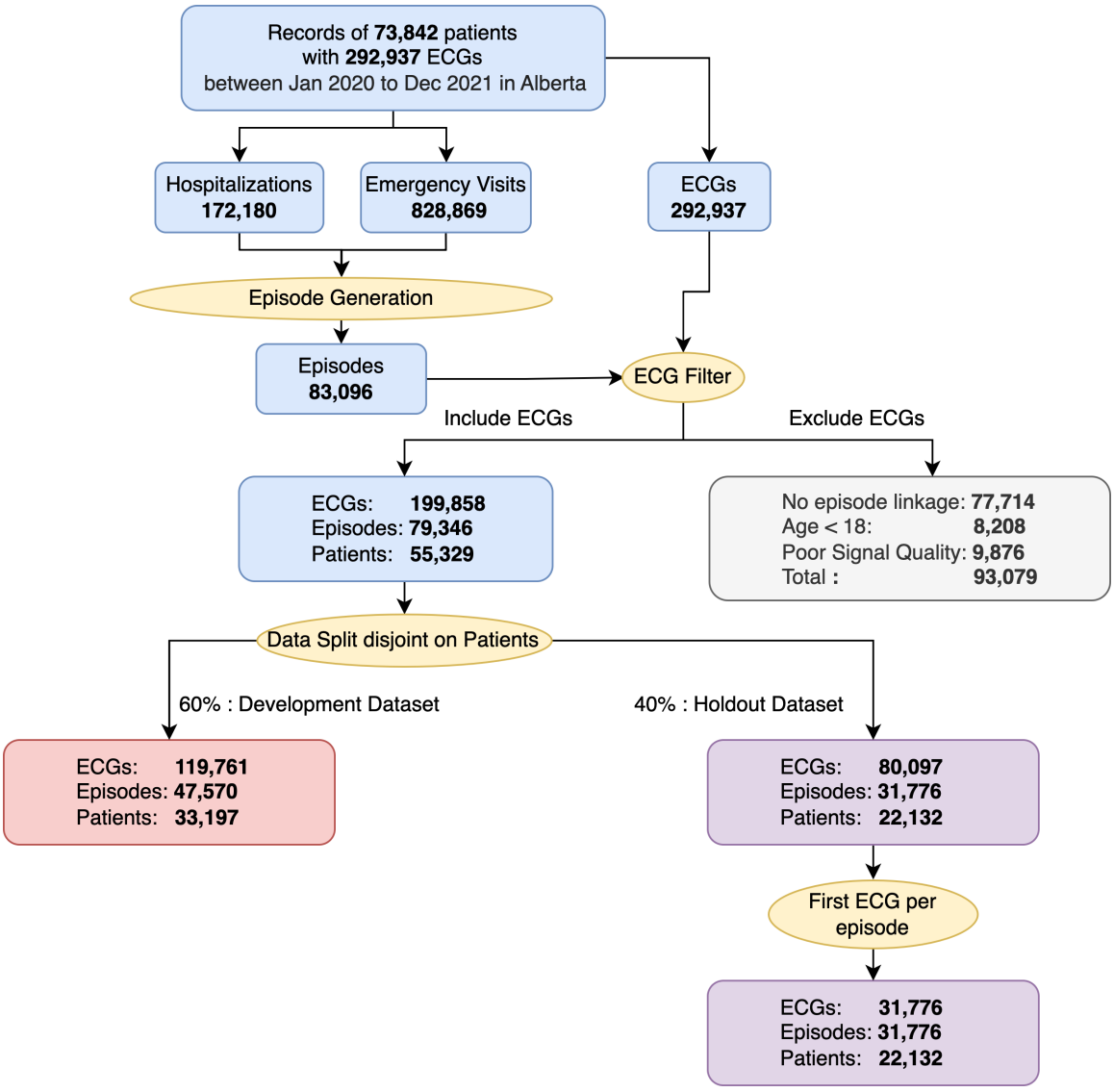

Figure 4.2: Flowchart of the study design for subset of ECGs in COVID-19 pandemic duration, showing the sample sizes for different splits.

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS CoV-2 virus. This worldwide pandemic broke out in Hubei Province, China, at the end of 2019. There are 546,363,985 confirmed cases and 6,336,802 deaths as of July 3, 2022 [65]. Initially regarded as a respiratory infection, COVID-19 is now known to affect all major systems in the body, including the cardiovascular system by causing myocardial damage, vascular inflammation, plaque instability, and myocardial infarction [66].

Kaliyaperumal et al. [67] report that COVID-19 patients can have various ECG abnormalities – ischemic changes, rate, rhythm abnormalities, and conduction defects and can also express a diverse set of comorbidities with ECG involvement. In this context, ECG-based artificial intelligence, along with the utilization of bedside rapid diagnostic tests to detect COVID-19, could prove helpful in recognizing patients who require urgent definitive management. Furthermore, multiple reports [68–74] show that DL models could perform well in diagnosing COVID-19 from publicly available camera-captured ECG image data [75].

Moreover, cardiac involvement in COVID-19 results in poor prognosis and adverse outcomes [76]. Hence, monitoring the cardiac function that identifies the need for prompt action is crucial. The ECG, which provides essential information about the heart's electrical activity, is a simple point-of-care diagnostic tool [77] that can be employed to assess cardiovascular involvement in COVID-19 patients. Unfortunately, despite the novel coronavirus's massive impacts, there are no available prediction models that have been verified on the population-scale ECG dataset that can accurately identify who has COVID-19.

In this study, we use population-scale administrative health records and large ECG datasets, with two-year coverage from the start of the pandemic (Jan 2020) to December 2021, to develop DL models to diagnose if a patient with ECG data has COVID-19.

Our AHS data includes 73,842 patients during the pandemic (2020-2021) who took 282,837 ECG tests. We exclude the ECGs that could not be linked to healthcare episodes, are of patients < 18 years old, or are poor-quality ECGs. Then, we split our ECG dataset into the development dataset B (random 60%: 33,197 patients with 119,761 ECGs, used for training and internal validation) and holdout dataset C (remaining 40%: 22,132 patients with 80,097 ECGs), while ensuring that we did not share ECGs from the same patient; for details; see Figure 4.3. We also have dataset A which includes 260,774 patients with 1,992,415 ECGs in the pre-pandemic period

(Feb, 2007 - Dec, 2019); See description in Figure 4.2.

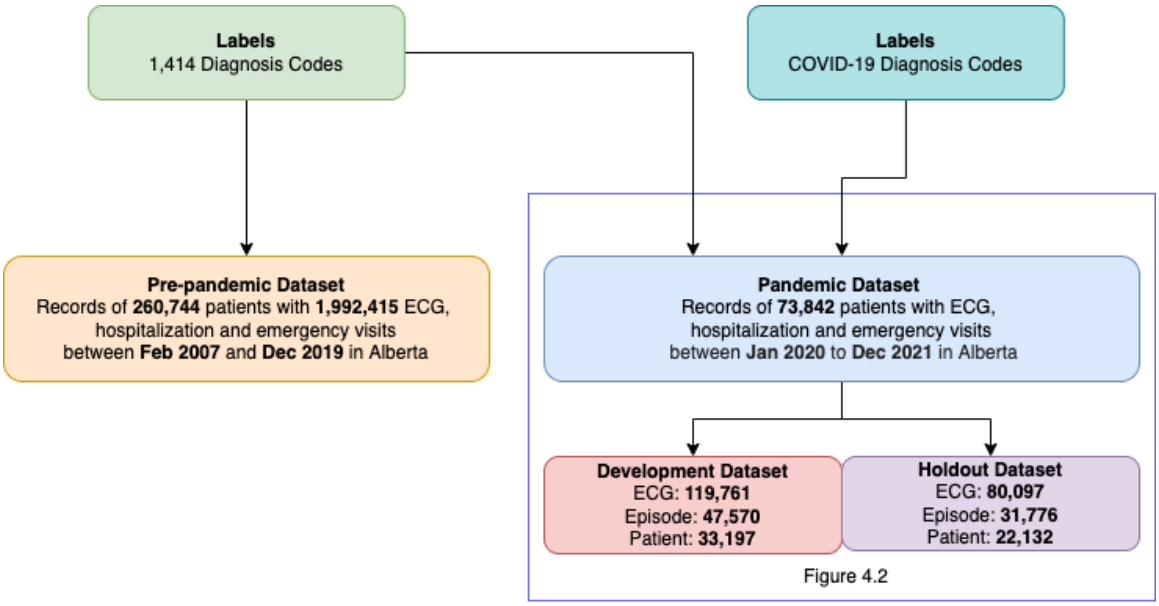

Figure 4.3: ECG data summary before pandemic, and after pandemic

According to forementioned ECG dataset splits, we design 4 different experiments in Table 4.3.

1. Model 1: we use instances in the development dataset to train with COVID-19 diagnosis labels. Then, we evaluate the model using the holdout dataset.

2. Model 2: we use instances in the development dataset to train with 1,415 labels: the 1,414 ICD-10 codes and COVID-19 diagnosis labels. Then we evaluate the model using the holdout dataset for just the COVID-19 diagnosis prediction.

3. Model 3: we use the development data to train the model with 1,414 diagnosis ICD-10 codes (as explained in Section 4.1). Then we froze the model's earlier layer weights and changed the last layer to predict COVID-19 labels. In this way, we employed the transfer learning from non-COVID diagnosis to train the COVID-19 diagnosis. Finally, we evaluated the model using a holdout dataset with COVID-19 labels.

	Pretrain Dataset		Train Dataset		Test Dataset	
	Instances	Label	Instances	Label	Instances	Label
Model 1	-	-	Development Data	Covid-19	Holdout Data	Covid-19
Model 2	-	-	Development Data	1414 ICD + Covid-19	Holdout Data	1414 ICD + Covid-19
Model 3	Development Data	1414 ICD	Development Data	Covid-19	Holdout Data	Covid-19
Model 4	Pre-pandemic Data	1414 ICD	Development Data	Covid-19	Holdout Data	Covid-19

Table 4.3: Description of models

4. We used the pre-pandemic data to train the model with 1,414 diagnosis ICD-10 codes (as explained in Section 4.1)). Next, we froze the model's earlier layer weights and changed the output layer to predict COVID-19 labels. In this way, we used the transfer learning from non-COVID diagnoses and pre-pandemic data to train a model for the COVID-19 diagnosis. Finally, we evaluate the model using a holdout dataset with COVID-19 labels.

In the model comparison Table 4.4 and AUROC bar Figure 4.4, we take the first ECG per episode in evaluating model performance and run bootstrap comparison methods. We used model 1 alone to establish a baseline model performance, which had an AUROC of 0.6064 for predicting COVID-19. Models 2 and 3, which use the information for the other 1,414 diagnosis codes from development data, have a substantially higher performance with AUROC of 0.6235 and 0.6125, respectively. Finally, the best model performance is model 4 with AUROC of 0.7178, which used a model whose weights were "transfer learned" from pre-pandemic data.

Limitation: In our data, not all COVID-19 patients took an ECG test during the pandemic season because of the infectious nature of the disease. Also, it is likely that

	Model 1	Model 2	Model 3	Model 4
AUROC	0.606 (0.595 - 0.619)	0.613 (0.601 - 0.625)	0.624 (0.612 - 0.635)	**0.718 (0.707 - 0.729)**
AUPRC	0.0649 (0.0604 - 0.0699)	0.0638 (0.0597 - 0.068)	0.0716 (0.0663 - 0.0769)	**0.13 (0.118 - 0.141)**
AP	0.0654 (0.0609 - 0.0704)	0.0642 (0.06 - 0.0685)	0.0722 (0.0667 - 0.0776)	**0.131 (0.119 - 0.142)**
F1 Score	0.107 (0.102 - 0.112)	0.103 (0.0987 - 0.108)	0.108 (0.102 - 0.113)	**0.164 (0.156 - 0.172)**
Specificity	0.577 (0.572 - 0.581)	0.456 (0.452 - 0.46)	0.586 (0.581 - 0.59)	**0.797 (0.794 - 0.801)**
Recall	0.585 (0.565 - 0.604)	**0.708 (0.689 - 0.726)**	0.576 (0.555 - 0.596)	0.487 (0.467 - 0.508)
Precision	0.059 (0.0561 - 0.0621)	0.0558 (0.0531 - 0.0582)	0.0594 (0.0563 - 0.0624)	**0.0983 (0.0932 - 0.103)**
Accuracy	0.577 (0.573 - 0.581)	0.467 (0.463 - 0.471)	0.585 (0.581 - 0.589)	**0.784 (0.78 - 0.787)**
Brier Score	0.0416 (0.0402 - 0.0431)	0.0414 (0.04 - 0.0429)	**0.0413 (0.0399 - 0.0428)**	0.0415 (0.0401 - 0.0428)

Table 4.4: Evaluation of ECG COVID model performance, using in AUROC, AP, AUPRC, AP, F1-score, etc, expressed in mean (95% confidence interval) percentage.

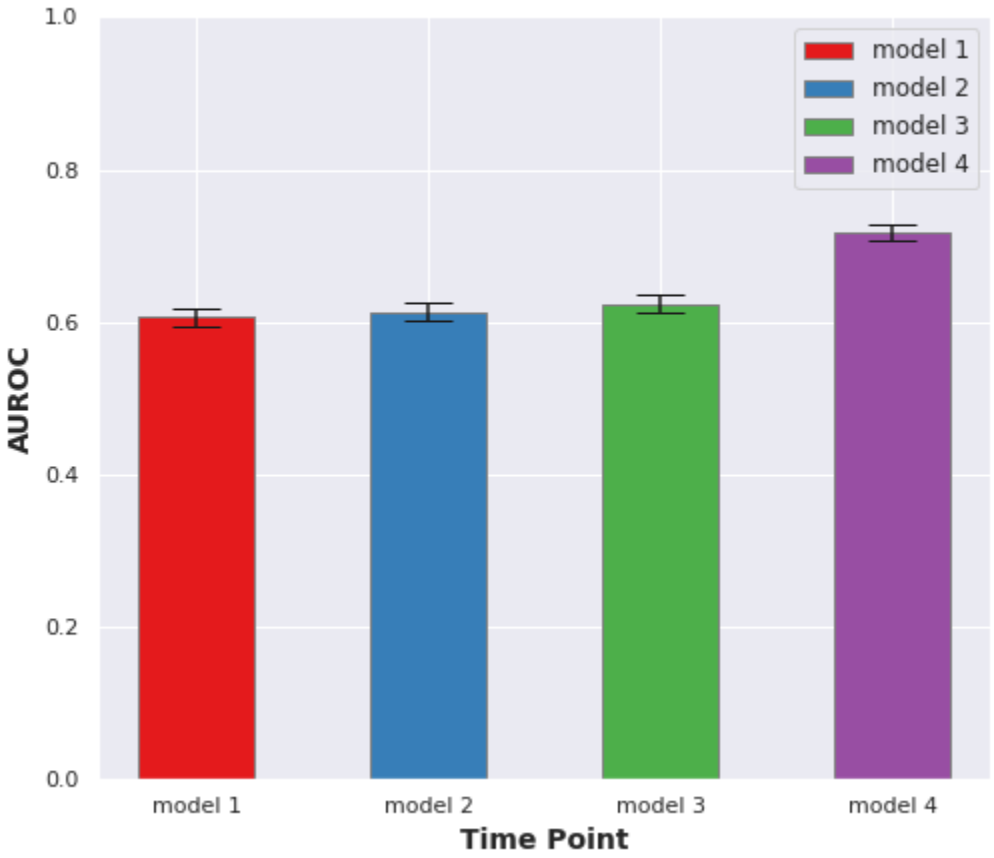

Figure 4.4: Comparison of AUROC performances for ECG COVID-19 models with ECG traces and the error bar is the lower bound and upper bound from 95% bootstrap confidence interval.

most of the COVID-19 patients, who took ECG tests, had severe symptoms.

Chapter 5

ECG Prognosis

This chapter explores the task of learning models for predicting prognosis using ECG data. Chapter 4 showed that we could learn diagnosis models that can accurately predict numerous health conditions based on ECG traces. Motivated by these findings and effectiveness of the patients' health condition to all cause mortality, we implement two different prognostic prediction methods. (1) Section 5.2.1 discusses the binarized mortality classification method, which predicts if a patient dies before a certain time point. (2) Section 5.2.2 introduces the individual survival distribution algorithm, which predicts a patient's survival distribution. Finally, Section 5.3 evaluates and compares the performances of the different prognosis models.

5.1 Method

5.1.1 Analysis Cohort

Table 5.1 describes the characteristics and comorbidities of patient cohorts used in the study. Table 5.2 gives statistics about the ECG measurements of the patients. Section 3.1 explains how we split our dataset into the development set, which is used as training set and tuning set, and holdout set, which forms the final evaluation set. We also make sure there are no patients in both the development set and holdout set. In our dataset, one patient could have multiple ECGs. However, patients with more severe illness are expected to undergo ECGs more frequently, which can cause a bias in the model performance due to the differential representation of patient phenotypes. To mitigate such bias, we evaluate our models using a single randomly-selected ECG per patient in the holdout set; below we refer to this as the "random evaluation set". At the time of the ECG, the average ages of patients in the development and holdout sets are each 65.8 years. However, the average age in the random evaluation set is slightly lower, 62.6 years, because older patients had more ECGs than younger ones, and here each is only counted once. Similarly, men had more ECGs, so the proportion of men is slightly lower in the random evaluation set than in the development set (54.7% vs 56.7%). This pattern is observed for some of the ECG measurements (e.g., mean of QRS duration: 97.9 vs 101.3 ms; QT interval 395.1 vs 399.8ms) and comorbidities (e.g., ECG-wise [1] frequency for Heart Failure: 4.1% vs 6.2%; Atrial Fibrillation 9.2% vs 15.5%).

5.1.2 Prediction Task

In recent years, exponential advancements in computational resources and machine learning technologies, coupled with big digitized ECG datasets, have opened up opportunities for ECG-based prognostic predictions. Here, we used a large cohort of

[1] Here, each instance is an ECG, as opposed to a patient.

	Full Data	Development set	Holdout set	Random ECG per patient in holdout set*
ECG Number	1605268	964741	640527	97631
Age (years)	65.80 ± 17.25	65.77 ± 17.22	65.85 ± 17.29	62.57 ± 18.59
Sex (Male in %)	56.73	56.81	56.6	54.73
Peripheral Vascular Disease	33518 (2.09%)	19714 (2.04%)	13804 (2.16%)	2144 (2.20%)
Cerebrovascular Disease	54349 (3.39%)	33191 (3.44%)	21158 (3.30%)	4252 (4.36%)
Hypertension	350859 (21.86%)	210275 (21.80%)	140584 (21.95%)	15387 (15.76%)
Dementia	133963 (8.35%)	80037 (8.30%)	53926 (8.42%)	8849 (9.06%)
Chronic Pulmonary Disease	31764 (1.98%)	19215 (1.99%)	12549 (1.96%)	2078 (2.13%)
Diabetes Mellitus	120260 (7.49%)	71860 (7.45%)	48400 (7.56%)	5684 (5.82%)
Renal Disease	163262 (10.17%)	96924 (10.05%)	66338 (10.36%)	8800 (9.01%)
Liver Disease	20268 (1.26%)	12062 (1.25%)	8206 (1.28%)	1079 (1.11%)
Cancer	18905 (1.18%)	11707 (1.21%)	7198 (1.12%)	1346 (1.38%)
NSTEMI	93946 (5.85%)	55632 (5.77%)	38314 (5.98%)	8699 (8.91%)
STEMI	162274 (10.11%)	96828 (10.04%)	65446 (10.22%)	6534 (6.69%)
Heart Failure	100206 (6.24%)	60381 (6.26%)	39825 (6.22%)	4049 (4.15%)
Atrial Fibrillation	249325 (15.53%)	150055 (15.55%)	99270 (15.50%)	8958 (9.18%)

Table 5.1: Characteristics of patient cohorts used in the study. For age, we expressed as mean (± standard deviation). For the comorbidities, we expressed as count (percentage).

ECG measurements	Full Data	Development set	Holdout set	Random ECG per patient in holdout set*
Atrial rate	85.60 ± 46.15	85.56 ± 46.11	85.67 ± 46.20	84.06 ± 40.30
P duration	155.92 ± 116.60	156.00 ± 116.39	155.79 ± 116.91	163.96 ± 114.06
RR interval	790.81 ± 213.11	790.90 ± 212.63	790.68 ± 213.83	790.89 ± 204.35
Q wave onset	508.84 ± 6.51	508.82 ± 6.29	508.87 ± 6.82	509.04 ± 6.17
Fridericia Rate-Corrected QT interval	434.86 ± 38.05	434.96 ± 38.04	434.71 ± 38.07	429.55 ± 35.23
Heart Rate	81.64 ± 23.22	81.61 ± 23.18	81.69 ± 23.28	81.13 ± 21.94
PR interval	169.34 ± 38.46	169.44 ± 37.66	169.18 ± 39.65	165.99 ± 33.65
QRS duration	101.36 ± 24.26	101.40 ± 24.23	101.31 ± 24.30	97.89 ± 21.66
QT interval	399.81 ± 54.83	399.94 ± 54.74	399.63 ± 54.96	395.10 ± 51.41
Bazett's Rate-Corrected QT interval	455.02 ± 40.09	455.10 ± 40.09	454.89 ± 40.09	449.23 ± 37.16
Frontal P axis	44.85 ± 35.52	44.81 ± 35.41	44.91 ± 35.69	45.66 ± 32.19
Frontal QRS axis in Initial 40 ms	27.50 ± 46.30	27.44 ± 46.37	27.59 ± 46.20	28.48 ± 42.12
Frontal QRS axis in Terminal 40 ms	45.36 ± 88.15	45.74 ± 88.28	44.80 ± 87.96	46.24 ± 84.64
Frontal QRS axis	19.98 ± 54.37	20.04 ± 54.38	19.88 ± 54.37	22.68 ± 49.81
Frontal ST wave axis	90.94 ± 88.23	90.98 ± 88.07	90.87 ± 88.48	79.23 ± 85.11
Frontal T axis	55.70 ± 67.76	55.48 ± 67.60	56.03 ± 68.00	47.97 ± 59.90
Horizontal P axis	20.69 ± 47.30	20.63 ± 47.14	20.77 ± 47.52	21.01 ± 41.15
Horizontal QRS axis in Initial 40 ms	27.79 ± 48.39	27.86 ± 48.17	27.69 ± 48.71	30.20 ± 42.68
Horizontal QRS axis in Terminal 40 ms	34.10 ± 129.50	33.94 ± 129.40	34.35 ± 129.66	26.67 ± 125.58
Horizontal QRS axis	-0.91 ± 78.19	-1.11 ± 77.91	-0.61 ± 78.62	-4.03 ± 69.27
Horizontal ST wave axis	97.02 ± 64.99	96.98 ± 65.00	97.09 ± 64.97	91.75 ± 60.20
Horizontal T axis	64.46 ± 58.98	64.27 ± 58.96	64.73 ± 59.01	59.30 ± 53.05

Table 5.2: ECG measurements of patient cohorts used in the study expressed as mean (± standard deviation)

universal health insurance patients with emergency department (ED) or hospitalization visits in a defined geographic area to develop DL models based on ECG tracings and XGB models based on ECG measurements to predict short-term (30 days) and long-term mortality (1 year and 5 years) and individual survival prediction starting from the day of ECG acquisition. We calculate the time until death based on the difference between the death date and the ECG acquisition date. The goal of the prognostic prediction model is to provide the calibrated probability of patients' mortality, which could assist the clinical system in (a) evaluating the patients' risk; (b) managing or allocating the clinical resources during patients' stay in the hospital, and (c) planning of patients' subsequent visits after they were discharged from the hospital. Figure 5.1 summarizes the number of ECGs, episodes, and patients used for modeling for each of the 3 tasks in overall data and experimental splits.

5.1.3 Pre-processing Binary Mortality Prediction Data

In our dataset, the study interval is from 2007 to 2020. Here, we consider 3 tasks, asking respectively if a given patient is dead within 30 days (resp., 1 year, and 5 years) after the ECG acquisition date. For each study, we exclude the patients who left the study. Figure 5.2 shows three groups of patients. We will consider that all ECGs, which have time intervals between ECG's acquisition dates and study end less than binary mortality prediction time interval, do not have enough follow-up. For example, our study interval ends at 2020-12-31, and patient ABC's ECG's acquisition date is 2018-12-10. We will consider this ECG does not have enough 5 years follow-up interval.

1. Group A: Patients with enough follow-up time interval. For example, in the 5-year binary mortality task, if they take ECG tests before 2015-01-01 (5 years before study ended), then we consider patients to have enough follow-up time interval.

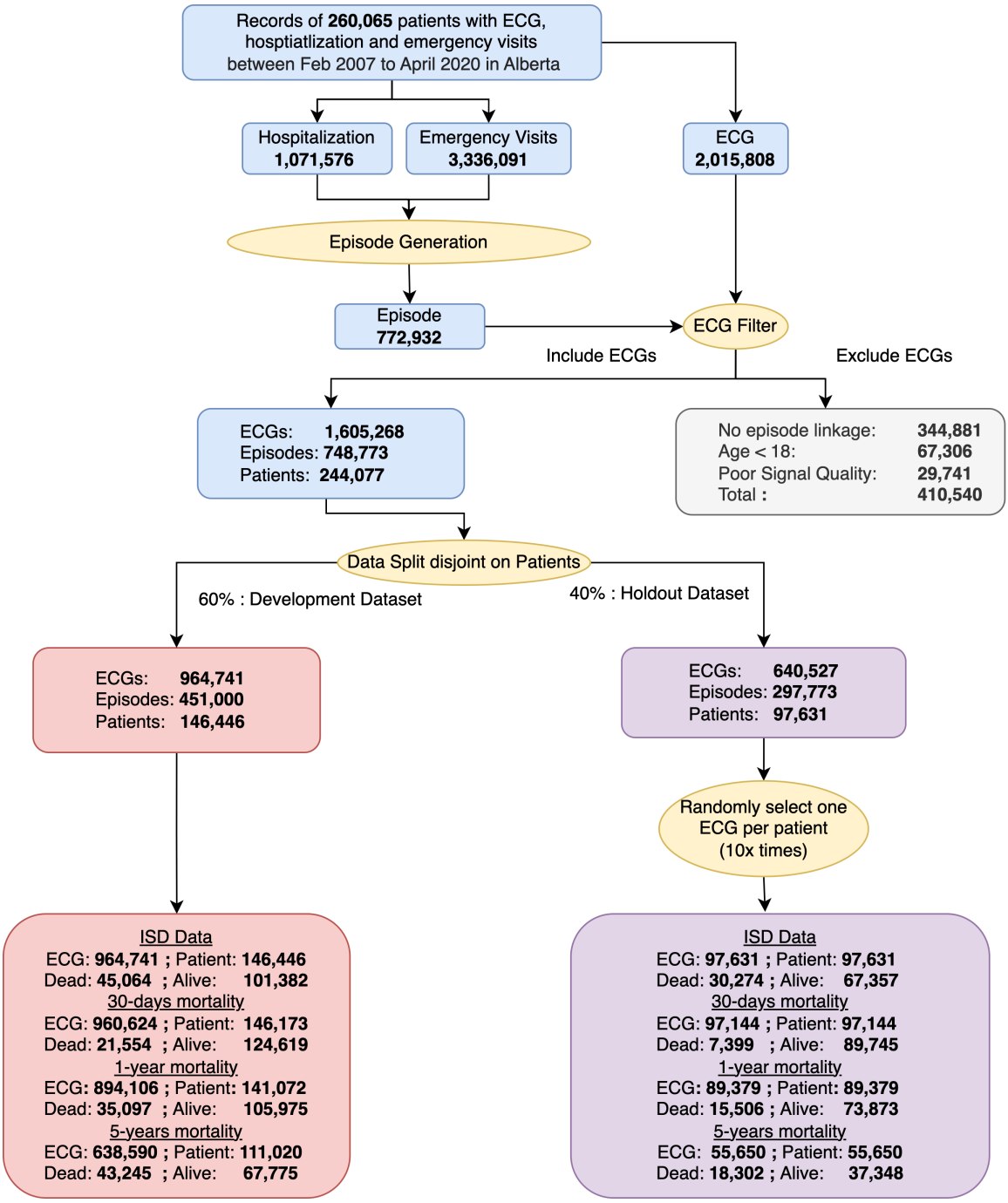

Figure 5.1: Flowchart of the ECG prognosis study design showing the sample sizes for different splits and outcomes

2. Group B: Patients are uncensored, but without enough follow-up intervals, For example, patient Alex had an ECG test on 2018-12-10 and he died on 2019-01-12. Then, he is uncensored but without enough follow-up interval in 5 year

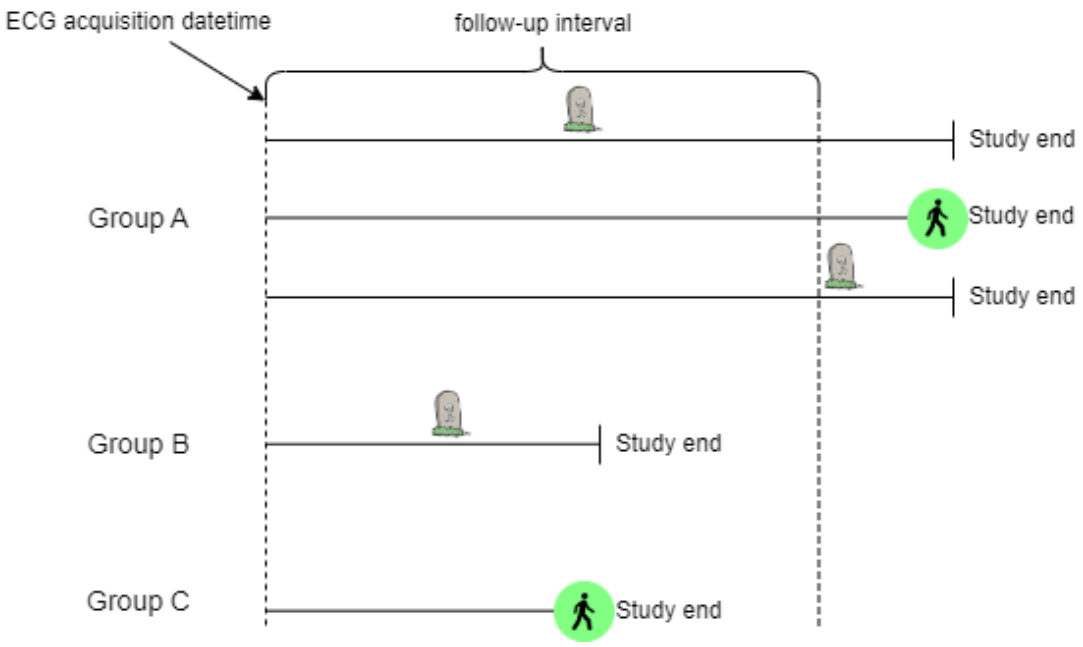

Figure 5.2: Groups of patients in binary mortality prediction task

binary mortality task.

3. Group C: Patients are alive at the study end and they don't have enough follow-up time intervals. For example, patient David had an ECG test on 2018-10-12 and he survived until the study ended. Then, he is censored but without enough follow-up interval in the 5 year binary mortality task.

In the training phase, we use group A and group B patients in the training data to retain the maximum number of instances for training. However, in the evaluation phase, all ECGs without complete follow-up are excluded, irrespective of their death or censoring status. Therefore, we use group A patients only.

5.1.4 Pre-processing for ISD model

In the prognosis task, the binary mortality prediction provides limited information to clinical communities. Since clinicians only know the predicted survival probability at a certain time-point, they might not offer the optimal decision to balance the costs of maintaining routine medical care and an aggressive treatment plan. Therefore,

the time until death and survival probability distribution, which ISD models provide, could inform the clinicians about the risk for patients at each time point and the estimated survival time.

According Section 3.2.3, we convert data to the format $D = \{[\vec{X}_i, T_i, \delta_i]\}_i$, which provides i-th patient's vector features $\vec{X}_i$; a non-negative value T_i which is the time until i-th patient censor or event occur, and $\delta_i \in 0, 1$ indicate if i-th patient is dead, at the time T_i. In our dataset, we have 25.2% of the patients (corresponding to 36.35% of the ECGs) died before the end of study interval. We show the details in the Figure 5.1 and the Kaplan Meier curve in the Figure 5.3.

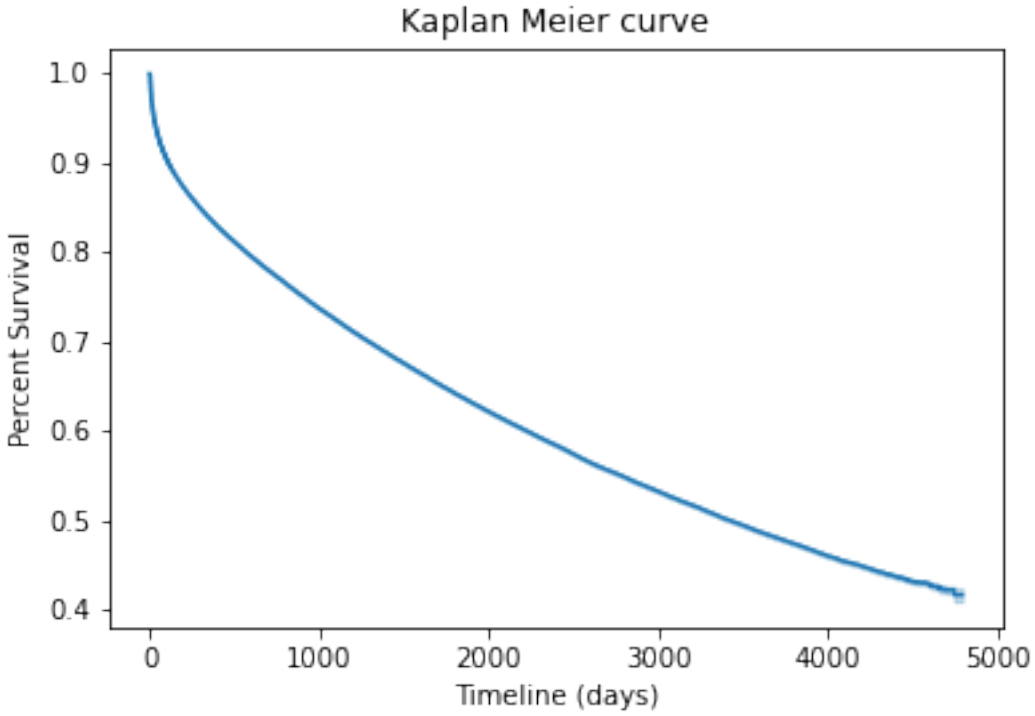

Figure 5.3: Kaplan Meier curve in study dataset, where X-axis is number of days, and Y-axis is percent of survival.

5.2 Learning Algorithm

We use two approaches to detect the mortality risk in the prognosis task. Firstly, we use a binary classification model in binary mortality tasks. To accommodate the generally used ECG formats – ECG measurements and digitized ECG waveform – we use both XGB and ResNet models, respectively. Secondly, we use the individual survival distribution (ISD) method. We also use different ISD model architectures for ECG measurements and digitized ECG waveforms.

5.2.1 Binary Mortality Classification

In this classification task, the input of the learned XGB model is 22 ECG measurements, age, and sex. The XGB hyperparameter and structure have been described in Section 3.2.1. Meanwhile, the input of the learned ResNet model is 12 leads digitized ECG waveform, age, and sex. The hyperparameter and architecture of the ResNet model have been described in 3.2.2.

5.2.2 Individual Survival Distribution

Individual survival analysis corresponds to converting personal data to estimate the time until an event of interest will occur. As shown in Figure 5.4, we design the ISD models separately from ECG measurements and 12-lead digitized ECG waveforms. We use three different ECG feature representations as inputs for our ISD algorithms: (A) We design the end-to-end ISD algorithm where the input is 12 lead ECGs and demographic features, and the output is individual survival distribution. (B) We also develop the two-step model. Here we use the 1,414 diagnoses prediction model from Section 4.2 as the ECG feature extractor [2]. Then, we used the predicted diagnosis probabilities and demographic features to train an ISD model. (C) To accommodate the hand-crafted ECG features, we directly use the ECG measurements, age, and sex

[2]We use 1,414 diagnosis codes instead of 275 top-performing ICD codes because we want comprehensive ECG representations for better training

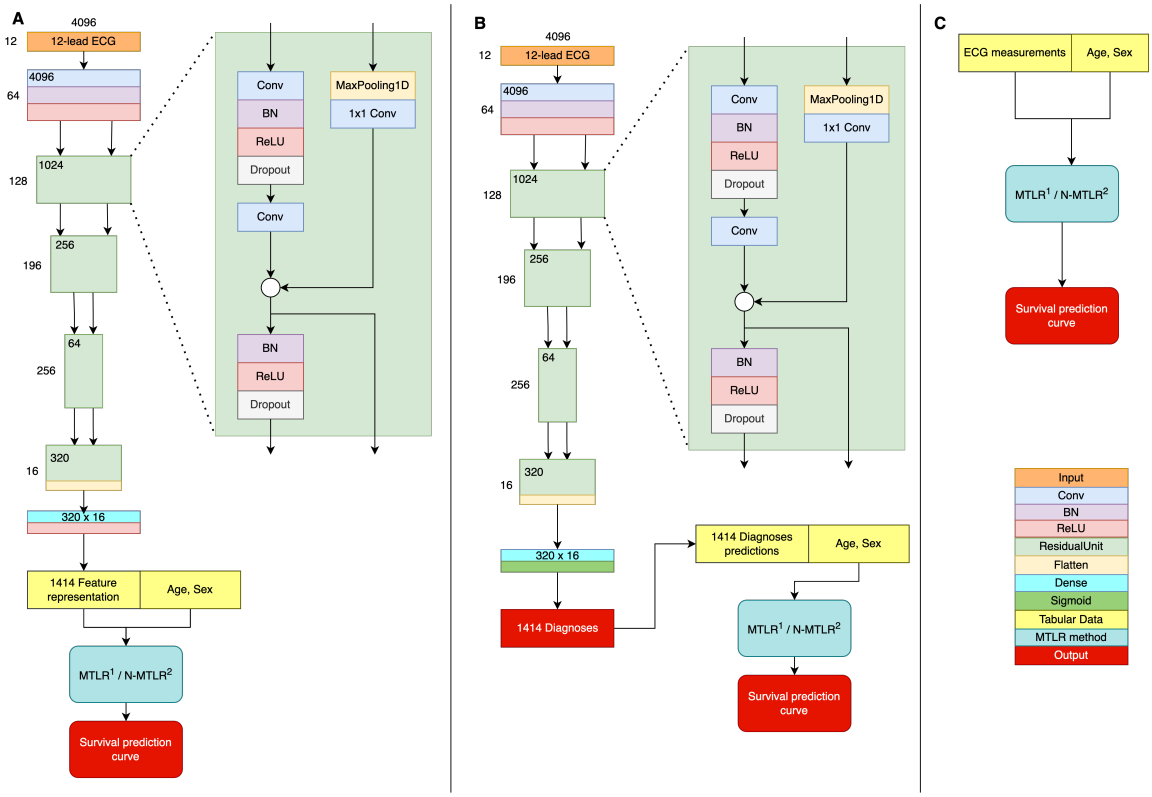

Figure 5.4: Schematic of ISD models. Three ECG feature representation: Model A input is 12 lead ECG waveform and output is 1,414 ECG feature representation; Model B input is 12 lead ECG waveform and output is 1,414 ICD diagnosis prediction values; Model C input is ECG measurements. Two ISD algorithms: 1 is MTLR and 2 is N-MTLR

as the input features to train an ISD algorithm. We use (1) MTLR and (2) N-MTLR, described in Section 3.2.3, as our choice of ISD algorithms for each of the above three scenarios and compared six different ECG ISD models in result Section 5.3.3.

5.3 Result

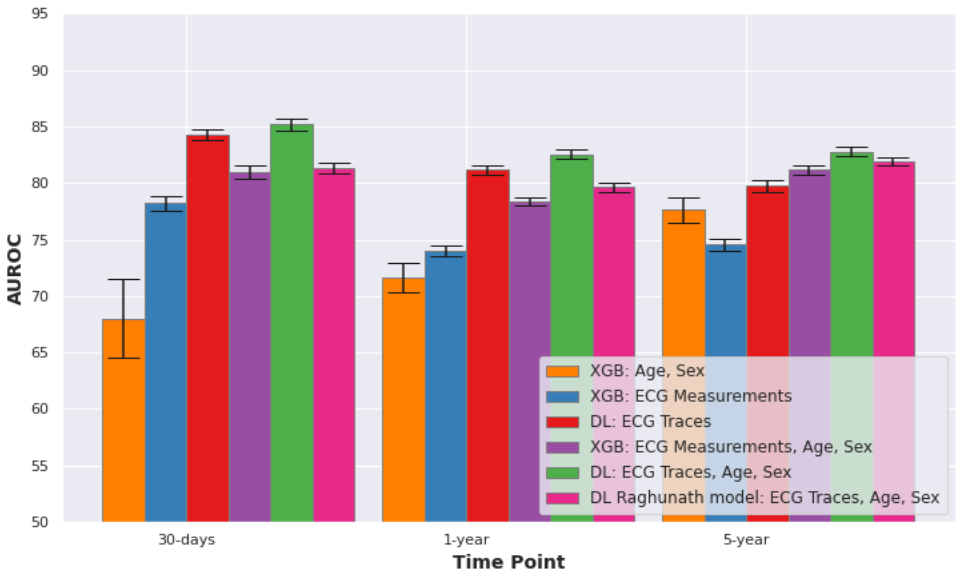

Figure 5.5: Comparison of AUROC model performances for ResNet, XGB and comparable models with ECG traces and measurements

We use the evaluation methods described in Section 3.3 to estimate the model performances. In addition, according to Section 5.1.1, we evaluate our model using a single ECG randomly selected from multiple episodes for each patient in the holdout set. This evaluation sampling strategy is reasonable and representative of deploying the model in a real-world scenario on a recent ECG from a new patient, rather than using the high mortality risk and high frequency ECGs from the same patient.

5.3.1 Binary Mortality Model comparison

Figure 5.5 and 5.6 present the comparisons of models' performances. We use age, and sex features alone to establish a baseline model performance, which had an AUROC of 0.680 for 30-days, 0.716 for 1-year, and 0.776 for 5-year mortality. The ResNet model with ECGs traces alone had a substantially higher performance with AUROC of 0.843, 0.812, and 0.798 for 30-days, 1-year, and 5-year predictions, respectively.

Age, sex, and ECG traces show further small but significant improvements with AUROC of 0.852, 0.826, and 0.828 for the three-time points. ResNet with ECG traces performs significantly better than XGB with ECG measurements for all three time points. ResNet with ECG traces, age, and sex is the best model in this comparison, with AUROCs consistently higher than 82%. For 5-year outcomes, XGB with ECG measurements did not perform better than just age and sex. However, ECG traces still provided relevant information to the prediction. They significantly outperformed the baseline age and sex model, emphasizing the prognostic utility of ECG traces over typically used ECG measurements. Figure 5.6 shows the superior performance of ResNet models with ECG traces in terms of AUROC, AUPRC, F1-Score, and other measures.

Time-point	Features	Model	AUROC	AUPRC	F1 Score	Specificity	Recall	Precision	Accuracy	Brier Score
30-days	Age, Sex	XGB	67.99 (64.54-71.47)	2.95 (2.16-4.06)	3.99 (3.37-4.59)	60.52 (59.86-61.11)	66.38 (60.38-72.32)	2.06 (1.73-2.37)	60.59 (59.93-61.21)	1.55 (1.41-1.68)
	ECG	DL	84.32 (83.83-84.82)	34.54 (33.24-35.81)	34.67 (33.83-35.46)	80.25 (79.96-80.56)	71.9 (70.76-73.08)	22.85 (22.17-23.46)	79.63 (79.33-79.92)	5.97 (5.83-6.12)
		XGB	78.24 (77.59-78.87)	22.07 (21.14-22.96)	30.09 (29.14-31.01)	84.54 (84.21-84.82)	51.38 (49.8-52.87)	21.27 (20.53-22.01)	82.05 (81.74-82.33)	6.44 (6.29-6.61)
	ECG, Age, Sex	DL	85.19 (84.7-85.68)	35.6 (34.31-36.84)	36.25 (35.37-37.09)	81.75 (81.45-82.05)	71.82 (70.67-73.07)	24.24 (23.56-24.93)	81.0 (80.72-81.29)	5.75 (5.62-5.89)
		XGB	81.02 (80.45-81.61)	25.72 (24.57-26.75)	32.66 (31.67-33.67)	84.71 (84.43-84.98)	56.23 (54.81-57.81)	23.02 (22.19-23.83)	82.57 (82.29-82.83)	6.28 (6.12-6.44)
1-year	Age, Sex	XGB	71.59 (70.39-72.89)	22.59 (20.93-24.19)	28.14 (26.89-29.38)	64.62 (63.97-65.32)	67.23 (64.98-69.36)	17.8 (16.89-18.69)	64.89 (64.28-65.55)	9.21 (8.91-9.46)
	ECG	DL	81.2 (80.77-81.61)	48.17 (47.11-49.11)	50.31 (49.62-51.01)	78.86 (78.54-79.18)	67.46 (66.56-68.32)	40.12 (39.45-40.82)	76.88 (76.61-77.17)	11.44 (11.3-11.6)
		XGB	74.02 (73.48-74.53)	35.92 (35.02-36.83)	41.88 (41.16-42.61)	80.88 (80.54-81.22)	50.62 (49.68-51.55)	35.72 (34.98-36.42)	75.63 (75.28-75.99)	12.86 (12.69-13.03)
	ECG, Age, Sex	DL	82.58 (82.18-82.97)	51.21 (50.33-52.22)	52.04 (51.29-52.75)	80.41 (80.1-80.71)	68.0 (66.98-68.94)	42.14 (41.43-42.83)	78.26 (77.97-78.56)	11.21 (11.05-11.37)
		XGB	78.39 (77.98-78.79)	42.03 (41.03-43.05)	46.53 (45.88-47.23)	81.96 (81.66-82.24)	56.39 (55.56-57.28)	39.61 (38.93-40.29)	77.52 (77.21-77.81)	12.19 (12.03-12.36)
5-years	Age, Sex	XGB	77.63 (76.53-78.7)	59.03 (56.73-61.62)	57.32 (55.78-58.98)	72.52 (71.35-73.6)	69.19 (67.34-70.99)	48.93 (47.19-50.89)	71.6 (70.61-72.59)	18.49 (18.13-18.86)
	ECG	DL	79.82 (79.16-80.29)	65.88 (64.89-66.74)	63.14 (62.45-63.78)	78.37 (77.89-78.84)	66.15 (65.26-66.88)	60.4 (59.68-61.17)	74.3 (73.84-74.73)	17.52 (17.29-17.83)
		XGB	74.61 (74.02-75.1)	56.21 (55.15-57.08)	54.78 (53.99-55.56)	81.04 (80.56-81.5)	52.07 (51.22-53.01)	57.78 (56.83-58.57)	71.4 (70.9-71.88)	19.05 (18.85-19.28)
	ECG, Age, Sex	DL	82.8 (82.35-83.24)	70.02 (69.17-70.91)	66.5 (65.84-67.18)	80.18 (79.69-80.63)	69.61 (68.85-70.36)	63.66 (62.86-64.48)	76.67 (76.22-77.1)	16.57 (16.34-16.81)
		XGB	81.18 (80.74-81.58)	67.26 (66.37-68.05)	63.84 (63.22-64.49)	81.68 (81.22-82.12)	64.12 (63.31-64.91)	63.56 (62.8-64.27)	75.84 (75.41-76.21)	16.5 (16.31-16.72)

Figure 5.6: Evaluation of various model performances expressed in mean (95% confidence interval) percentage

Risk Groups: We derive five risk groups - 'very low', 'low', 'medium', 'high', and 'very high' risk groups based on 20 percent cut-points (0 - 20%, 20% - 40%, etc.,) of

predicted probability of death from our main models (ResNet: ECG Trace, age, sex) in the holdout set (Figure 5.8 for 1-year mortality, Figure 5.7 for 30-day and Figure 5.9 5-year mortality). The percentage of observed deaths in each predicted risk group show good calibration with a steady increase across the risk groups (8.6%, 34.6%, 52.3%, 70.9%, and 78.9% death in the 'very low', 'low', 'medium', 'high', and 'very high' risk groups, respectively).

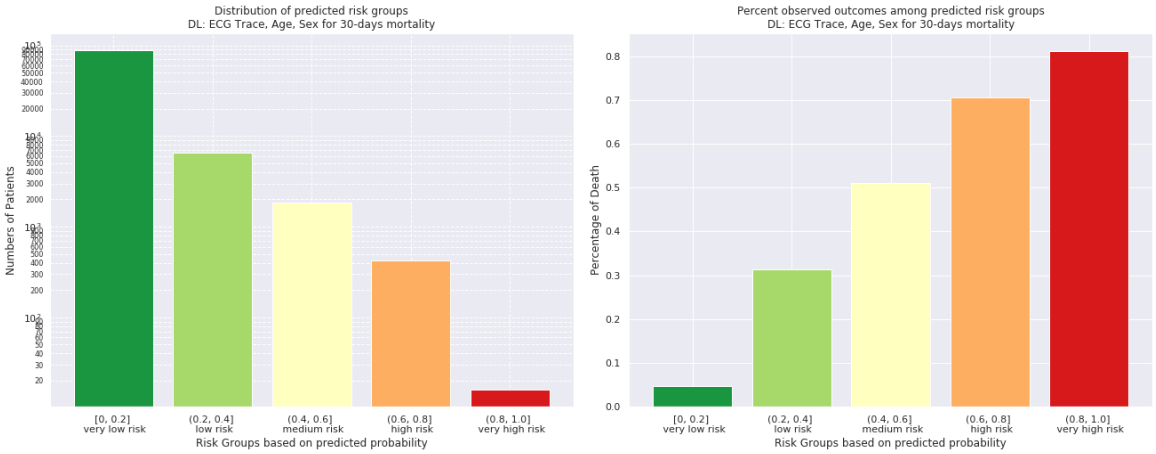

Figure 5.7: Predicted risk groups in the evaluation set for 30-days mortality with ResNet: ECG traces, Age, Sex

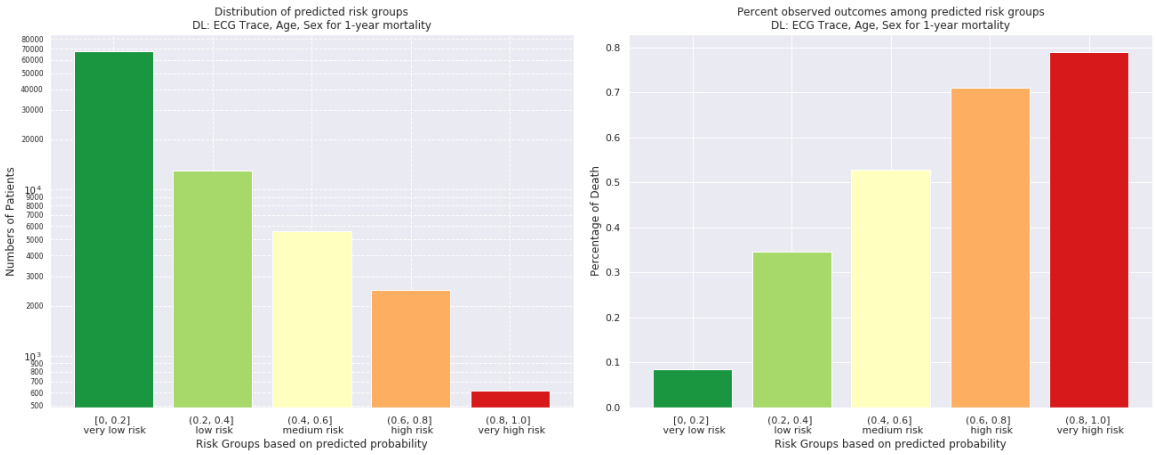

Figure 5.8: Predicted risk groups in the evaluation set for 1-year mortality with ResNet: ECG traces, Age, Sex

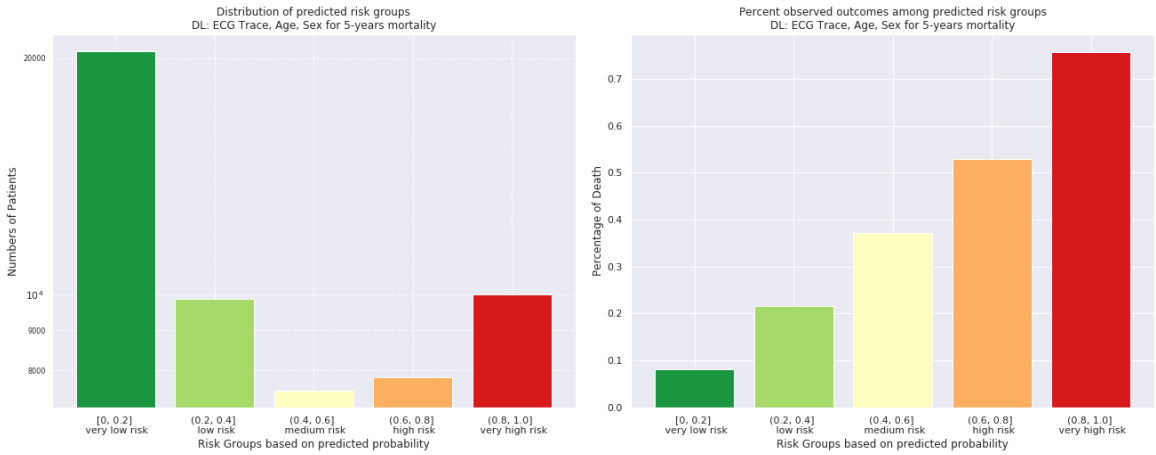

Figure 5.9: Predicted risk groups in the evaluation set for 5-years mortality with ResNet: ECG traces, Age, Sex

Diagnoses	ICD Codes
Non-ST elevation myocardial infarction (NSTEMI)	I214
ST elevation myocardial infarction (STEMI)	I210, I211, I212, I213
Heart Failure	I50, I43, I099, I110, I130, I132, I255, I420, I425, I426, I427, I428, I429, P290
Atrial Fibrillation	I48
Diabetes Mellitus	E10, E11, E12, E13, E14
Hypertension	I10, I11, I12, I13, I15

Table 5.3: ICD 10 codes used for the identifying diagnostic subgroups.

5.3.2 ECG subgroups

We also investigate the learned models' performance for specific subgroups based on patients' sex or disease conditions during the ECG test. We use the following diseases based on the primary diagnosis codes with ICD-10: Non-ST elevation myocardial infarction (NSTEMI), ST-elevation myocardial infarction (STEMI), Heart Failure, Atrial Fibrillation, Diabetes Mellitus, and Hypertension. We present ICD-10 codes for these six diagnostic subgroups in Table 5.3.

Mortality rates differ significantly across the diagnostic groups of interest (Figure 5.10), with patients with heart failure having the highest mortality at each time point.

Meanwhile, most prognostic models perform slightly better in men than in women. Figure 5.11 shows the performance of our models in these different diagnosis subgroups. The models perform better in patients with STEMI and NSTEMI (AUROC of 0.87 and 0.88 for 1-year mortality, respectively) than in the overall cohort. The model's performance in the other subgroups is lower than in the overall holdout cohort. The model performs poorly in heart failure patients (AUROC of 0.75 for 1-year mortality). The results are shown in Figure 5.12.

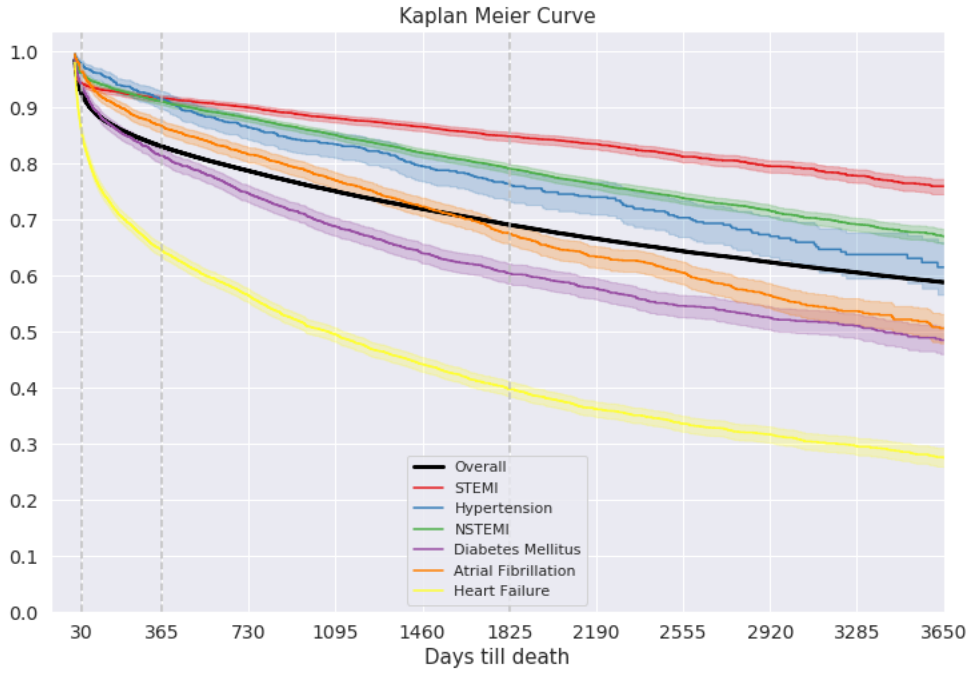

Figure 5.10: Kaplan Meier curves for diagnostic subgroups in the study dataset

5.3.3 ISD models comparison

This section compares the performance of the ISD models using the following evaluation metrics: C-index, hinge L1 loss, marginal L1 loss, and Integral Brier Score (IBS) (described in Section 3.3.2). We consider six ISD models described in Section 5.2.2.

Table 5.4 and C-index bar Figure 5.13 show the superior performance of ISD B2 models in terms of C-index, hinge L1 loss, marginal L1 loss, and Integral Brier Score

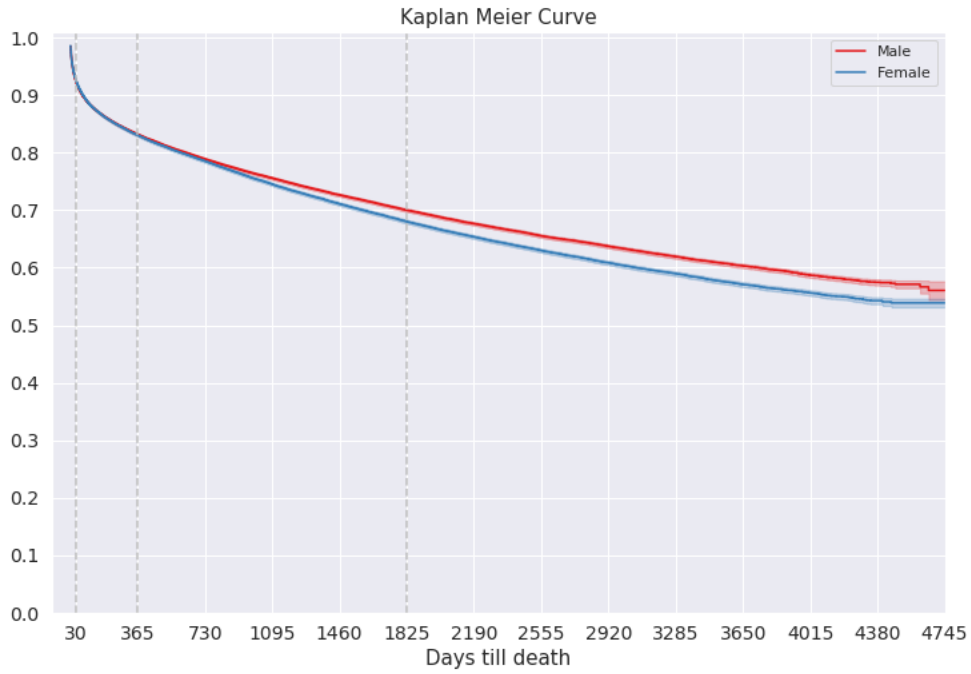

Figure 5.11: Kaplan Meier curves for males and females in the study dataset

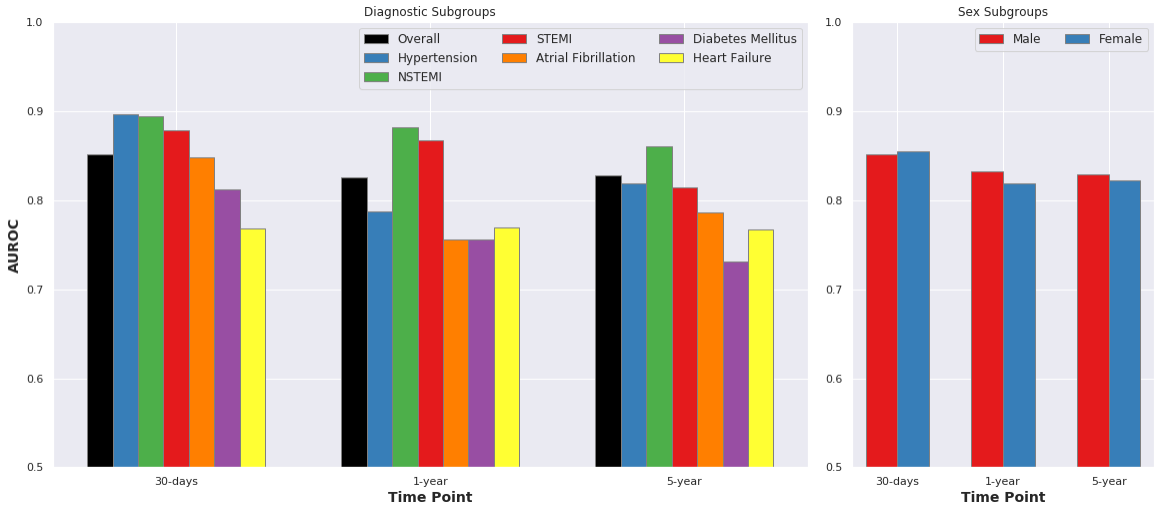

Figure 5.12: AUROC model peformances in primary diagnostic and sex based subpopulations for 1 year mortality with ResNet: ECG traces, Age, Sex.

(IBS). The B2 model (two-step learning with ICD-10 based feature extractor) has a significantly higher C-index of 0.8004 and significantly lower hinge L1 loss of 514.78, marginal L1 loss of 2116.31, and Integral Brier Score (IBS) of 0.14 than the other

	Input	Features	ISD Method	hinge L1 loss	marginal L1 loss	C index	IBS
Model A1	12 lead ECG + Age, Sex	Trace + Age, Sex	MTLR	551.6670 (550.5716 - 552.8586)	2237.7006 (2235.0975 - 2239.9825)	0.7503 (0.7492 - 0.7518)	0.1476 (0.1466 - 0.1487)
Model A2	12 lead ECG + Age, Sex	Trace + Age, Sex	N-MTLR	547.5019 (545.5068 - 549.2004)	2260.5984 (2256.2395 - 2263.8624)	0.7643 (0.7627 - 0.7660)	0.1503 (0.1490 - 0.1518)
Model B1	12 lead ECG + Age, Sex	1414 ICD predictions + Age, Sex	MTLR	518.5863 (516.8676 - 520.5421)	2166.3025 (2162.5818 - 2170.5418)	0.7940 (0.7927 - 0.7947)	0.1419 (0.1406 - 0.1431)
Model B2	12 lead ECG + Age, Sex	1414 ICD predictions + Age, Sex	N-MTLR	514.7825 (513.0864 - 516.7966)	2116.3125 (2112.1232 - 2120.3248)	0.8004 (0.7995 - 0.8011)	0.1368 (0.1355 - 0.1382)
Model C1	ECG measurements + Age, Sex	same as input	MTLR	646.7728 (644.265 - 650.2352)	2325.8173 (2322.8872 - 2330.3700)	0.6785 (0.6776 - 0.6795)	0.2946 (0.2936 - 0.2962)
Model C2	ECG measurements + Age, Sex	same as input	N-MTLR	564.239 (563.0365 - 566.0258)	2304.6902 (2302.1593 - 2307.9460)	0.7589 (0.7576 - 0.7597)	0.1508 (0.1495 - 0.1517)

Table 5.4: Evaluation of ECG ISD models' (described in Section 5.2.2) performance in hinge L1 loss, marginal L1 loss, C-index, and integrated brier score expressed in mean (95% confidence interval) percentage

5 models. The N-MTLR ISD algorithm performs significantly better than MTLR for all model architectures and is particularly pronounced in models C1 and C2. Overall, models C1 and C2 (shallow models without ResNet module) show the lowest performance.

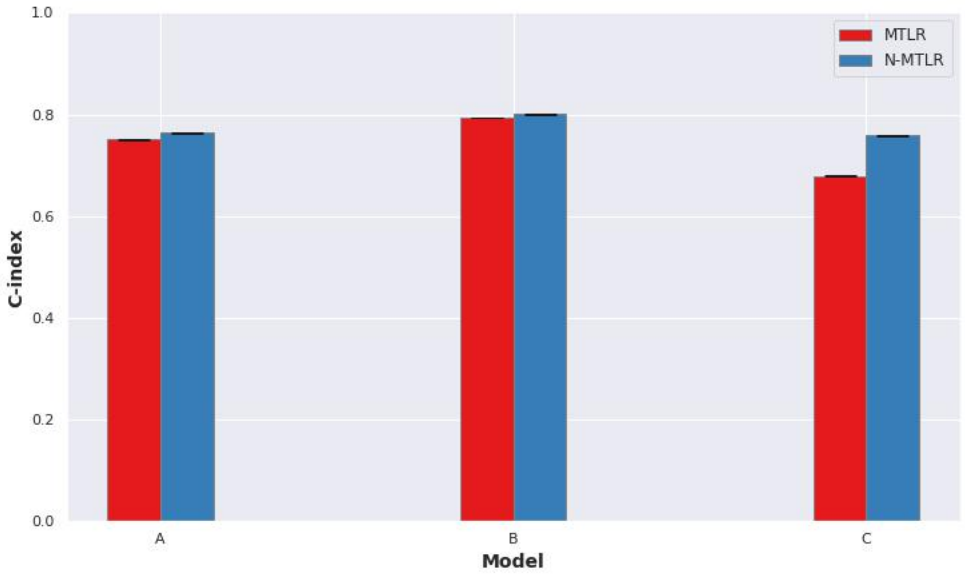

Figure 5.13: Evaluation C-index expressed in mean (95% confidence interval) percentage.

5.3.4 Comparison between binary mortality and ISD models in time points

Table 5.5 lists the AUROC on 30 days (1 year, and 5 years) time points for ISD models and binary mortality models [3]. When the models' inputs are raw ECG traces, age, and sex, the ISD (A1 and A2) end-to-end models perform worse than the binary mortality ResNet model. However, when the models' inputs are ECG measurements, age, and sex, the ISD model (C2) performs better than the binary mortality XGB model. Moreover, for ISD (C1 and C2) models outperform all other models. However, this uses the pretrained weights which is not fair in comparison with binary mortality models.

[3] According to Section 5.1.3, in binary mortality tasks, we have no follow-up or censor data missing in training and evaluation sets.

	30-days AUROC	1-year AUROC	5-years AUROC
A1	0.8303	0.8092	0.8311
A2	0.8241	0.8089	0.8316
B1	0.8626	0.8368	0.8537
B2	0.8619	0.8343	0.849
C1	0.7342	0.728	0.7622
C2	0.8099	0.7897	0.824
12-lead ECG traces, Age, Sex with ResNet	0.8519	0.8258	0.828
ECG measurements, Age, Sex with XGB	0.6799	0.7159	0.7763

Table 5.5: Comparison for all ISD models from Section 5.3.3 and binary mortality models with ECG, age, and sex from Section 5.3.1.

Chapter 6

Discussion and Conclusion

There are two main goals for this thesis. (1) Chapter 4 shows we can learn models that can use a patient's ECG to effectively predict multiple diseases and improve the performance in diagnosing COVID-19 with ECG by using transfer learning technology. (2) Chapter 5 shows that the all-cause binary mortality classification and survival prediction models in ECG prognosis models for all-cause mortality have good performance.

6.1 Future Work

Below we describe three future extensions for our study.

First, we plan to implement more DL models instead of ResNet only because, in machine learning literature, multiple state-of-the-art DL models have shown superior performance in ECG diagnosis of ECG abnormalities. For example, Oh Shu Lih et al. [12] developed the U-Net model that predicted normal sinus beats, atrial premature beats (APB), premature ventricular contractions (PVC), left bundle branch block (LBBB), and right bundle branch block (RBBB) from ECG signals. Peng Xiong et al. [13] used the DenseNet model to diagnose Myocardial Infarction (MI) from 12-lead ECG Traces.

Second, we could measure the generalizability of our models and correct model biases towards certain gender or ethnic groups, because the current study demonstrates

an exciting potential for state-of-the-art DL models trained on a ubiquitous diagnostic test (ECG) linked to routinely collected health data to transform high-throughput diagnostics for a wide range of diseases.

Finally, we might explore other study design methods for prognosis tasks. Then we could compare the performance in binary mortality classification methods and ISD algorithms in short- and long- time points in a fair comparison from the comparable training and evaluation set.

6.2 Diagnosis Discussion and Conclusion

To the best of our knowledge, this is the first study that explores the ECG-based predictability of multiple diseases over the ICD-wide diagnostic landscape. Our DL models, which are trained and validated using population scale datasets, demonstrate excellent AUROC (i.e high sensitivity and specificity) for several diseases; however their precision (PPV) might be limited, partially owing to their low prevalence rates (89.8% of diseases had $< 1\%$ of occurrence) [78]. Therefore, model predictions for such diseases might be more suitable for 'rule out' screening rather than 'rule in' diagnostics. In addition, population-based records enable learning from high-volume healthcare data. However, diagnostic labels obtained from these records may not be considered ground truth without proper adjudication. Like any other supervised machine learning model, the latent ECG features used for prediction in our models may not be directly related to the underlying pathology of diseases. They could be attributed to patients' comorbidities, medication usage, and lifestyle factors that are naturally correlated with disease states in the population.

Labels used in our ECG dataset are ICD-10 based medical diagnosis, whereas most publicly available ECG datasets such as physionet used ECG abnormalities (e.g. Abnormal T Wave) or SNOMED-CT codes (eg: 102594003). Therefore, there is no direct way to evaluate prediction models that we developed, on the external datasets for most of the labels. [79]. However, we compare the performance in

ICD-10 codes associated with cardiovascular diseases (e.g. Atrial Fibrillation, and STEMI), where it is close to the performance in SNOMED-CT in other reports, with AUROC $\geq$ 90% [80, 81]. Finally, although most adult patients get an ECG at some point during their lifetime, there is a potential for selection bias in our cohort as it is restricted to patients who had undergone at least one ECG in the 13 years (2007-2020). Therefore, these results should be considered preliminary proof-of-concept for further investigation of specific diseases by future studies.

Moreover, this ICD-wise diagnosis model (ResNet 1414Dx) not only helps in future studies in ECG abnormality related to non-cardiovascular diseases but also improves the performance in learning ECG models from small samples using transfer learning with pre-train weight. Section 4.3 shows a way to use the ResNet 1414Dx pre-train weight that improves our ECG diagnosis COVID-19 model's performance.

Several other studies [68–74] claimed excellent performance in diagnosing COVID-19 with ECG scanned images. However, we found that all of these studies are based on the same camera-captured ECG image data [75], where the scanner used for COVID-19 patients was different from the scanner used for the other patients. Note this scanner difference might boost the models' performance; see batch effects [82]. We, however, use the 12-lead Digitized voltage-time series ECG waveform (the numerical format of ECG scanned images), which is much less dependent on the scanner used. Our study focuses on raw ECG waveforms that are not available in the publicly available camera-captured ECG image data [75]. (See the ECG formats in Section 2.1.) In the future research, we will convert our ECG waveforms to scanned image format, which allows us to compare the models' performance.

6.3 Prognosis Discussion and Conclusion

Our study is based on a large, population-based cohort of patients with universal access to healthcare. It demonstrates that, for patients at high risk upon their arrival at a hospital, machine learned models could use a patient's ECG data to predict

short- and long-term binary mortality as well as the survival probability for all future time points.

Our binary mortality classification study found that ResNet models based on 12-lead ECG traces perform better than gradient-boosting models (XGB) based on routinely-reported ECG measurements in predicting binary mortality. We also demonstrate that ECG-based ResNet models can be used to identify patients at high risk for short- or long- term mortality. In addition, these models perform equally well in males and females.

We use the ISD models that can generate survival probability for all future time points for each patient, to solve the limitations in binary mortality classification: (1) fully leverage both censored and uncensored data; (2) provide survival probabilities in all time points; (3) resolve the issue that exists patient's survival probability is lower in the short term than the long term. To our knowledge, this is the first study that predicts individual survival curves with ECG data. Section 5.3.3 shows that the 2-step model – beginning with pre-train weight from the ICD-wise diagnosis model (ResNet - 1414Dx) and then using the N-MTLR algorithm – significantly outperforms the other models in terms of C-index, hinge L1 loss, marginal L1 loss, Integral Brier Score (IBS). The 2-step models are better than others due to the pre-train weight of the feature extraction part, which learns the correlation between ECGs and patients' comorbidity and health conditions from a training set.

Bibliography

[1] N. Varma, "Role of the surface electrocardiogram in developing countries," *Journal of electrocardiology*, vol. 43, no. 6, pp. 612–614, 2010.

[2] B. Surawicz and T. Knilans, *Chou's electrocardiography in clinical practice: adult and pediatric*. Elsevier Health Sciences, 2008.

[3] Z. Ebrahimi, M. Loni, M. Daneshtalab, and A. Gharehbaghi, "A review on deep learning methods for ecg arrhythmia classification," *Expert Systems with Applications: X*, vol. 7, p. 100 033, 2020.

[4] G. R. Brämer, "International statistical classification of diseases and related health problems. tenth revision.," *World health statistics quarterly. Rapport trimestriel de statistiques sanitaires mondiales*, vol. 41, no. 1, pp. 32–36, 1988.

[5] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[6] W Amadi and G Kabari, "Analyzing electrocardiograph (ecg) using signal processing technique.,"

[7] D. Kasper, A. Fauci, S. Hauser, D. Longo, J Jameson, and J. Loscalzo, *Harrison's principles of internal medicine, 19e*, 2. Mcgraw-hill New York, NY, USA: 2015, vol. 1.

[8] W. TER and L. HARTFORDHOSPITA, "Your heart's electrical system," *Health*, vol. 860, pp. 545–5000,

[9] E. A. Ashley and J. Niebauer, "Cardiology explained," 2004.

[10] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, 2021.

[11] P. M. Systems, *Philips dxl ecg algorithm physician's guide for ph100b, edition 2*, April, 2009, 2009.

[12] S. L. Oh, E. Y. Ng, R. San Tan, and U. R. Acharya, "Automated beat-wise arrhythmia diagnosis using modified u-net on extended electrocardiographic recordings with heterogeneous arrhythmia types," *Computers in biology and medicine*, vol. 105, pp. 92–101, 2019.

[13] P. Xiong, Y. Xue, J. Zhang, M. Liu, H. Du, H. Zhang, Z. Hou, H. Wang, and X. Liu, "Localization of myocardial infarction with multi-lead ecg based on densenet," *Computer Methods and Programs in Biomedicine*, vol. 203, p. 106 024, 2021.

[14] A. H. Ribeiro, M. H. Ribeiro, G. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira Jr, *et al.*, "Automatic diagnosis of the 12-lead ecg using a deep neural network," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

[15] S. Raghunath, A. E. Ulloa Cerna, L. Jing, D. P. VanMaanen, J. Stough, D. N. Hartzel, J. B. Leader, H. L. Kirchner, M. C. Stumpe, A. Hafez, *et al.*, "Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network," *Nature medicine*, vol. 26, no. 6, pp. 886–891, 2020.

[16] W3Techs, *12-lead ecg placement*, JANUARY 11, 2019, 2019.

[17] A. Lyon, A. Mincholé, J. P. Martínez, P. Laguna, and B. Rodriguez, "Computational techniques for ecg analysis and interpretation in light of their contribution to medical advances," *Journal of The Royal Society Interface*, vol. 15, no. 138, p. 20 170 821, 2018.

[18] E. K. Wang, L. Xi, R. P. Sun, F. Wang, L. Pan, C. Cheng, A. Dimitrakopoulou-Srauss, N. Zhe, and Y. Li, "A new deep learning model for assisted diagnosis on electrocardiogram," *Mathematical Biosciences and Engineering: MBE*, vol. 16, no. 4, pp. 2481–2491, 2019.

[19] R. Avanzato and F. Beritelli, "Automatic ecg diagnosis using convolutional neural network," *Electronics*, vol. 9, no. 6, p. 951, 2020.

[20] E. A. P. Alday, A. Gu, A. J. Shah, C. Robichaux, A.-K. I. Wong, C. Liu, F. Liu, A. B. Rad, A. Elola, S. Seyedi, *et al.*, "Classification of 12-lead ecgs: The physionet/computing in cardiology challenge 2020," *Physiological measurement*, vol. 41, no. 12, p. 124 003, 2020.

[21] S. Somani, A. J. Russak, F. Richter, S. Zhao, A. Vaid, F. Chaudhry, J. K. De Freitas, N. Naik, R. Miotto, G. N. Nadkarni, *et al.*, "Deep learning and the electrocardiogram: Review of the current state-of-the-art," *EP Europace*, vol. 23, no. 8, pp. 1179–1191, 2021.

[22] Y. Wang, X. Zhao, A. O'Neil, A. Turner, X. Liu, and M. Berk, "Altered cardiac autonomic nervous function in depression," *BMC psychiatry*, vol. 13, no. 1, pp. 1–7, 2013.

[23] B. Hage, B. Britton, D. Daniels, K. Heilman, S. W. Porges, and A. Halaris, "Low cardiac vagal tone index by heart rate variability differentiates bipolar from major depression," *The World Journal of Biological Psychiatry*, vol. 20, no. 5, pp. 359–367, 2019.

[24] E. Z. Soliman, R. J. Prineas, M. P. Roediger, D. A. Duprez, F. Boccara, C. Boesecke, C. Stephan, S. Hodder, J. H. Stein, J. D. Lundgren, *et al.*, "Prevalence and prognostic significance of ECG abnormalities in hiv–infected patients: Results from the strategies for management of antiretroviral therapy study," *Journal of electrocardiology*, vol. 44, no. 6, pp. 779–785, 2011.

[25] S. P. Shashikumar, M. D. Stanley, I. Sadiq, Q. Li, A. Holder, G. D. Clifford, and S. Nemati, "Early sepsis detection in critical care patients using multiscale blood pressure and heart rate dynamics," *Journal of electrocardiology*, vol. 50, no. 6, pp. 739–743, 2017.

[26] P. P. Harms, A. A. van der Heijden, F. Rutters, H. L. Tan, J. W. Beulens, G. Nijpels, P. Elders, *et al.*, "Prevalence of ECG abnormalities in people with type 2 diabetes: The Hoorn diabetes care system cohort," *Journal of Diabetes and its Complications*, vol. 35, no. 2, p. 107 810, 2021.

[27] Z. Cheng, K. Zhu, Z. Tian, D. Zhao, Q. Cui, and Q. Fang, "The findings of electrocardiography in patients with cardiac amyloidosis," *Annals of Noninvasive Electrocardiology*, vol. 18, no. 2, pp. 157–162, 2013.

[28] C. Polcwiartek, K. Kragholm, S. M. Hansen, B. D. Atwater, D. J. Friedman, C. A. Barcella, C. Graff, J. B. Nielsen, A. Pietersen, J. Nielsen, *et al.*, "Electrocardiogram characteristics and their association with psychotropic drugs among patients with schizophrenia," *Schizophrenia bulletin*, vol. 46, no. 2, pp. 354–362, 2020.

[29] E. Yahud, G. Paul, M. Rahkovich, L. Vasilenko, Y. Kogan, E. Lev, and A. Laish-Farkash, "Cannabis induced cardiac arrhythmias: A case series," *European Heart Journal-Case Reports*, vol. 4, no. 6, pp. 1–9, 2020.

[30] R. Zulli, F. Nicosia, B. Borroni, C. Agosti, P. Prometti, P. Donati, M. De Vecchi, G. Romanelli, V. Grassi, and A. Padovani, "Qt dispersion and heart rate variability abnormalities in alzheimer's disease and in mild cognitive impairment," *Journal of the American Geriatrics Society*, vol. 53, no. 12, pp. 2135–2139, 2005.

[31] C. A. Pastore, N. Samesima, R. Imada, M. Reis, M. T. Santos, M. C. Ferreira, C. Grupi, F. Fumagalli, J. Wagenfuhr, and M. Chammas, "Characterization of the electrocardiographic pattern of individuals with cerebral palsy," *Journal of electrocardiology*, vol. 44, no. 2, pp. 138–141, 2011.

[32] Q. Wu, L. Han, M. Xu, H. Zhang, B. Ding, and B. Zhu, "Effects of occupational exposure to dust on chest radiograph, pulmonary function, blood pressure and electrocardiogram among coal miners in an eastern province, china," *BMC public health*, vol. 19, no. 1, pp. 1–8, 2019.

[33] M. S. Larssen, K. Steine, J. M. Hilde, I. Skjørten, C. Hodnesdal, K. Liestøl, and K. Gjesdal, "Mechanisms of ECG signs in chronic obstructive pulmonary disease," *Open heart*, vol. 4, no. 1, e000552, 2017.

[34] L. Toma, A. M. Stanciu, A. Zgura, N. Bacalbasa, C. Diaconu, and L. Iliescu, "Electrocardiographic changes in liver cirrhosis—clues for cirrhotic cardiomyopathy," *Medicina*, vol. 56, no. 2, p. 68, 2020.

[35] M Wehr, J Hess, B Noll, and J. Bode, "Cardiac findings in alcoholic liver disease," *Medizinische Klinik (Munich, Germany: 1983)*, vol. 85, no. 11, pp. 629–36, 1990.

[36] S. Shafi, M. Saleem, R. Anjum, W. Abdullah, and T. Shafi, "ECG abnormalities in patients with chronic kidney disease," *Journal of Ayub Medical College Abbottabad*, vol. 29, no. 1, pp. 61–64, 2017.

[37] M Gemelli, F De Luca, R Manganaro, R Leonardi, F Rando, A Agnetti, C Mami, and G Di Pasquale, "Transient electrocardiographic changes suggesting myocardial ischaemia in newborn infants following tocolysis with beta-sympathomimetics," *European journal of pediatrics*, vol. 149, no. 10, pp. 730–733, 1990.

[38] G. Myung, L. J. Forbess, M. L. Ishimori, S. Chugh, D. Wallace, and M. H. Weisman, "Prevalence of resting-ECG abnormalities in systemic lupus erythematosus: A single-center experience," *Clinical rheumatology*, vol. 36, no. 6, pp. 1311–1316, 2017.

[39] G. A. Roth, G. A. Mensah, C. O. Johnson, G. Addolorato, E. Ammirati, L. M. Baddour, N. C. Barengo, A. Z. Beaton, E. J. Benjamin, C. P. Benziger, *et al.*, "Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the gbd 2019 study," *Journal of the American College of Cardiology*, vol. 76, no. 25, pp. 2982–3021, 2020.

[40] J.-m. Kwon, K.-H. Kim, K.-H. Jeon, S. E. Lee, H.-Y. Lee, H.-J. Cho, J. O. Choi, E.-S. Jeon, M.-S. Kim, J.-J. Kim, *et al.*, "Artificial intelligence algorithm for predicting mortality of patients with acute heart failure," *PloS one*, vol. 14, no. 7, e0219302, 2019.

[41] R. Van de Leur, H Bleijendaal, K Taha, T Mast, J. Gho, M Linschoten, B van Rees, M. Henkens, S Heymans, N Sturkenboom, *et al.*, "Electrocardiogram-based mortality prediction in patients with covid-19 using machine learning," *Netherlands Heart Journal*, pp. 1–7, 2022.

[42] J. M. McGinnis, H. V. Fineberg, and V. J. Dzau, "Advancing the learning health system," *The New England Journal of Medicine*, pp. 1–5, 2021.

[43] R. M. Califf, "The benefits of moving quality to a national level," *American heart journal*, vol. 156, no. 6, p. 1019, 2008.

[44] C.-N. Yu, R. Greiner, H.-C. Lin, and V. Baracos, "Learning patient-specific cancer survival distributions as a sequence of dependent regressors," *Advances in neural information processing systems*, vol. 24, 2011.

[45] S. Fotso, "Deep neural networks for survival analysis based on a multi-task framework," *arXiv preprint arXiv:1801.05512*, 2018.

[46] S Ramraj, N. Uzir, R Sunil, and S. Banerjee, "Experimenting xgboost algorithm for prediction and classification of different datasets," *International Journal of Control Theory and Applications*, vol. 9, no. 40, 2016.

[47] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, PMLR, 2015, pp. 448–456.

[48] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[50] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.

[51] G. W. Brier *et al.*, "Verification of forecasts expressed in terms of probability," *Monthly weather review*, vol. 78, no. 1, pp. 1–3, 1950.

[52] M. Zhu, "Recall, precision and average precision," *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, vol. 2, no. 30, p. 6, 2004.

[53] K. Hajian-Tilaki, "Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation," *Caspian journal of internal medicine*, vol. 4, no. 2, p. 627, 2013.

[54] M. Goldstein, X. Han, A. Puli, A. Perotte, and R. Ranganath, "X-cal: Explicit calibration for survival analysis," *Advances in neural information processing systems*, vol. 33, pp. 18 296–18 307, 2020.

[55] Y. Sasaki *et al.*, "The truth of the f-measure," *Teach tutor mater*, vol. 1, no. 5, pp. 1–5, 2007.

[56] H. Haider, B. Hoehn, S. Davis, and R. Greiner, "Effective ways to build and evaluate individual survival distributions.," *J. Mach. Learn. Res.*, vol. 21, no. 85, pp. 1–63, 2020.

[57] J. Wang, J. Sareen, S. Patten, J. Bolton, N. Schmitz, and A. Birney, "A prediction algorithm for first onset of major depression in the general population: Development and validation," *J Epidemiol Community Health*, vol. 68, no. 5, pp. 418–424, 2014.

[58] L. Antolini, P. Boracchi, and E. Biganzoli, "A time-dependent discrimination index for survival data," *Statistics in medicine*, vol. 24, no. 24, pp. 3927–3944, 2005.

[59] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[60] G. W. Brier and R. A. Allen, "Verification of weather forecasts," in *Compendium of meteorology*, Springer, 1951, pp. 841–848.

[61] A. H. Murphy, "Scalar and vector partitions of the probability score: Part i. two-state situation," *Journal of Applied Meteorology (1962-1982)*, pp. 273–282, 1972.

[62] A. H. Murphy, "A new vector partition of the probability score," *Journal of Applied Meteorology and Climatology*, vol. 12, no. 4, pp. 595–600, 1973.

[63] M. H. DeGroot and S. E. Fienberg, "The comparison and evaluation of forecasters," *Journal of the Royal Statistical Society: Series D (The Statistician)*, vol. 32, no. 1-2, pp. 12–22, 1983.

[64] E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher, "Assessment and comparison of prognostic classification schemes for survival data," *Statistics in medicine*, vol. 18, no. 17-18, pp. 2529–2545, 1999.

[65] W. H. Organization *et al.*, "Covid-19 weekly epidemiological update, edition 99, 6 july 2022," 2022.

[66] R. M. Inciardi, L. Lupi, G. Zaccone, L. Italia, M. Raffo, D. Tomasoni, D. S. Cani, M. Cerini, D. Farina, E. Gavazzi, *et al.*, "Cardiac involvement in a patient with coronavirus disease 2019 (covid-19)," *JAMA cardiology*, vol. 5, no. 7, pp. 819–824, 2020.

[67] D. Kaliyaperumal, K. Bhargavi, K. Ramaraju, K. S. Nair, S. Ramalingam, and M. Alagesan, "Electrocardiographic changes in covid-19 patients: A hospital-based descriptive study," *Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine*, vol. 26, no. 1, p. 43, 2022.

[68] O. Attallah, "An intelligent ecg-based tool for diagnosing covid-19 via ensemble deep learning techniques," *Biosensors*, vol. 12, no. 5, p. 299, 2022.

[69] E. Irmak, "Covid-19 disease diagnosis from paper-based ecg trace image data using a novel convolutional neural network model," *Physical and Engineering Sciences in Medicine*, vol. 45, no. 1, pp. 167–179, 2022.

[70] T. Rahman, A. Akinbi, M. E. Chowdhury, T. A. Rashid, A. Şengür, A. Khandakar, K. R. Islam, and A. M. Ismael, "Cov-ecgnet: Covid-19 detection using ecg trace images with deep convolutional neural network," *Health Information Science and Systems*, vol. 10, no. 1, pp. 1–16, 2022.

[71] O. Attallah, "Ecg-biconet: An ecg-based pipeline for covid-19 diagnosis using bi-layers of deep features integration," *Computers in biology and medicine*, vol. 142, p. 105 210, 2022.

[72] J. C. Gomes, M. A. de Santana, A. I. Masood, C. L. de Lima, and W. P. dos Santos, "Covid-19's influence on cardiac function: A machine learning perspective on ecg analysis," 2022.

[73] M. M. Bassiouni, I. Hegazy, N. Rizk, E.-S. A. El-Dahshan, and A. M. Salem, "Automated detection of covid-19 using deep learning approaches with paper-based ecg reports," *Circuits, Systems, and Signal Processing*, pp. 1–43, 2022.

[74] N. Sobahi, A. Sengur, R.-S. Tan, and U. R. Acharya, "Attention-based 3d cnn with residual connections for efficient ecg-based covid-19 detection," *Computers in Biology and Medicine*, vol. 143, p. 105 335, 2022.

[75] A. H. Khan, M. Hussain, and M. K. Malik, "Ecg images dataset of cardiac and covid-19 patients," *Data in Brief*, vol. 34, p. 106 762, 2021.

[76] A. Magadum and R. Kishore, "Cardiovascular manifestations of covid-19 infection," *Cells*, vol. 9, no. 11, p. 2508, 2020.

[77] N. D. Maleki, A. E. Afshar, and P. W. Armstrong, "Use of electrocardiogram indices of myocardial ischemia for risk stratification and decision making of reperfusion strategies," *Journal of Electrocardiology*, vol. 47, no. 4, pp. 520–524, 2014.

[78] A. W. Wong, "Classification and analysis of 12-lead electrocardiograms," 2021.

[79] M. A. Reyna, N. Sadr, E. A. P. Alday, A. Gu, A. J. Shah, C. Robichaux, A. B. Rad, A. Elola, S. Seyedi, S. Ansari, *et al.*, "Will two do? varying dimensions in electrocardiography: The physionet/computing in cardiology challenge 2021," in *2021 Computing in Cardiology (CinC)*, IEEE, vol. 48, 2021, pp. 1–4.

[80] D. U. Jeong and K. M. Lim, "Convolutional neural network for classification of eight types of arrhythmia using 2d time–frequency feature map from standard 12-lead electrocardiogram," *Scientific reports*, vol. 11, no. 1, pp. 1–9, 2021.

[81] M. Bodini, M. W. Rivolta, and R. Sassi, "Classification of 12-lead ecg with an ensemble machine learning approach," in *2020 Computing in Cardiology*, IEEE, 2020, pp. 1–4.

[82] W. W. B. Goh, C. H. Yong, and L. Wong, "Are batch effects still relevant in the age of big data?" *Trends in Biotechnology*, 2022.

Appendix A: Appendix A: AUROC plots for list of categories with top performing Diagnosis ICD-codes tasks

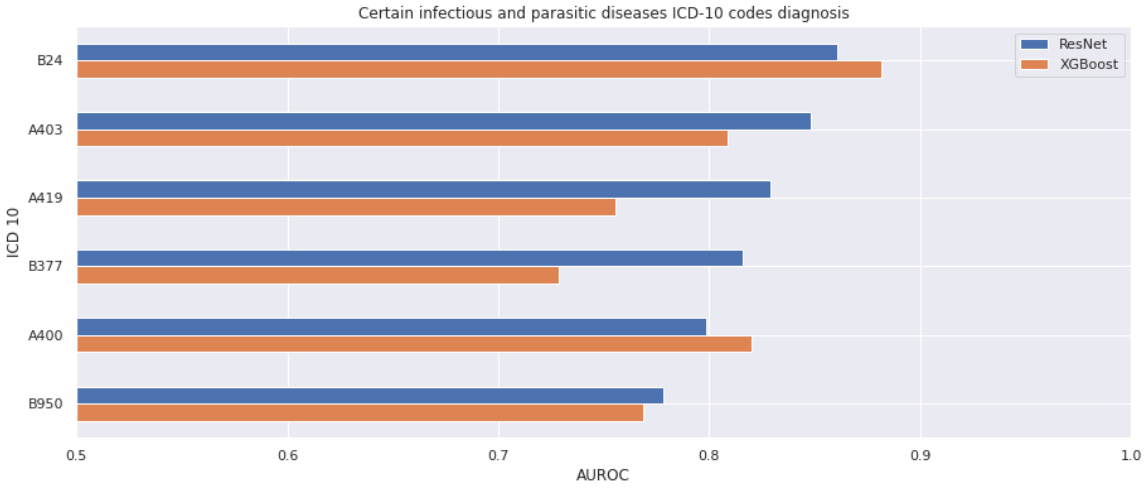

Figure A.1: Certain infectious and parasitic diseases AUROC plot

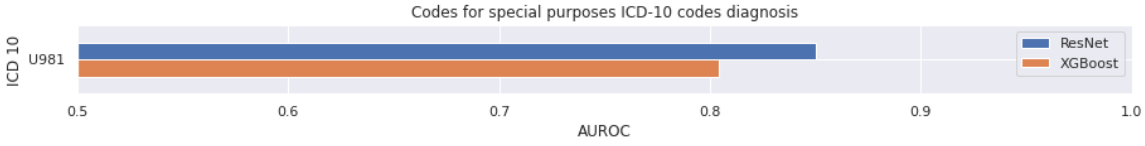

Figure A.2: Codes for special purposes AUROC plot

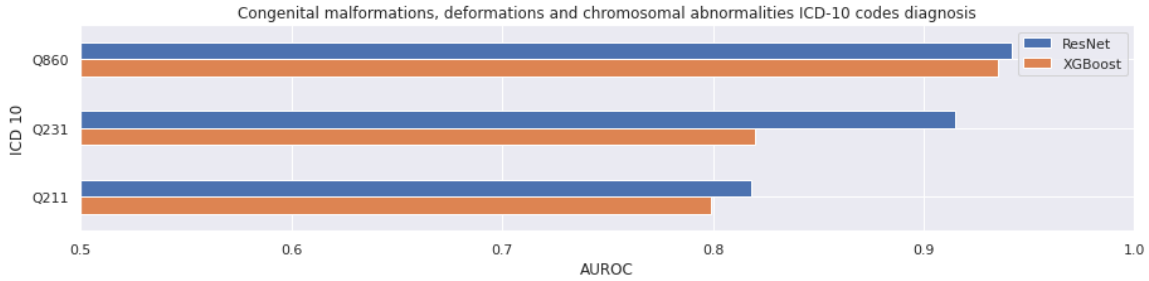

Figure A.3: Congenital malformations, deformations and chromosomal abnormalities AUROC plot

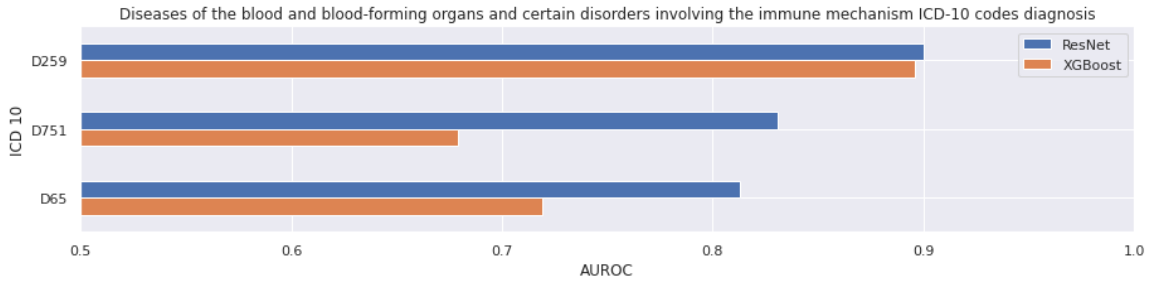

Figure A.4: Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism AUROC plot

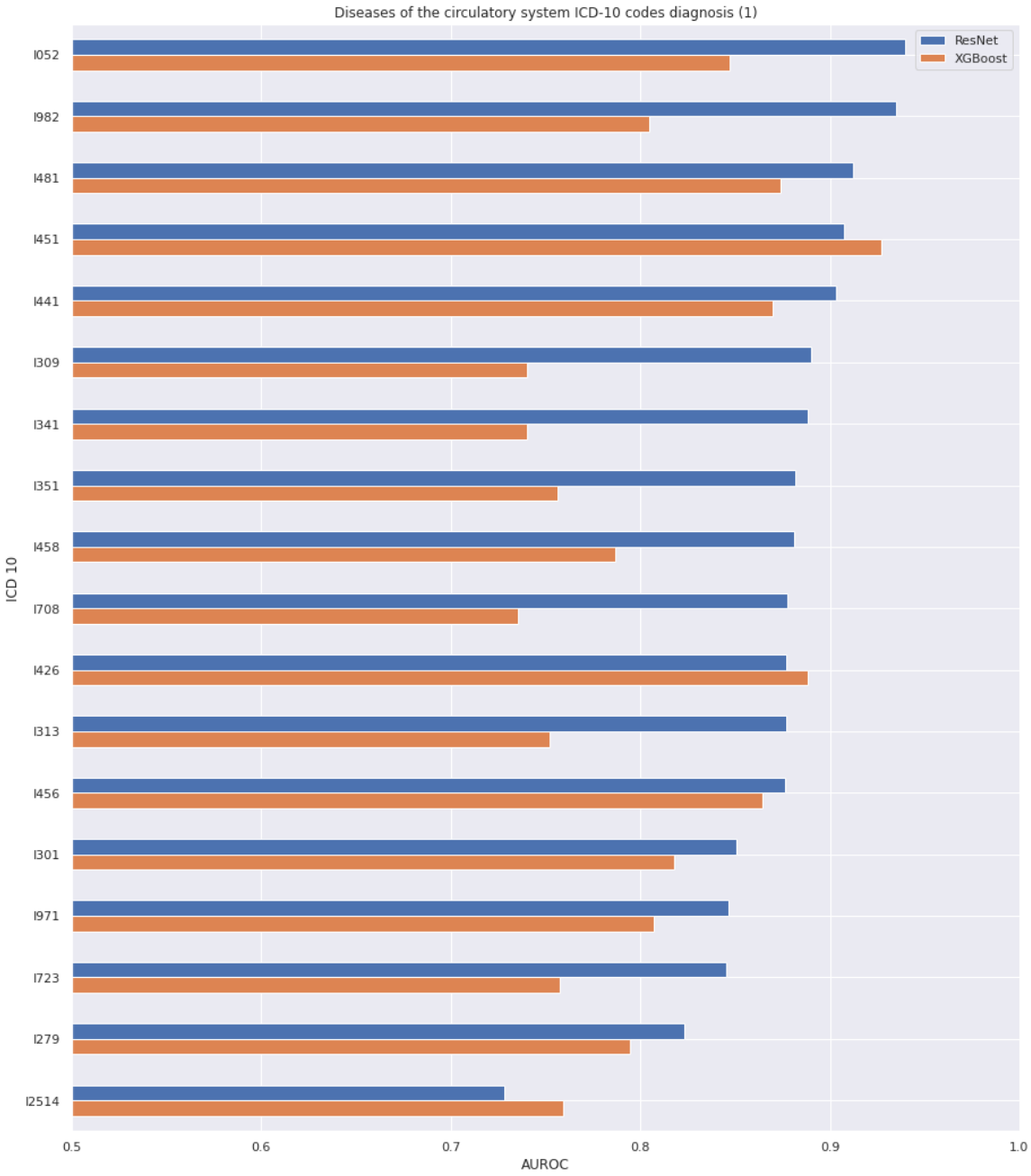

Figure A.5: Diseases of the circulatory system AUROC (1) plot

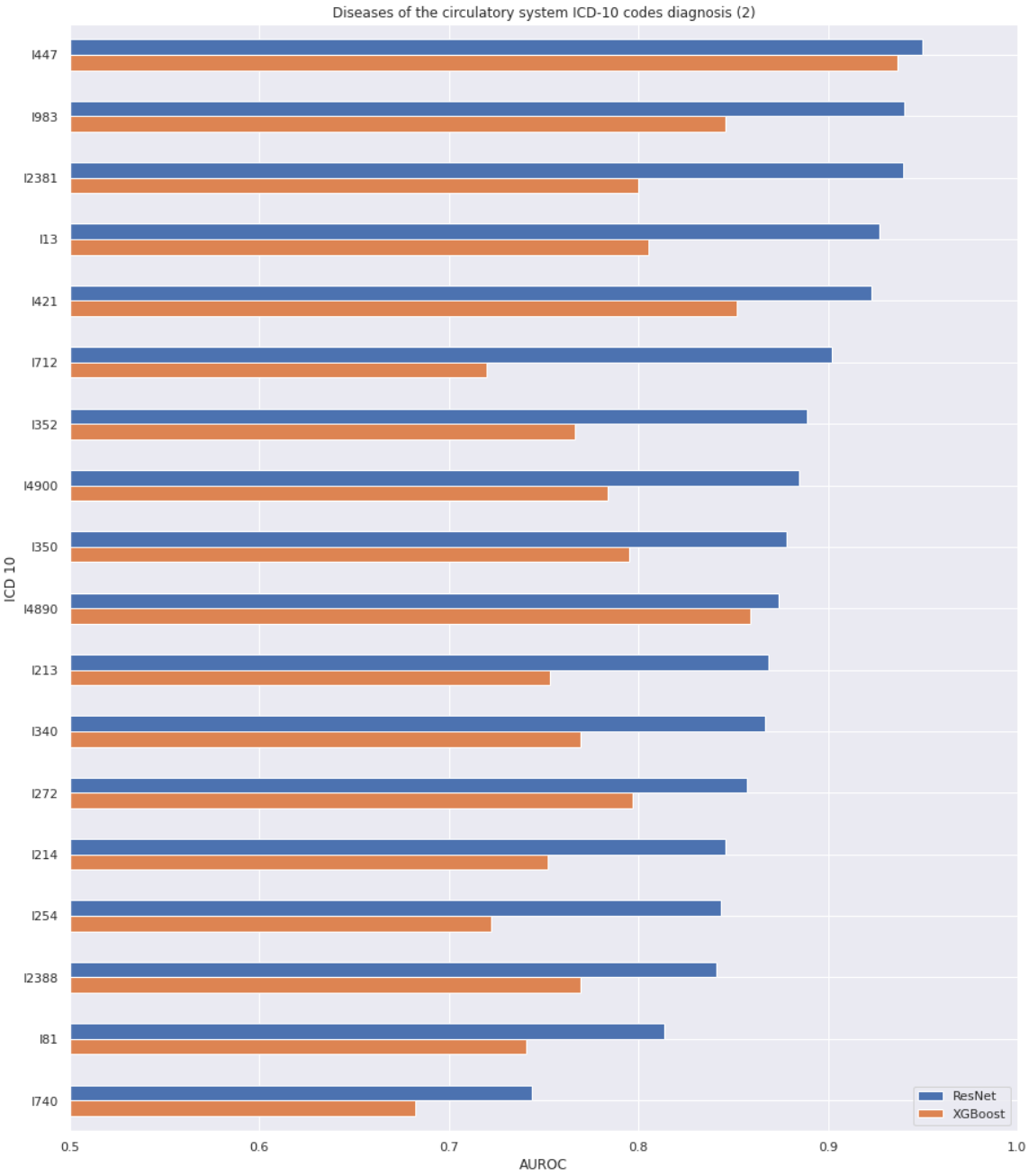

Figure A.6: Diseases of the circulatory system AUROC (2) plot

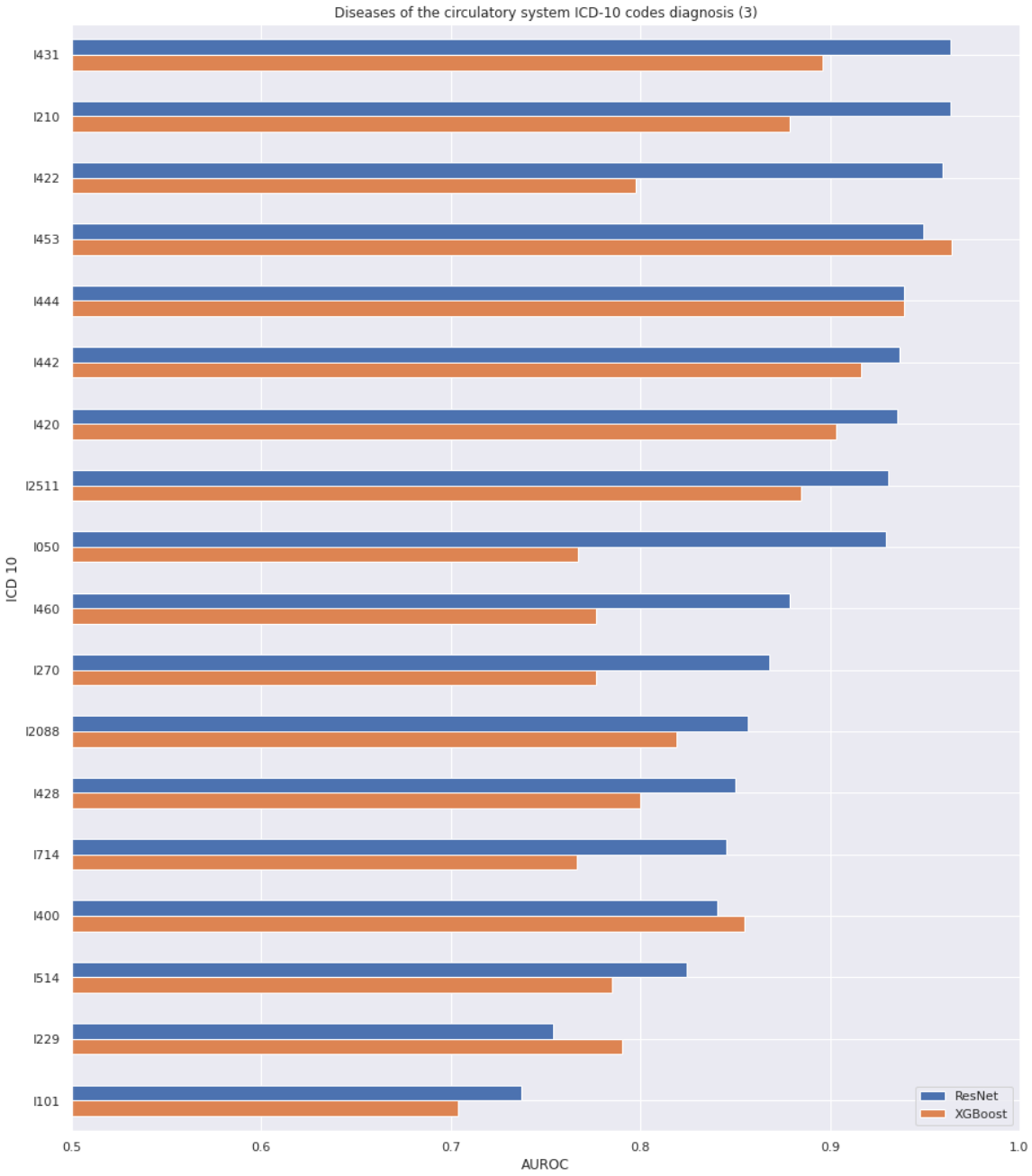

Figure A.7: Diseases of the circulatory system AUROC (3) plot

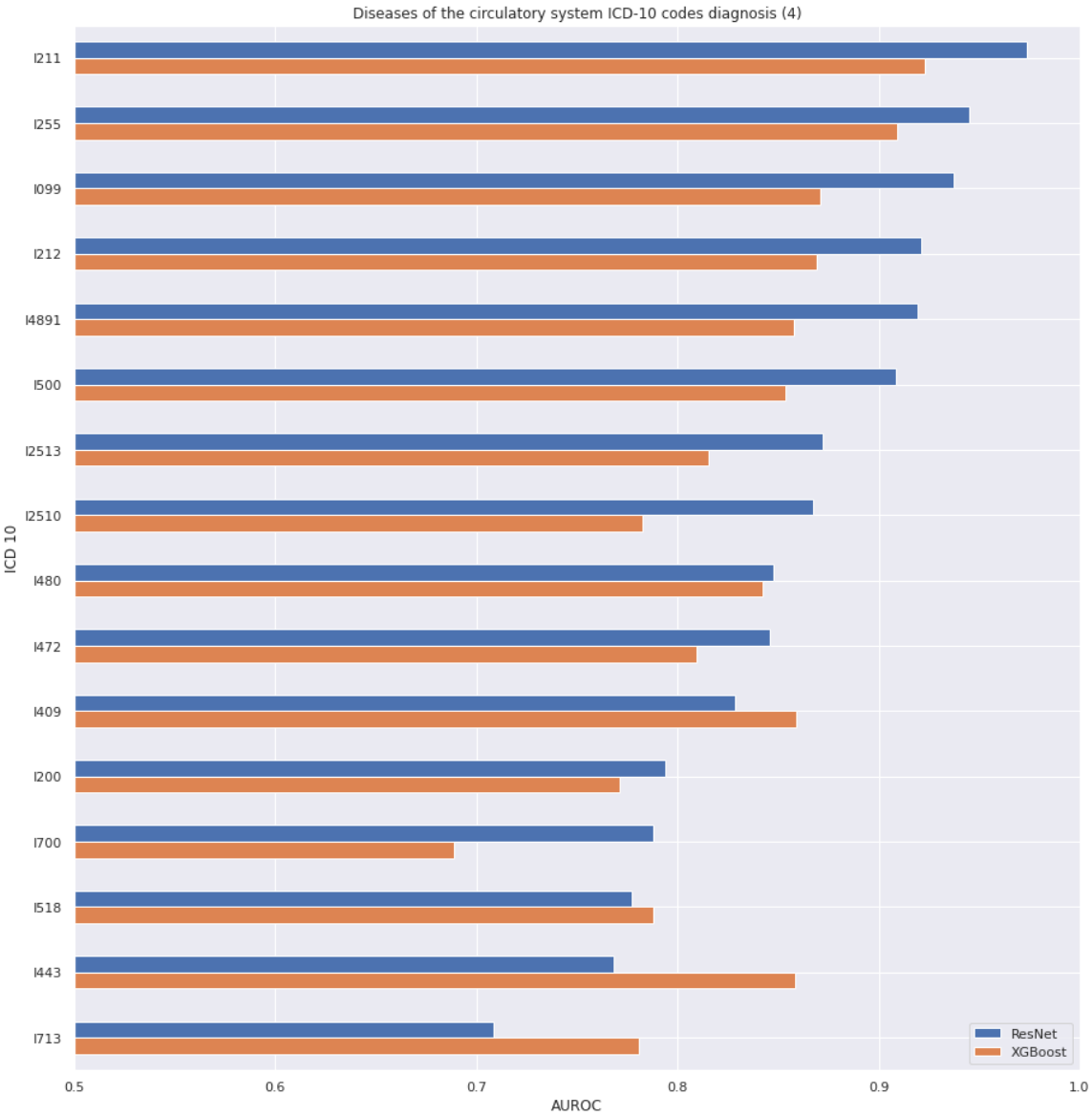

Figure A.8: Diseases of the circulatory system AUROC (4) plot

Figure A.9: Diseases of the digestive system AUROC plot

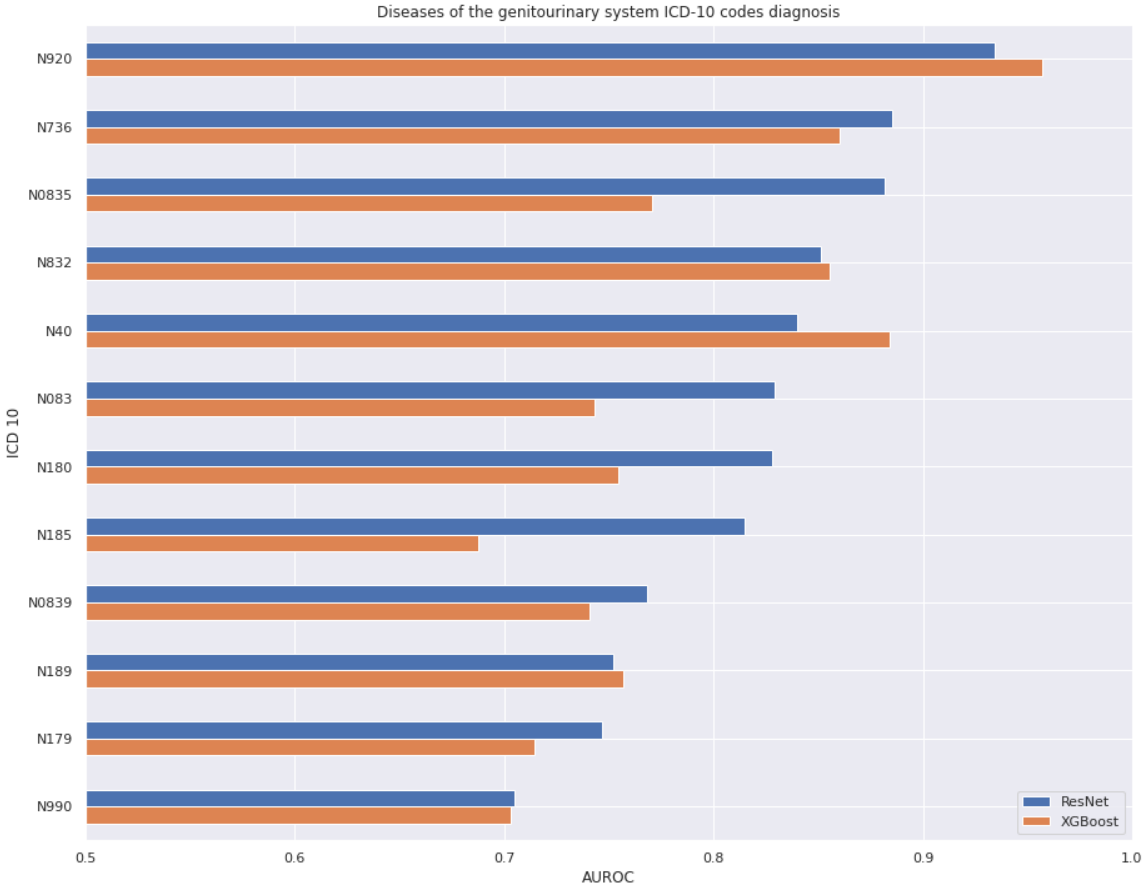

Figure A.10: Diseases of the genitourinary system AUROC plot

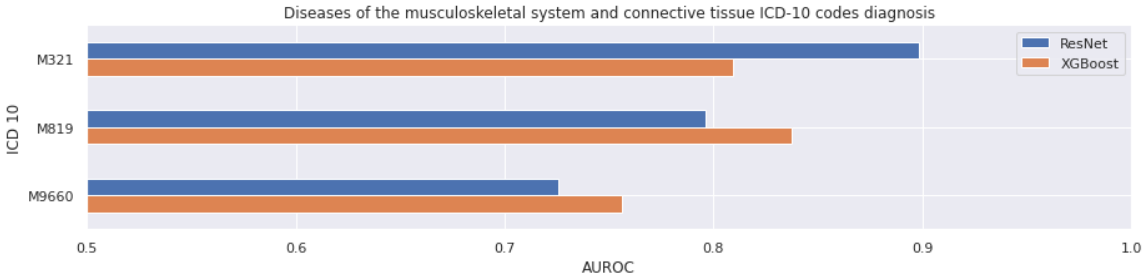

Figure A.11: Diseases of the musculoskeletal system and connective tissue AUROC plot

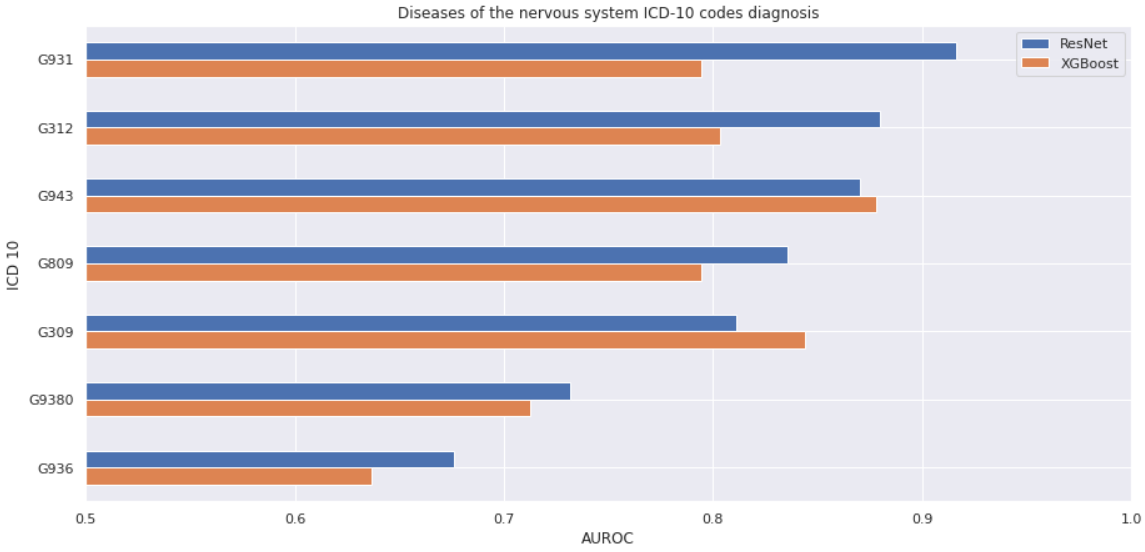

Figure A.12: Diseases of the nervous system AUROC plot

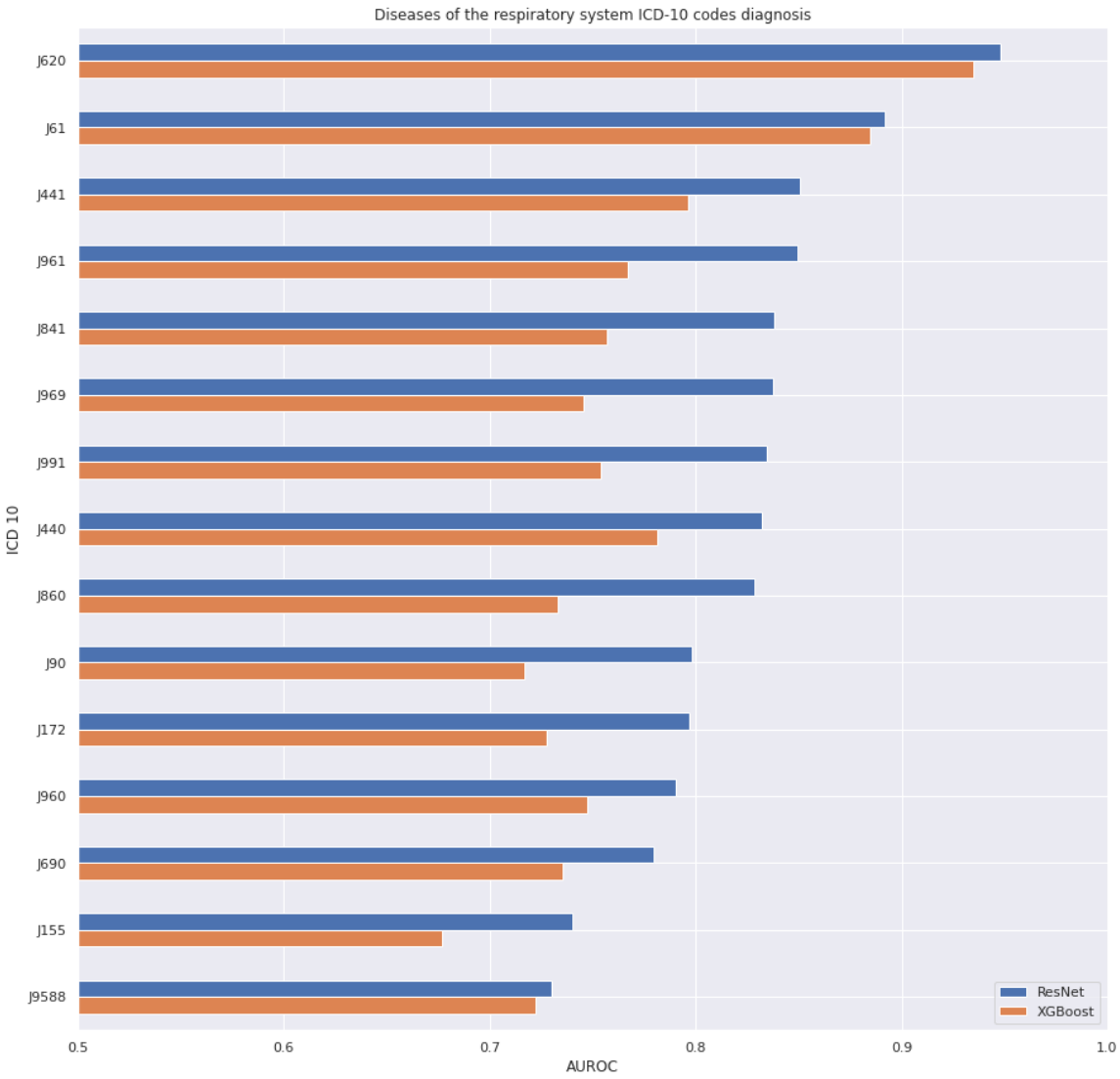

Figure A.13: Diseases of the respiratory system AUROC plot

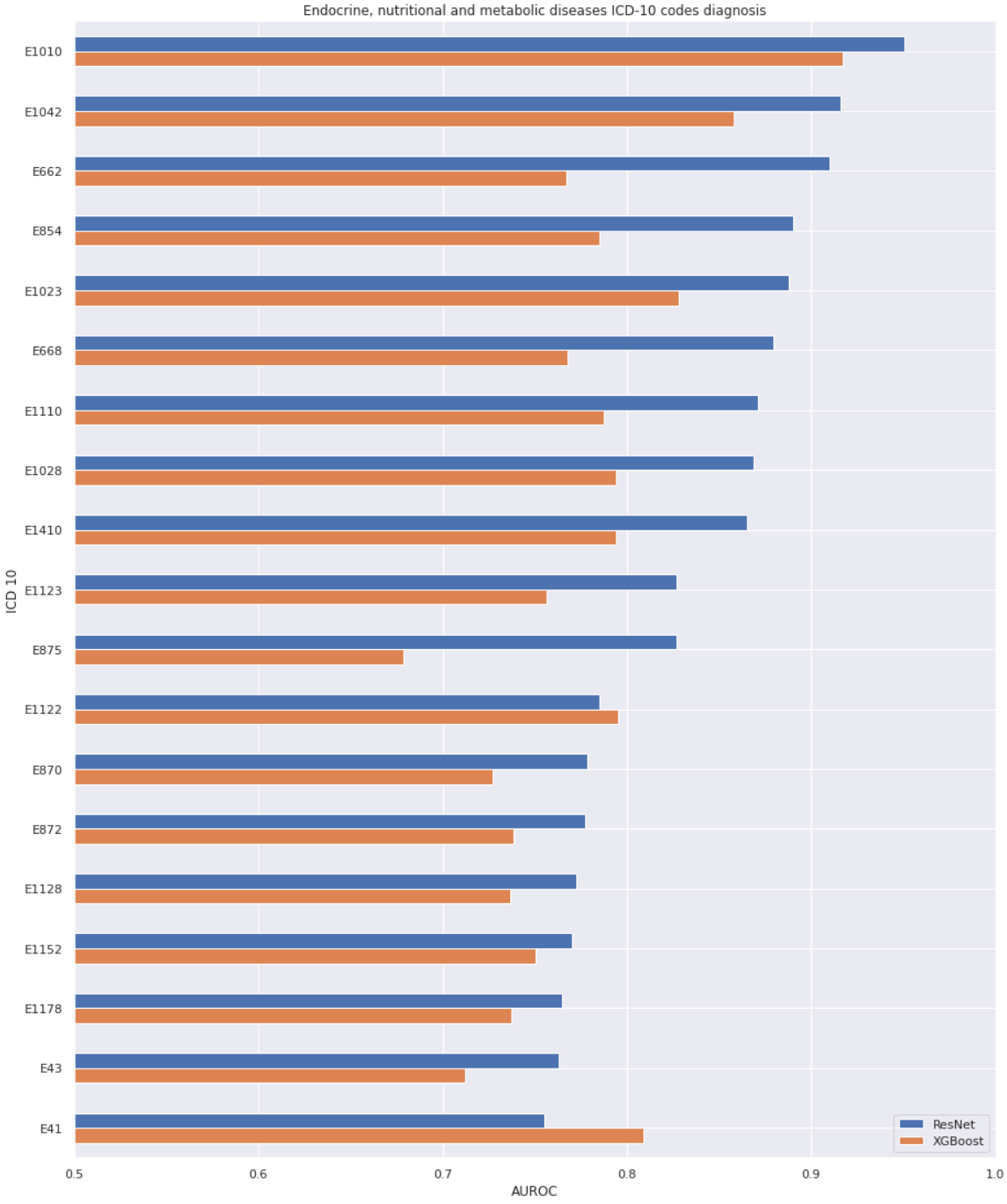

Figure A.14: Endocrine, nutritional and metabolic diseases AUROC plot

Figure A.15: External causes of morbidity and mortality AUROC plot

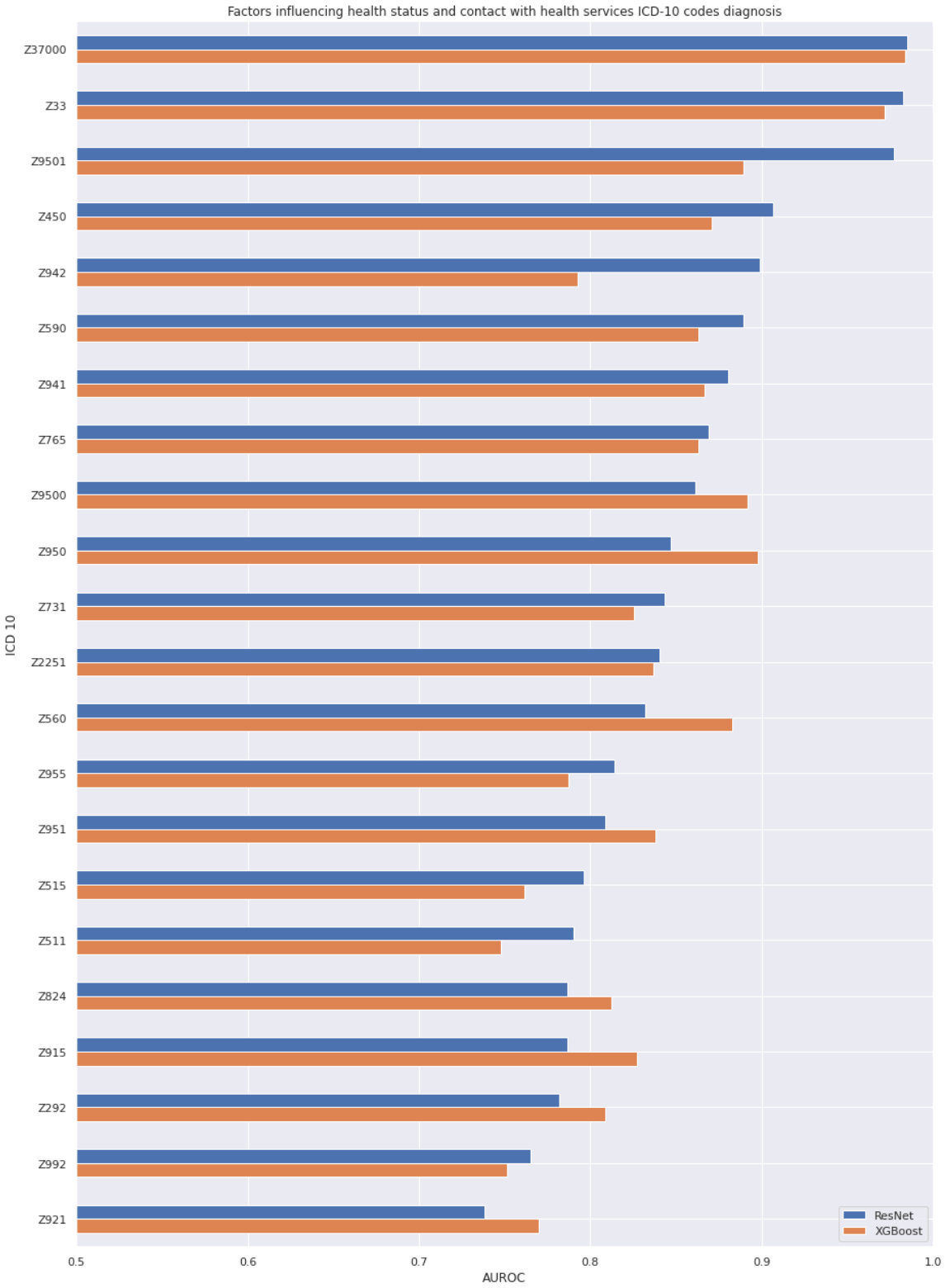

Figure A.16: Factors influencing health status and contact with health services AUROC plot

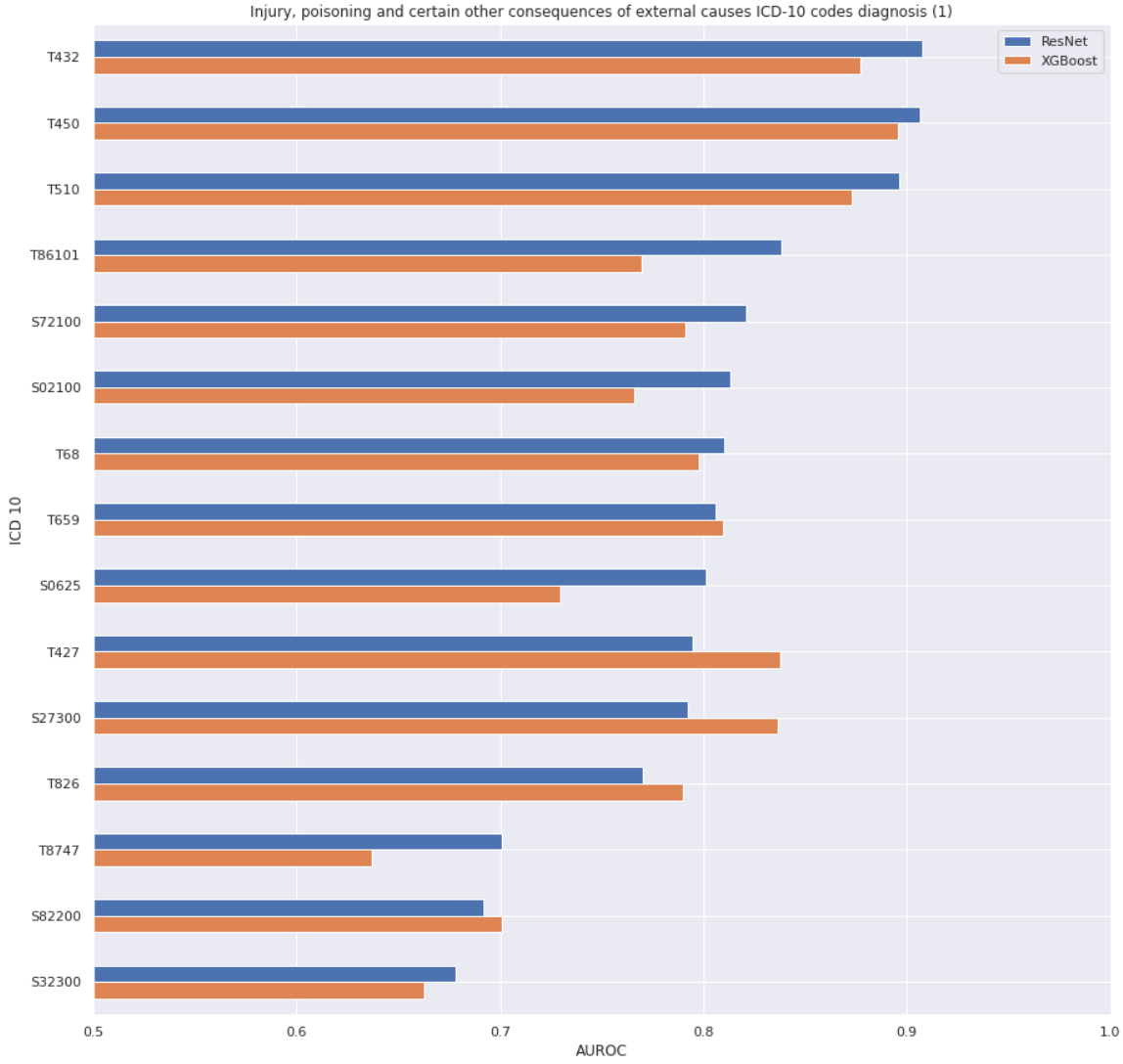

Figure A.17: Injury, poisoning and certain other consequences of external causes (1) AUROC plot

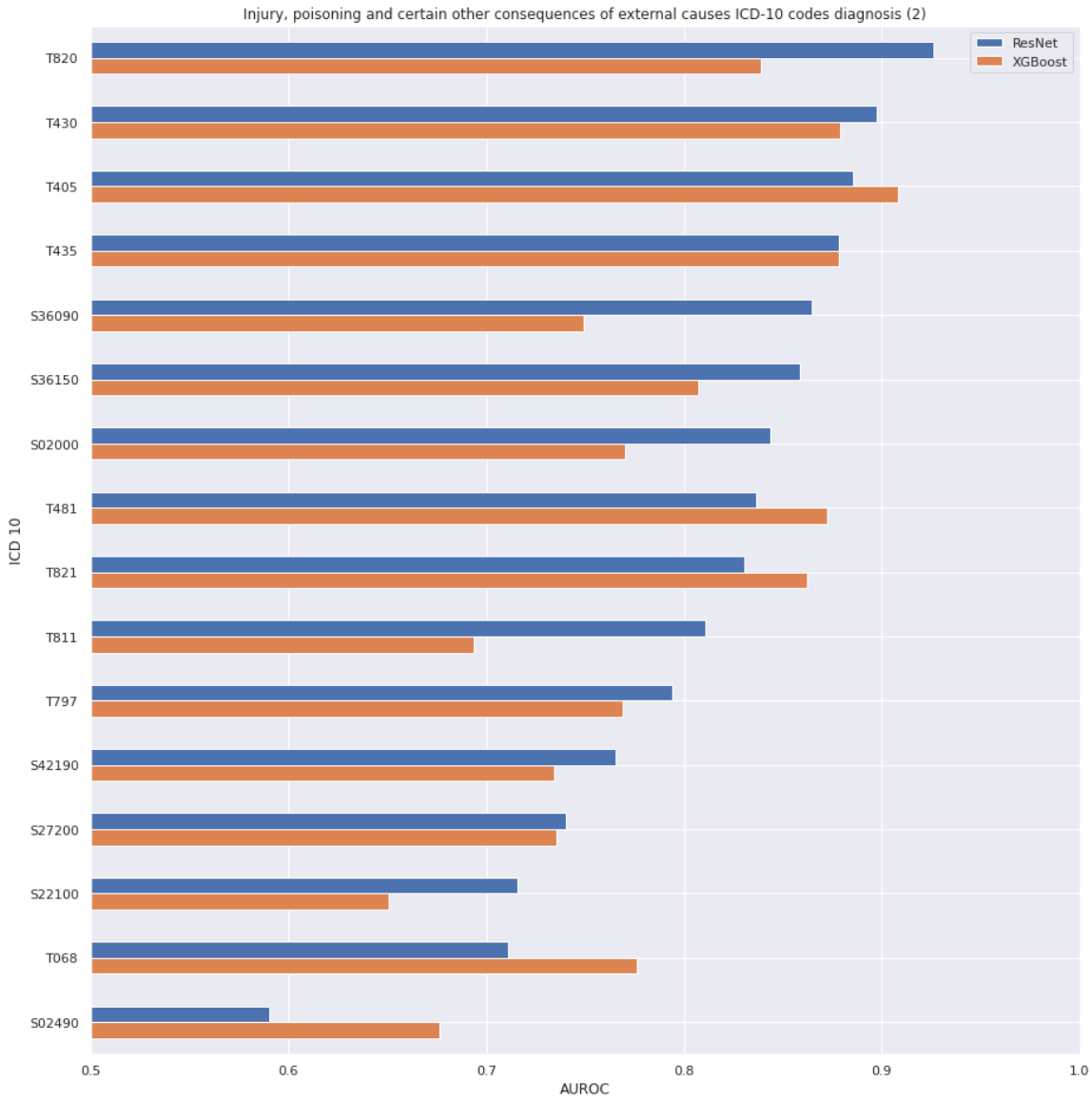

Figure A.18: Injury, poisoning and certain other consequences of external causes (2) AUROC plot

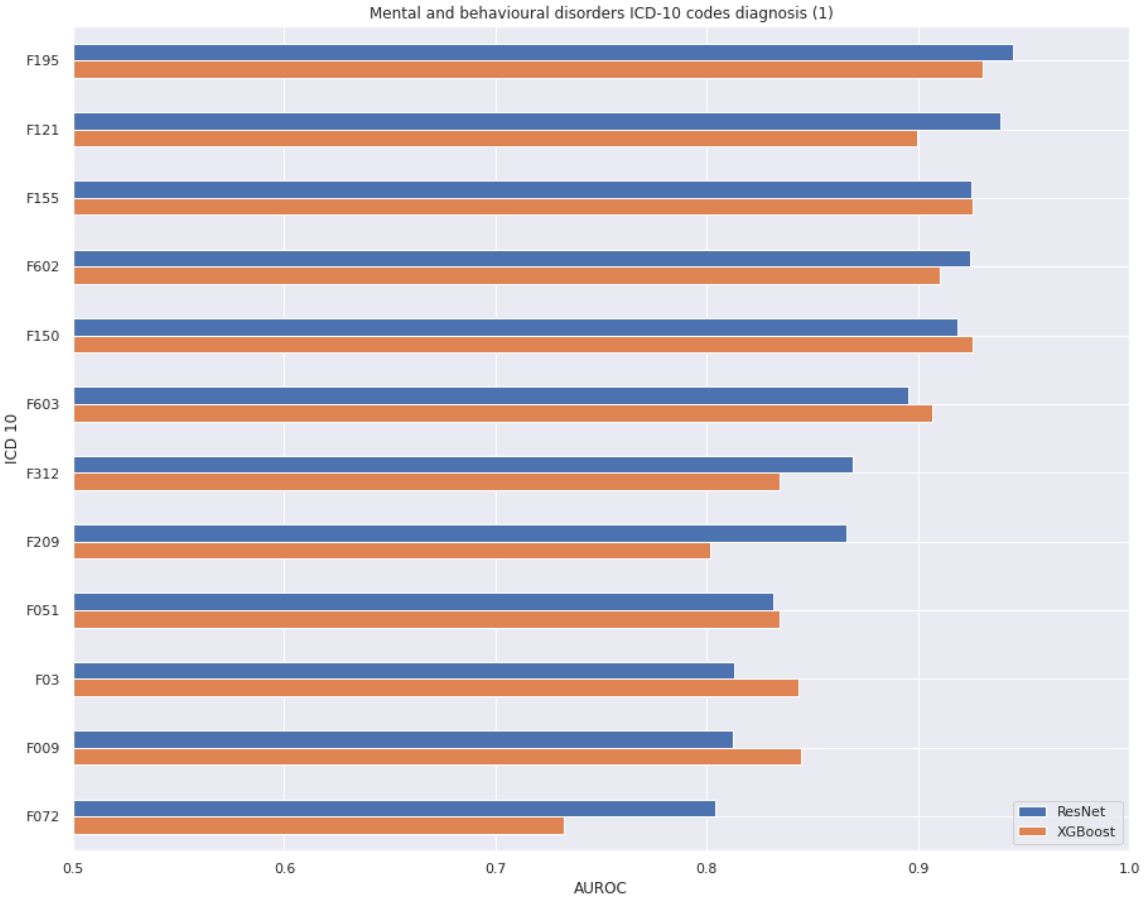

Figure A.19: Mental and behavioural disorders (1) AUROC plot

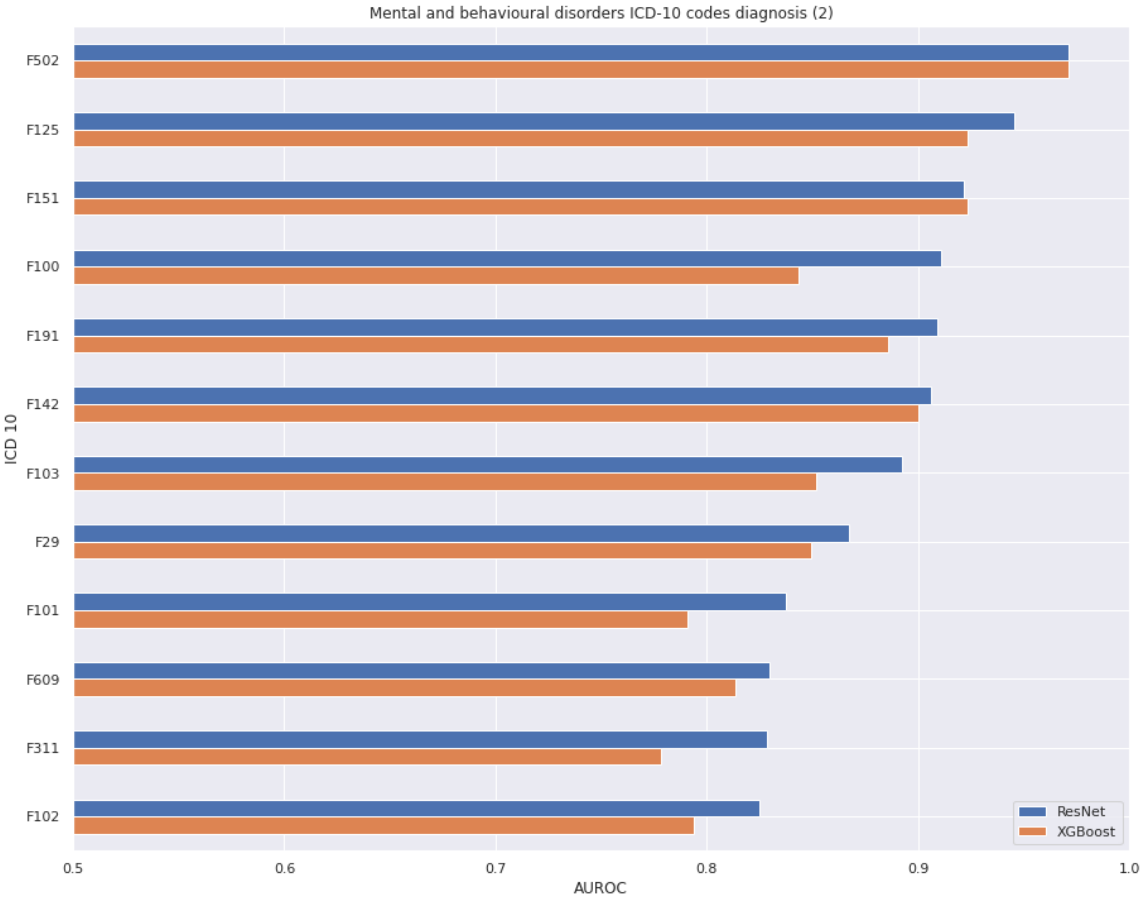

Figure A.20: Mental and behavioural disorders (2) AUROC plot

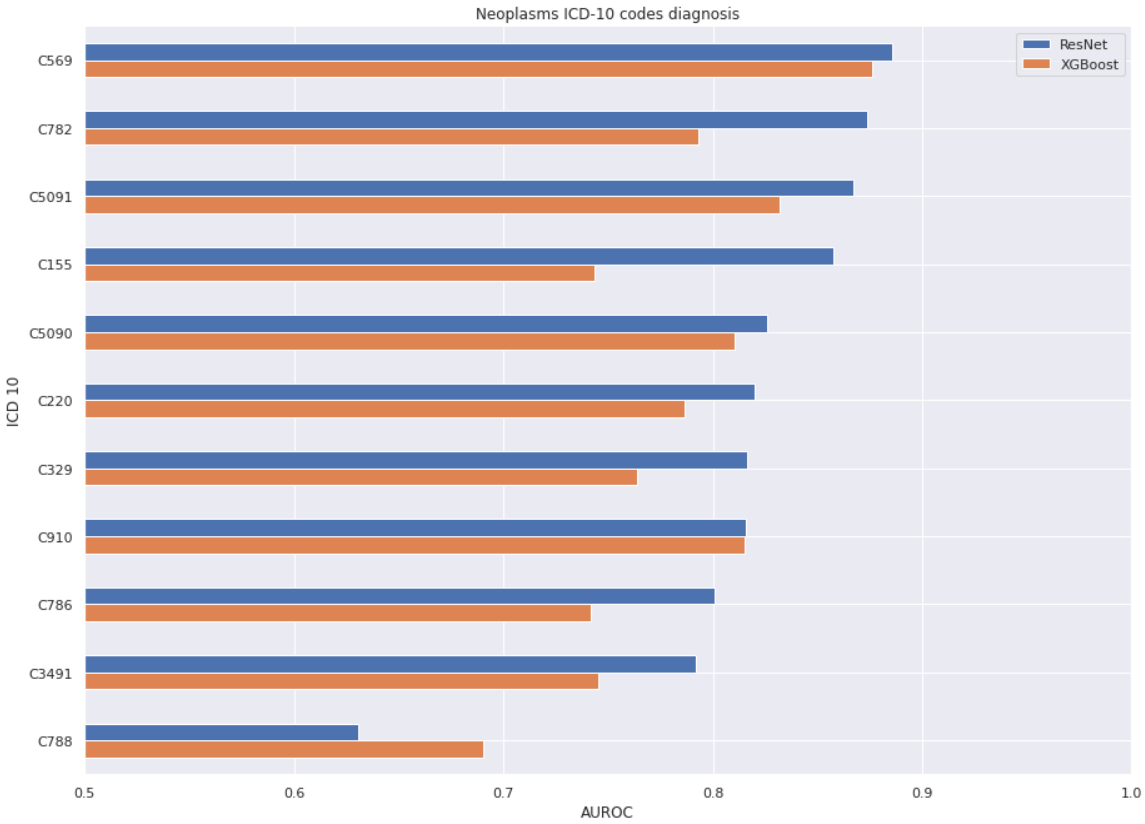

Figure A.21: Neoplasms AUROC plot

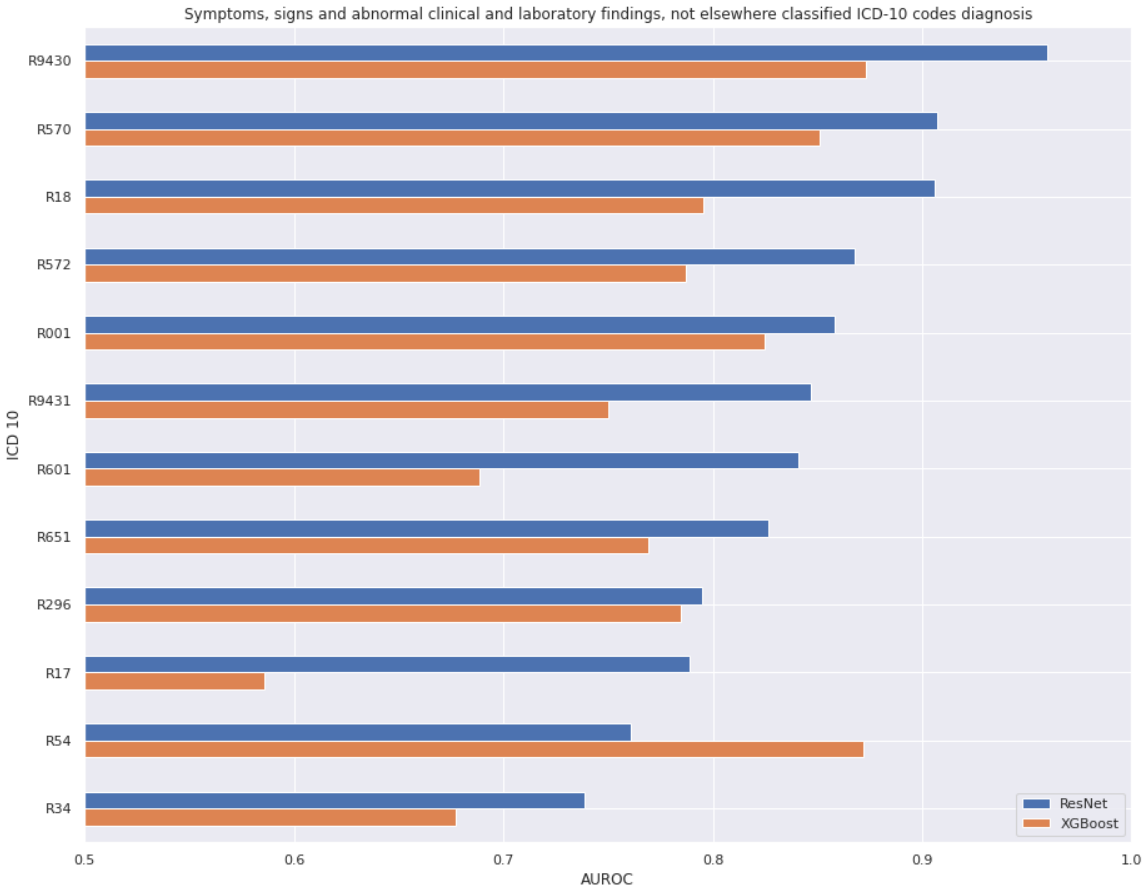

Figure A.22: Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified AUROC plot