Advancing ECG Analysis through Machine Learning: A Study on Data Generation for ECG Classification and Feature Selection For Individual Survival Prediction

by

Yousef Nademi

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

Master of Science

Department of Computing Science

University of Alberta

Abstract

Electrocardiograms (ECGs) are a valuable and easily-collected measurement of heart health, reflecting its morphology (R peak, QRS duration,..) and rhythm (sequence of multiple heartbeats). With the advance of machine learning, many studies utilize electrocardiogram (ECG) signals for various purposes, such as detecting ECG abnormalities, predicting patient mortality and other supervised tasks. In this thesis, we used the Alberta Hospital ECG Dataset consisting of more than 1.6 million ECG collected from 244,077 patients, for two objectives: (1) To produce a generative model, which can then be used to generate synthetic ECGs for a specific abnormality, which can augment a dataset of real ECGs, in order to improve the performance of a machine learned model for ECG abnormality classification. (2) To explore and compare different approaches for extracting high-level features from ECG signals and determine which approach is most effective in estimating patient-specific survival curves for accurately predicting time-to-death.

For the first objective, we used this ECG dataset, where each 12-lead ECG is labeled with one of 15 diagnoses abnormalities, to train an unsupervised beta variational AutoEncoder (β-VAE) model, that could generate synthetic 12-lead ECG signals time series, with each specified abnormalities. We then used this generative model to generate ECGs with the abnormality of ST-segment Elevated (STE). These generated ECGs were then added to the public dataset from the China Physiological Signal Challenge 2018, which contained 6,877 real ECGs. This dataset included healthy controls (sinus rhythm) and 8 different

abnormalities. We found that a learner trained on this extended dataset performed better than one trained on only the original data on the targeted STE label but also enhanced its performance for the classification of 4 other labels. For the second objective, we explored ways to obtain useful high-level features from ECG traces through various approaches, including supervised with clinical diagnoses, unsupervised approaches, and knowledge-based ECG features. Using these ECG features, along with age and sex, we trained models to estimate patient-specific individual survival distributions (ISD) to predict each patient's time-to-death. The results showed that ECG features produced by supervised learning approaches led to models that were superior in estimating patient-specific time until death than ECG features obtained from unsupervised and knowledge-based methods. In fact, the supervised ECG features required fewer training instances (as few as 500) to learn ISD models that performed better than models that only used age and sex. On the other hand, unsupervised and knowledge-based ECG features required over 5000 training samples to produce ISD models that performed better than ones using only age and sex. These findings may assist researchers in selecting the most appropriate approach for extracting high-level features from ECG signals to estimate patient-specific ISD curves.

Preface

The research conducted over the course of my thesis has directly led to two scholarly articles. The following are publications closely related to this work: Chapter 4 is derived from our work titled 'Generative Data by β-Variational Autoencoders Help Build Stronger Classifiers: ECG Use Case.' This research was not only orally presented at the 19th International Symposium on Medical Information Processing and Analysis (SIPAIM) in 2023, but also published in the symposium's proceedings. The full citation of the publication is Nademi, Yousef, et al. "Generative Data by -Variational Autoencoders Help Build Stronger Classifiers: ECG Use Case." 2023 19th International Symposium on Medical Information Processing and Analysis (SIPAIM). IEEE, 2023.

Chapter 5 is based on the paper 'Supervised Electrocardiogram (ECG) Features Outperform Knowledge-Based and Unsupervised Features in Individualized Survival Prediction.' This work was accepted for poster presentation in the Symposium on Machine Learning for Health (ML4H) in 2023. It was published in the symposium's proceedings, highlighting the importance and relevance of our research in this area. The full citation of the published work is Nademi, Yousef, et al. "Supervised Electrocardiogram (ECG) Features Outperform Knowledge-based And Unsupervised Features In Individualized Survival Prediction." Machine Learning for Health (ML4H). PMLR, 2023.

Acknowledgements

To begin, I want to express my appreciation to Dr. Russell Greiner, my supervisor during my MSc years, for their exceptional support, expertise, trust, and patience, as well as for being an amazing person. Additionally, I would like to thank Dr. Sunil V. Kalmady for their guidance and invaluable advice during our regular meetings. Furthermore, I am grateful for the expertise and support of Weijie Sun, Amir Salimi, Dr. Abram Hindle, and Dr. Padma Kaul. Lastly, I want to acknowledge the unconditional love and support of my parents, wife, and friends.

Contents

List of Tables

List of Figures

Abbreviations & Acronyms

β-VAE Beta Variational Autoencoder

AE Autoencoder

AFIB Atrial Fibrillation

AUROC Area Under The Receiver Operating Characteristic Curve

COX-PH Cox-Proportional Hazard

CPSC 2018 China Physiological Signal Challenge 2018 Dataset

CVDs Cardiovascular Diseases

DL Deep Learning

E2E DL End-To-End Deep Learning

ECG Electrocardiogram

FN False Negative

FP False Positive

GANs Generative Adversarial Networks

IAVB First-degree Atrioventricular Block

ICD-10 International Classification of Diseases, 10th revision

IPCW Inverse Probability of Censoring Weights

ISD Individual Survival Distribution

KM Kaplan-Meier

LBBB Left Bundle Branch Block

LR Logistic Regression

MSE Mean Squared Error

MTLR Multi-Task Logistic Regression

N-MTLR Neural Multi-Task Logistic Regression

PAC Premature Atrial Contraction

PVC Premature Ventricular Contraction

RBBB Right Bundle Branch Block

SR Sinus Rhythm

STD ST-segment Depression

STE ST-segment Elevated

TN True Negative

TP True Positive

VAE Variational Autoencoder

XGB Gradient Boosted Tree Ensembles

Chapter 1

Introduction

1.1 Motivation

Cardiovascular diseases (CVDs) are a major cause of death globally, and the likelihood of developing a CVD increases with age [11]. Early detection of CVDs is important, as it can help patients to manage their condition effectively and increase their lifespan. One of the main tools that cardiologists use to identify CVDs is electrocardiogram (ECG) measurements. ECGs measure the electrical activity of the heart, which can be analyzed to assess heart health, as it provides data on both the morphology (heartbeat) and the rhythm (sequence of multiple heartbeats). The advent of portable and wearable ECG devices, such as smartwatches, has significantly increased the volume of ECG data, underscoring the need for effective and efficient ECG data analysis. [21] ECG signals can vary between patients who have the same abnormalities, which can make it challenging to diagnose cardiovascular conditions accurately [6]. In addition to inter-patient variability, some ECG abnormalities are temporal, which means they may not appear in every heartbeat. For example, Atrial Fibrillation, a common heart rhythm disorder, may not display its morphological characteristics in every heartbeat of an ECG. Identifying such abnormalities requires long-term monitoring of the patient's heart [55]. The continuous recording of the patient's ECG includes multiple seconds of data, consisting of multiple beats. This extended observation period allows for a comprehensive analysis of the heart's activity. To gain a thorough understanding of the patient's heart health, the standard procedure involves using

12 leads. Each lead provides valuable insight into the heart's condition from a different angle, enhancing the accuracy and effectiveness of the diagnostic process. This is because certain abnormalities may only show their characteristics in specific leads of the 12-lead ECG [6], [33]. Despite the use of 12 leads, diagnosing ECG abnormalities remains a challenging task due to inter-patient variability and temporal changes in the signals, and now the expertise of an experienced cardiologist is required to accurately read and interpret ECG signals.

In recent years, the advancement of computational power and machine learning algorithms has led many researchers to explore their potential applications in the medical field, particularly in the areas of diagnosis and prognosis. To this end, two main categories of machine learning approaches have been used: supervised and unsupervised methods. Supervised learning involves training a model on the labeled data. Unsupervised learning, on the other hand, involves training a model on the unlabeled data. Our group has previously developed a diagnosis model for 15 different cardiovascular diseases by applying an end-to-end supervised deep learning (DL) model to 12-lead ECG signals. The model achieved a Area Under the Receiver Operating Characteristics (AUROC) performance of approximately 80% for 12 of the 15 CVD labels [**unpublished data** [1]]. One major drawback of deep learning models is their lack of interpretability, making it difficult to understand how the model is making its decisions. Some researchers have attempted to address this by using a two-step process. Firstly, they use unsupervised methods, such as variational autoencoders (VAE), to encode ECG signals into lower-dimensional embeddings. Secondly, they use these embeddings for downstream tasks such as multi-label classification of ECG abnormalities [24], [31]. Unsupervised approach, which utilizes unsupervised methods like Beta VAE (β-VAE) to encode ECG signals into lower-dimensional embeddings, provides more explainable results compared to deep learning approaches (See discussion in Section 4.4.1). In these studies, researchers have used β-VAE to encode either 1-lead ECG signals or

[1]This paper has been submitted to the 'npj Digital Medicine' journal and is currently under review.

1 beat of 12-lead ECG signals. Additionally, β-VAE models have the added advantage of being used for data generation, and these synthetic data can be used for data augmentation.

Bridging the gap between the need for interpretable models and the availability of sufficiently diverse training data, a notable advancement in ECG analysis is the development of algorithms capable of generating synthetic ECGs. Generative algorithms, such as Generative Adversarial Networks (GANs), have been utilized to create these synthetic ECG signals. These synthetic signals offer multiple advantages: they augment real datasets by enriching them, thus achieving a better balance between normal and abnormal cases in the training data. Also, synthetic ECGs circumvent the privacy and regulatory constraints often associated with the use of real ECG data [4]. This innovation is particularly critical in light of the prevalent issue of class imbalance in ECG datasets, where normal signals typically outnumber abnormal cases. Therefore, the generation of synthetic ECG signals not only addresses the challenges in machine learning model training but also paves new pathways for advancing ECG analysis methodologies.

Survival prediction models aim to estimate the time until a specific event occurs, such as hospital admission, death, or onset of a disease. These models can be binary (e.g., alive or dead after a certain time) or probabilistic. In addition, the model can provide a single estimate, or a set of estimates, one for each of several time points. For example, the model might provide estimates of survival probabilities at 1 month, 3 months, and 5 months after diagnosis. Haidar et al. [17] provide a comprehensive overview of survival distribution models, including individual survival distributions (ISDs) and also summarize various methods to evaluate such models. Unlike traditional regression approaches, it is challenging to learn survival prediction models as the dataset includes right censored instances which provides only a lower bound of the time. For instance, if the study period is 5 years, but a patient leaves the study after 3 years or is still alive at the end of the study, their exact survival time remains unknown. He/she could live just a day or several years beyond the study period, but this data is not captured, resulting in a right-censored

instance. ISD models, which incorporate patient-specific characteristics, are more useful for decision-making compared to survival prediction models that predict survival at a single time point (e.g., 1 year).

ECG signals carry information about the health of a patient's heart that can be used to estimate patient-specific ISD curves for different events of interest. For example, in the case of cardiac death as the event of interest, analyzing ISD curves might reveal a shorter expected survival time for a certain patient. This insight could prompt consideration of more aggressive treatment or medication strategies for this patient. Such tailored interventions, based on the anticipated survival time, might potentially result in a lower overall mortality rate. One approach to estimating patient-specific ISD curves for any event of interest using ECG signals is to first obtain embeddings/features from ECG signals. There are various approaches that can be used to extract these ECG features:

- Supervised model: A supervised model for feature extraction can be a viable approach in extracting ECG features. We can leverage a pre-trained neural network, which has already learned to classify ECG signals for a specific task such as multi-label classification, for a different downstream task.

- Unsupervised models: These unsupervised models (i.e. AEs, VAEs) can be used to extract features from ECG signals. In these models, the ECG signal is first encoded into a low-dimensional representation, and then this representation is used to estimate patient-specific ISD.

- Knowledge-based ECG features such as heart rate variability, QRS duration, QT interval, and others can be extracted during ECG collection; they can then be used as ECG features for other supervised tasks.

1.2 Contributions of the Thesis

My contributions to this thesis include:

1. We developed a β-VAE model to generate synthetic 12-lead ECG signals with specific abnormalities. This advancement facilitates the augmentation of real, labeled ECG datasets, particularly for minority labels. See Chapter 4

2. We demonstrate the value of synthetic ECGs, generated by the β-VAE model, in enhancing the performance of classifiers for ECG abnormality detection. This finding underscores the significance of synthetic data in training more robust diagnostic models. Refer to Chapter 4.

3. We leveraged a substantial dataset consisting of over 1.6 million 12-lead ECGs from 244,077 patients. This extensive dataset provides a solid foundation for analyzing and comparing the efficacy of various feature extraction methods.

4. We found that supervised ECG feature extraction methods, especially those utilizing 1414 ICD-10 codes, yield more accurate ISD predictions than unsupervised or knowledge-based methods. This discovery is crucial in guiding future research and practice towards more effective ECG-based prognostic tools. See Chapter 5.

5. We established that supervised ECG features require a significantly smaller training sample size to surpass the performance of the baseline model, which only uses age and sex for estimating ISD for survival prediction. This efficiency in training could have significant implications for the practical application of ISD models. See Chapter 5.

6. The findings from this research pave the way for future studies to explore and develop advanced ECG feature extraction methods, particularly those based on supervised or semi-unsupervised learning. This could lead to more accurate and efficient tools for ISD estimation using ECG data.

1.3 Thesis Outline

Chapter 2 discusses ECG measurements and reviews the architecture of the β-VAE model. Chapter 3 outlines the datasets, models, the procedure for generating data using the trained temporal convolutional network (TCN)-based β-VAE, and evaluation metrics used in the study. Chapter 4 discusses the experiments involving data generation using a public dataset. Chapter 5 compares the performance of models developed using various ECG features, obtained from different models, to estimate ISDs for the prediction of time until death. Chapter 6 provides a summary of the key findings of the thesis and outlines potential directions for future research.

Chapter 2

Background And Related Literature

This chapter offers essential introductory details for the next chapters – Section 2.1 discusses ECG measurements, then Section 2.2 describes the autoencoder (AE) and a variation of AE known as variational AE (VAE).

2.1 ECG Measurements

The ECG collection involves the placement of electrodes on a person's chest, arms, and legs to detect the heart's electrical activity from various angles, typically using 12 leads (Figure 2.1). The ECG exam captures the intensity and timing of electrical impulses as they travel through the heart, which can help investigate symptoms related to heart issues [18].

The anatomy of a heart, including its chambers, is shown in Figure 2.2. The heart consists of four chambers: two upper chambers called atria and two lower chambers called ventricles. Blood flows from the body into the right atrium, to the right ventricle, then to the lungs; oxygenated blood returns to the left atrium, moves into the left ventricle, and is then pumped throughout the body. This requires the heart to contract and relax in a coordinated cycle, corresponding to the P wave for atrial contraction, the QRS complex for ventricular contraction, and the T wave for the ventricles' return to a resting state.

Figure 2.3 shows the cycles of an ECG, which consist of various waves,

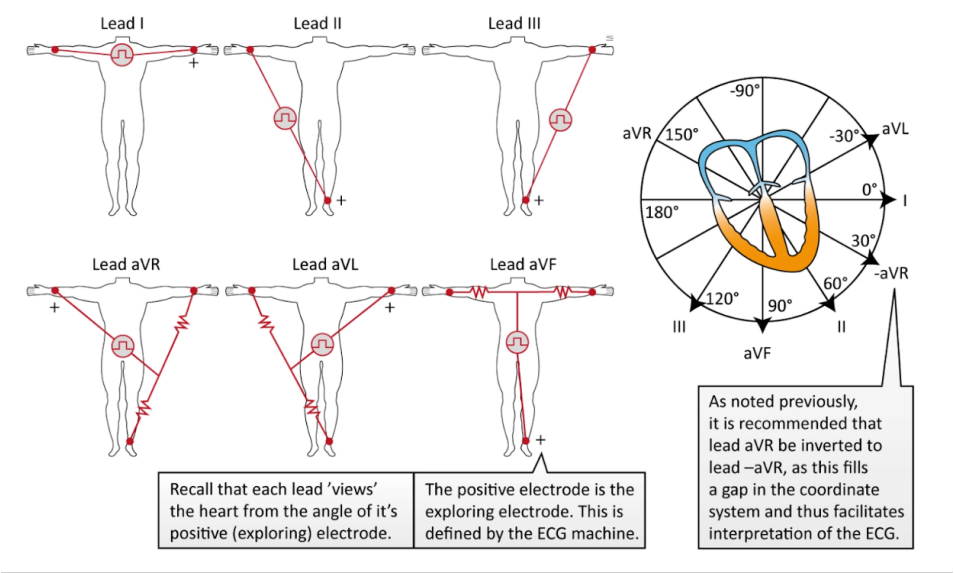

Figure 2.1: The electrodes and the corresponding angles that are linked to each ECG lead. We directly obtained the figure from [1].

here P, Q, R, S, T. The shape, location, and duration of each wave serves as an indicator of the heart's well-being [18]. The P wave is a slight drop in voltage from the baseline that occurs when the atria depolarize before contracting. The QRS complex indicates the coordinated activation of both the right and left ventricles. However, the majority of the waveform comes from the larger muscles in the left ventricle [18]. After the QRS complex, the T wave appears and signifies ventricular repolarization, which is the passage of electrical current sequentially within the heart muscle, followed by resting polarized state (no electrical activity). This process readies the cardiac muscle for the upcoming ECG cycle [6]. For further explanation of these terms, refer to [36].

The device used for ECG collection records signal information from the patient. Then, using those signals, the measurement device algorithm (in our case, the Philips DXL ECG Algorithm) provides some summary statistics of 12-lead ECGs. We obtained our ECG data using Philips IntelliSpace systems, which utilizes knowledge-based algorithms to generate ECG measurements commonly employed for clinical interpretation. Table 2.1 summarizes these features along with their descriptions. As examples and based on all 12-leads,

Table 2.1: Description of Philips' ECG measurements and their definition.

Feature	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	Milivolts	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	Milivolts	T onset

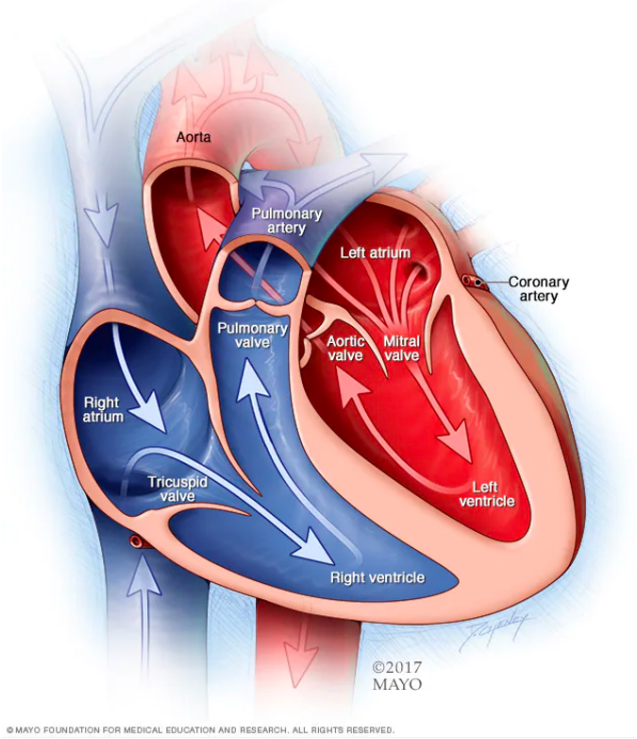

Figure 2.2: The anatomy of the heart consists of four chambers: two upper chambers called atria and two lower chambers called ventricles. Image source from [2]. Blood flows from the body into the right atrium, to the right ventricle, then to the lungs; oxygenated blood returns to the left atrium, moves into the left ventricle, and is then pumped throughout the body. This requires the heart to contract and relax in a coordinated cycle, corresponding to the P wave for atrial contraction, the QRS complex for ventricular contraction, and the T wave for the ventricles' return to a resting state.

Pdur is the duration of P wave in the ECG cycles, and **RRint** is the average R-R distance between two subsequent beats for an ECG signal. For more information on the definition of each feature, please refer to [3]. Also, there are some publicly available libraries that can extract these statistics for each individual beat of the signal, including Neurokit [35] and TSfresh [9] libraries.

The development of machine learning technology has led many researchers to use electrocardiogram (ECG) signals for various purposes. They have been used to detect ECG abnormalities [45], predict mortality [52], and identify individuals based on their ECG signals [42].

We consider three approaches for this task of using a patient's ECG signal

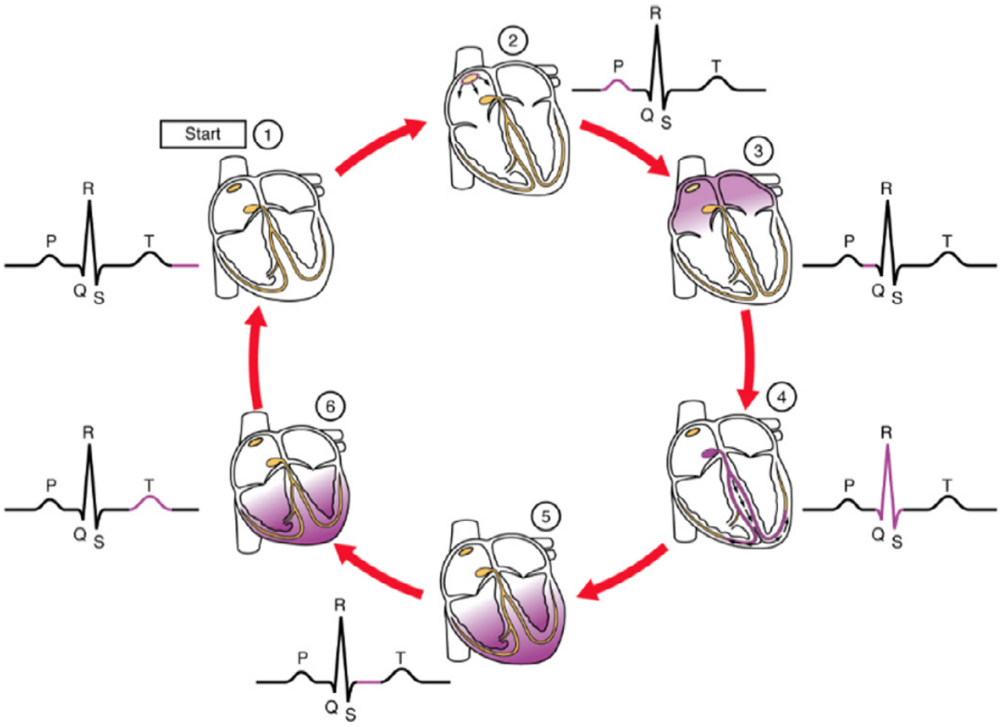

Figure 2.3: This schematic illustrates the waves of an ECG signal and the cardiac cycle of the heart, as detailed in the steps below. Permission to use this image has been granted by the publisher of Merone et al. [36]. 1) **Start of the ECG Cycle**: This initial stage does not correspond to any specific wave but signifies the beginning of the ECG cycle. 2) **P wave**: Reflecting atrial activation, the P wave, due to relatively weak atrial systole, displays a small size, with an amplitude not exceeding 0.4 mV and a duration of 60 to 120 ms. 3) **PQ** Stretch: Characterized by its flat nature, this segment marks the time from atrial activation to ventricular activation, typically lasting between 12 and 20 ms. 4) **QRS Complex**: Comprising Q, R, and S waves, this complex, varying in duration from 60 to 90 ms, provides insights into various heart functions. 5) **ST Stretch**: Extending from the end of the S wave to the start of the T wave, this segment, which generally lasts between 230 to 460 ms, corresponds to the ventricles' contraction and relaxation phase. 6) **T wave**: Indicating ventricular repolarization, this wave occurs as the ventricles complete activation and prepare for the next contraction, with a duration between 100 and 250 ms.

.

to predict the multi-label classification of cardiovascular diagnoses: (1) feed the knowledge-based features – here those provided by Philips Machine – directly to the downstream task (classification, mortality,..); (2) using unsupervised

approaches such as autoencoder (AE) or variational autoencoder (VAE) to encode ECG signals into the lower-dimensional features/embeddings and then use this encoding as input for the supervised task, and (3) utilizing end-to-end machine learning approaches using raw ECG signal for supervised tasks.

Each of these approaches has its own advantages and disadvantages. Models trained on these knowledge-based ECG features are generally less accurate than supervised or unsupervised methods, but the models are easier to interpret as features have clinical physical meaning [47]. Deep learning methods are shown to achieve superior performance, specifically in diagnosing ECG abnormalities, when applied to large datasets of voltage-time series from 12-lead ECG traces [45]. In contrast, shallow learning algorithms like XGBoost are more suitable for analyzing knowledge-based ECG features. These models provide interpretable results and require less training time for model's training. By 'interpretable results,' we mean that methods like Shapley Additive Explanations (SHAP) [58] can be used to shed insight on rational behind a model's predictions. It is important to note, however, that even though these shallow learning models require less time for training, their performance may not necessarily exceed that of deep learning methods [47].

2.2 Autoencoders

The goal of the unsupervised AE method is to encode input data into a lower dimension/embedding with minimal information loss, in that the input data should be reconstructed using these embeddings. AEs consist of three main components: the encoder, which compresses input data, the Bottleneck, which forms the lower embeddings/features, and the decoder, which reconstructs the original input data from the lower embeddings (Figure 2.4).The encoder and decoder components of AEs can be constructed using fully connected or convolutional layers. Convolutional layers, specifically, are designed to process data with a grid-like topology, such as images. They apply a convolution operation to the input, passing the result to the next layer. Mathematically, given the input data $x \in \mathbb{R}^r$, the encoder aims to learn the function $f(x)$ that encodes

important information in a lower representation $h \in \mathbb{R}^k$. The objective of the decoder function $g(h)$ is to reconstruct the input data from the embedding h, such that $g(h) \approx x$. During training, the AE model tries to minimize the dissimilarity between the original input data x and the reconstructed data $x' = g(h(x))$. For regression tasks, a commonly used loss function is the mean squared error (MSE),

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x'_i - x_i)^2 \tag{2.1}$$

where n is the number of data points, x_i and x'_i are input data and reconstructed data, respectively. Following the introduction of AE, several modified versions of AE have been developed, including but not limited to: Variational AE (VAEs), Beta-VAEs (β-VAE) [20], and Sparse AE [49]. These methods represent a spectrum of AE architectures, each with unique characteristics and applications. VAEs and their extension, β-VAE, fall under generative models that are suitable for learning latent representations. β-VAEs introduce a tunable hyperparameter β to the VAE framework, which helps in disentangling the latent representations. Here, "disentangling the latent representations" means the model's capability to separate distinct features in the latent space. This separation allows the model to learn representations where different dimensions in the latent space correspond to different attributes of the input data. Sparse AE introduces an L1 regularization to the loss function to learn a limited and more distinctive set of features by activating only a small number of neurons in the hidden layers. Sparse AEs are useful for feature selection and anomaly detection.

2.3 Variational Autoencoders

Unlike traditional AE, VAEs have the ability to generate data that is similar to the input data used for its training. This capability is due to the probabilistic nature of VAEs, which allows them to learn a probability distribution over the input data, enabling them to generate new samples that follow the same distribution. Here, we will provide a brief overview of our approach here. For

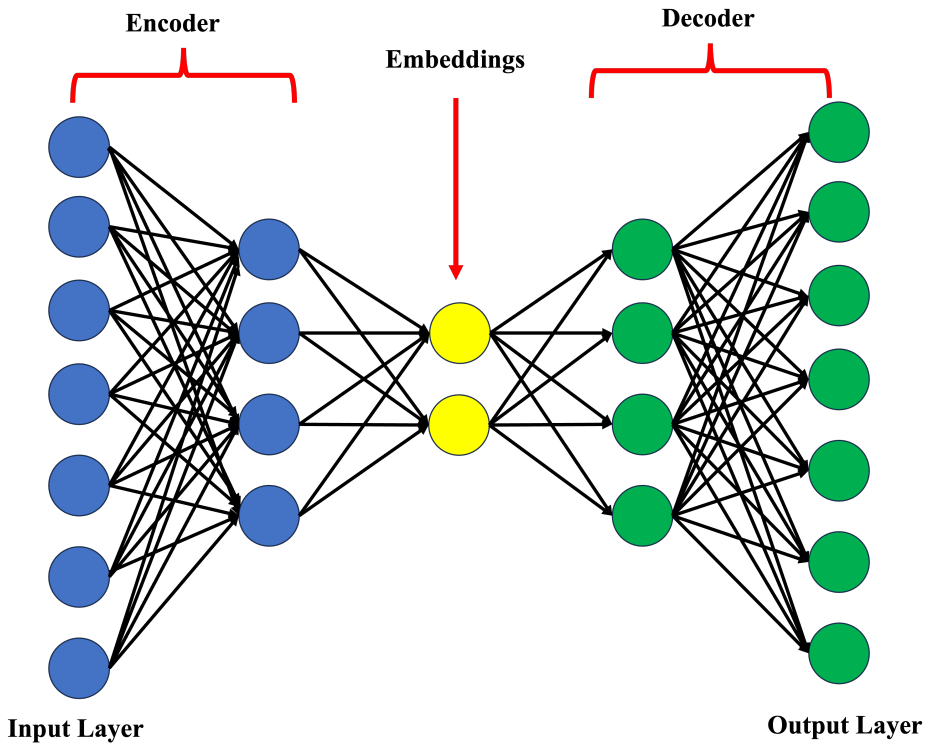

Figure 2.4: Schematic representation of AE.

a more detailed explanation of the methodology, please refer to the work by Kingma et al. [29].

VAE is a type of directed graphical model with continuous latent variables. The model consists of a generative component, represented by $p_\theta(z)p_\theta(x|z)$, and a variational approximation, denoted by $q_\phi(z|x)$, which is an approximate posterior distribution of the latent variables given the observed data. Here, $p_\theta(z)$ is the prior distribution over the latent variables z, and $p_\theta(x|z)$ is the conditional distribution of the observed data x given the latent variables. The encoder of VAE learns the variational approximation component, and the decoder learns the generative component. Now, consider that we have a data set $D = \{x_{(i)}\}_{i=1}^{N}$ that consists of N independent and identically distributed (i.i.d) samples of a variable x, which could be either continuous or discrete. The samples are assumed to be generated through a two-step random process involving an unobserved continuous random variable z, where:

- First, a value $z_{(i)}$ is generated from a prior distribution $p_{\theta^*}(z)$.

- Second, a value of $x_{(i)}$ is generated from a conditional distribution of $p_{\theta^*}(x|z)$

Both the prior distribution $p_{\theta^*}(z)$ and the likelihood $p_{\theta^*}(x|z)$ are assumed to be belong to parametric families of distributions of $p_\theta(z)$ and $p_\theta(x|z)$. $q_\phi(z|x)$ is refer to as probabilistic encoder, where given a data point x, the encoder outputs a distribution (often Gaussian) over the possible latent variable z, that could have generated x. The $p_\theta(x|z)$ is refer to as a probabilistic decoder, where given a latent variable z, the decoder outputs a a distribution over the possible corresponding values of x. The overall probability of observing the entire dataset is calculated as the product of the probabilities of each individual data point in that dataset

$$\log p_\theta(x_{(1)}, \ldots, x_{(N)}) = \sum_{i=1}^{N} \log p_\theta(x_{(i)}) \tag{2.2}$$

which can each be rewritten as:

$$\log p_\theta(x_{(i)}) = D_{KL}(q_\phi(z|x_{(i)})||p_\theta(z|x_{(i)})) + L(\theta, \phi; x_{(i)}) \tag{2.3}$$

where the first term, $D_{KL}(q_\phi(z|x_{(i)})||p_\theta(z|x_{(i)}))$, is KL divergence. It measures the divergence between the approximate posterior distribution $q_\phi(z|x_{(i)})$ and the true posterior distribution $p_\theta(z|x_{(i)})$. This term quantifies how well the variational approximation, parameterized by ϕ is performing in estimating the true distribution, parameterized by θ. The second term, $L(\theta, \phi; x_{(i)})$, is called the (variational) *lower bound* on the marginal likelihood of data point i, and can be written as follows:

$$L(\theta, \phi; x_{(i)}) = -D_{KL}(q_\phi(z|x_{(i)})||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x_{(i)})}[\log p_\theta(x_{(i)}|z)] \tag{2.4}$$

where the first term represents the negative KL divergence between the approximate posterior $q_\phi(z|x_{(i)})$ and the prior distribution $p_\theta(z)$. Minimizing this divergence is essential for ensuring that the approximate posterior is as close as possible to the prior. The second term, $\mathbb{E}_{q_\phi(z|x_{(i)})}[\log p_\theta(x_{(i)}|z)]$, is the expected log-likelihood under the approximate posterior distribution. It quantifies the expected fit of the model to the data point x_i given the latent variable z. The objective is to minimize this lower bound.

2.4 Beta-Variational Autoencoders (Beta-VAEs)

β-VAE [20] is a special case of VAE that adds the hyperparameter of β to the loss function (divergence loss term) to learn disentangled features. The adjustable parameters of β-VAE are the number of features and β. This modification enhances the model's ability to learn disentangled representations by adjusting the balance between the reconstruction of input data and the regularization of the latent space, enforced by the KL divergence.

2.4.1 Loss Function

The loss function for a β-VAE is formulated as:

$$L(\theta, \phi; x_{(i)}) = -\beta D_{KL}(q_\phi(z|x_{(i)})||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x_{(i)})}[\log p_\theta(x_{(i)}|z)] \tag{2.5}$$

Here, the first term represents the reconstruction loss, while the second term is the β-weighted KL divergence. This divergence quantifies the discrepancy between the encoder's distribution and the prior distribution of latent variables.

Chapter 3

Method and Evaluation Metrics

This chapter will first discuss the the ECG datasets and the methodology used for future chapters. We will also explain the preprocessing steps to prepare the data for our tasks (unsupervised learning of ECG signals and survival prediction). Second, we will explain the different architectures that we will use for training the unsupervised β-VAE algorithms. Third, we will discuss the Neural Multi-task Logistic Regression (N-MTLR) [12] and Cox-Proportional Hazard (COX-PH) [27] models that we will use to estimate patient-specific individual survival distributions (ISD). Then, we will present the process of ECG data generation using the learned β-VAE algorithm. Finally, we will explain various evaluation metrics that we will use for the multilabel classification of ECG abnormalities and ISD models.

3.1 Datasets

3.1.1 Alberta ECG Dataset

We will use the Alberta Hospital Dataset, which consisted of 12-lead ECG signals collected from 244,077 patients using a Philips IntelliSpace ECG machine with a sampling frequency of 500 Hz for a duration of 10 seconds. Each ECG is labeled with zero or more of 15 possible labels: Non-ST-Elevation Myocardial Infarction (NSTEMI), ST-elevation myocardial infarction (STEMI), Heart Failure (HF), Unstable Angina, Atrial Fibrillation (Afib), Ventricular Tachycardia, Cardiac Arrest, Supraventricular Tachycardia, Atrioventricular Block, Pulmonary Embolism, Aortic Stenosis, Pulmonary Hypertension, Hy-

pertrophic Cardiomyopathy, Mitral Valve Prolapse, and Mitral Valve Stenosis. The machine algorithms detected and removed ECGs with poor quality such as muscle artifacts, AC noise, baseline wander, QRS clipping, and leads-off during pre-processing [3]. For a more detailed exploration of the Alberta ECG dataset and its comprehensive nature, the reader is referred to Sun's thesis [50].

For supervised approaches, no further preprocessing is required. However, for unsupervised approaches, more preprocessing steps are required. Since we are using mean squared error (MSE), these unnormalized signals can lead to significant loss values and might disturb the learning process. So, for all unsupervised learning algorithms used here, we include filtering out baseline noise with a Butterworth filter (provided by Neurokit library [35]) and normalizing the ECG signals over time using the z-score normalization method. We also used common measurements of ECG provided by the Philips IntelliSpace ECG system (see Table 2.1) for the survival prediction task. To train the models, we split the dataset into a development set (964,741 ECG signals from 146,466 patients) and a test set (640,527 ECG signals from 97,631 patients – disjoint from the training patients), using 60% and 40% of the data, respectively.

The Health Research Ethics Board at the University of Alberta approved the use of ECG data in this study

3.1.2 China Physiological Signal Challenge 2018 Dataset

The dataset (CPSC 2018) is provided by the challenge competition; see Liu et al. [32] for the demographic details and a description of each label. The dataset consists of 12-lead ECGs collected from 11 hospitals using the frequency of 500 Hz, each with one or more of 9 possible labels: Sinus Rhythm (SR), Atrial Fibrillation (AFIB), First-degree Atrioventricular Block (IAVB), Left Bundle Branch Block (LBBB), Right Bundle Branch Block (RBBB), Premature Atrial Contraction (PAC), Premature Ventricular Contraction (PVC), ST-segment Depression (STD), and ST-segment Elevated (STE). The dataset was previously divided into the training set (6877 instances (female: 3178; male: 3699)) and test set (2954 instances (female: 1416; male: 1538)) by the competition, where the test set, which is still not public (both signals and their

labels), was used to rank participants. Table 3.1 shows the number of training set recordings for each label. The majority of the ECG training data has only 1 label (6401 samples), while 477 samples have multiple abnormalities in their ECGs.

Here, we divided the the training CPSC dataset into 80% training (5503 ECGs), 10% validation (687 ECGs), and 10% test set (687 ECGs). The test set was fixed for all experiments (with and without data augmentation).

3.2 Learning Algorithms

This section discusses the learning algorithms, including both supervised and unsupervised learning approaches, that are mentioned in the next chapters. Chapter 4 then uses an unsupervised temporal convolutional network (TCN)-based β-VAE algorithm to learn the characteristics of the Alberta ECG dataset. Our model then uses the learned algorithm to generate new ECG signals from this dataset, to augment the training dataset of CPSC 2018 (which is fed to a supervised learning algorithm for multi-label classification of ECG abnormalities.). Chapter 5 describes several supervised and unsupervised algorithms to extract features from ECG signals (Figure 3.1). Cox-Proportional Hazard (COX-PH) and Neural Multi-Task Logistic Regression (N-MTLR) algorithms then uses these ECG features to estimate patient-specific ISD. We then use these ISDs to estimate the time until death for each patient.

3.2.1 Unsupervised Models

TCN Based β-VAE

In Chapter 4, we will train a generative TCN-based β-VAE model in an unsueprvised manner using Alberta ECG dataset. Then, we will use this trained generative model to produce synthetic ECGs with specific diagnosis. In Chapter 5, we will use the extracted features obtained from this trained TCN-based β-VAE model to estimate patient-specific ISD (Figure 3.1-d). Here, For simplicity, we call this approach **TCN-β-VAE**. We used β-VAE architecture and code provided by van de Leur et al. [31], but modified it to train and recon-

Table 3.1: Total number of ECGs in the training set for various labels. The column "Total # of Recordings" indicates the total occurrences of each label. "Single Label Occurrence" shows how many ECGs have the label with no other labels. "With Exactly Two Labels" shows the count of ECGs where the label is present with exactly one other label, denoted as OtherLabel (Number), where Number is the count of ECGs with this specific pair.

Challenge Set	**Label**	**Total # of Recordings**	**Single Label Occurrence**	**With Exactly Two Labels**
Training	SR	918	918	0
	AFIB	1098	976	RBBB(86), PVC(4), LBBB(13), STE(1), PAC(2), STD(15)
	1-AVB	704	686	RBBB(5), PVC(2), LBBB(4), STE(2), PAC(2), STD(3)
	LBBB	207	179	PVC(3), AFIB(13), 1-AVB(4), STE(2), PAC(5)
	RBBB	1695	1533	PVC(26), AFIB(86), 1-AVB(5), STE(8), PAC(26), STD(10)
	PAC	574	533	PVC(1), RBBB(26), LBBB(5), AFIB(2), 1-AVB(3), STE(1), STD(3)
	PVC	653	607	RBBB(25), LBBB(3), AFIB(4), 1-AVB(3), STE(1), STD(9), PAC(1)
	STD	826	784	RBBB(10), PVC(9), AFIB(15), 1-AVB(3), STE(1), PAC(3)
	STE	202	185	RBBB(8), PVC(1), AFIB(1), RBBB(2), 1-AVB(2), PAC(1), STD(1)
Total		**6877**	**6401**	

struct 12-lead ECG traces of Alberta Dataset (see Figure 3.2). The adjustable parameters of β-VAE are the number of features and β, which we set to 32 and 8, respectively. These values were chosen the same as van de Leur et al. [31]

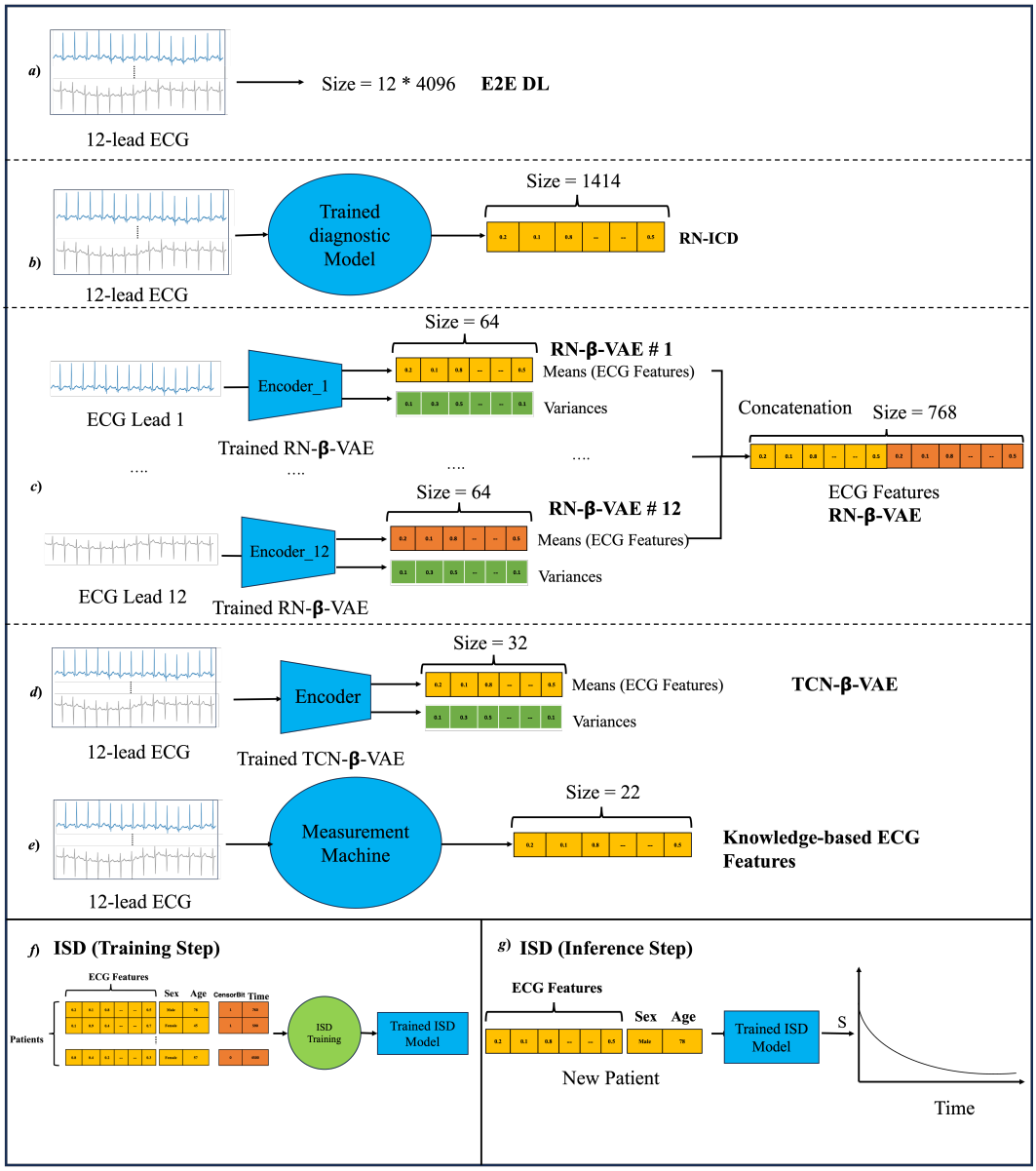

Figure 3.1: a-e) ECG feature extraction methods used to train ISD models. f) ECG features are utilized to train ISD models. g) In the ISD inference step, we use the trained ISD model to estimate the ISD for a new patient, taking into account his/her ECG features along with age and sex.

study, which performed a grid search to find these values; see that paper for the details of the model's architecture. In short, the temporal convolutional neural networks (TCN) [30] were used as layers of the encoder and the decoder. TCN represent a specific convolutional technique applied to temporal data, where it requires that the model maintain the data's temporal order. This means that

the model's prediction estimated at the next time step, cannot be influenced by any of the future time steps. To ensure disentanglement, the loss function put more weight (parameter of β) on the KL divergence term. The multivariate Gaussian distribution is selected as the prior for the latent variables. We added SoftPlus activation function [65] to the linear layer that is responsible for learning the variances, along with a small non-negative value=0.001, to ensure that the standard deviations remain positive and to prevent potential numerical instability that could arise if the value approaches zero. The encoder using a convolutional deep neural network encodes the input ECG signal into 32 pairs [means, variances]. The building blocks of the encoder is causal convolution block composed of causal convolutions [30], weight normalizations, leaky Rectified Linear Units (ReLUs) [59], and residual connections [19]. As explained by Van Den Oord et al. [40], the residual connection is only utilized when there is a change in the number of input channels/filters. The dilation parameter utilized in the causal convolutional layer is doubled in each subsequent causal convolution block. Weight Normalizations is a technique that normalizes the weights in neural networks, which can speed up training and lead to faster convergence. During the learning process (Figure 3.3), we fed 12-lead ECGs into the encoder section of β-VAE. Each ECG input consists of 12 leads, with each lead comprising 4096 32-bit floating point samples, representing 10 seconds of data per lead. The ECG data were stored and processed to maintain the correct order of the ECG leads. Thus, the input to the model has the dimensions of batch size $\times$ 12 $\times$ 4096. The encoder includes 7 serial layers, as indicated by the '7x' notation in Figure 3.2. After the encoding process, the signals are encoded into a batch size $\times$ 32-tuple. Using this intermediate representation, we drew a sample from a Gaussian distribution based on the computed means and variances. We then fed the resulting batch size $\times$ 32 samples into the decoder, which mirrors the encoder. The decoder's task is to reconstruct the input ECG signal from this lower-dimensional representation. We used a batch size of 32 and trained the model for a duration of 40 epochs. Early stopping was employed, and we stoped the training if the loss did not decrease after 3 consecutive epochs on the validation set. We trained

the model using the Adam optmimizer [28] with the learning rate of 0.001.

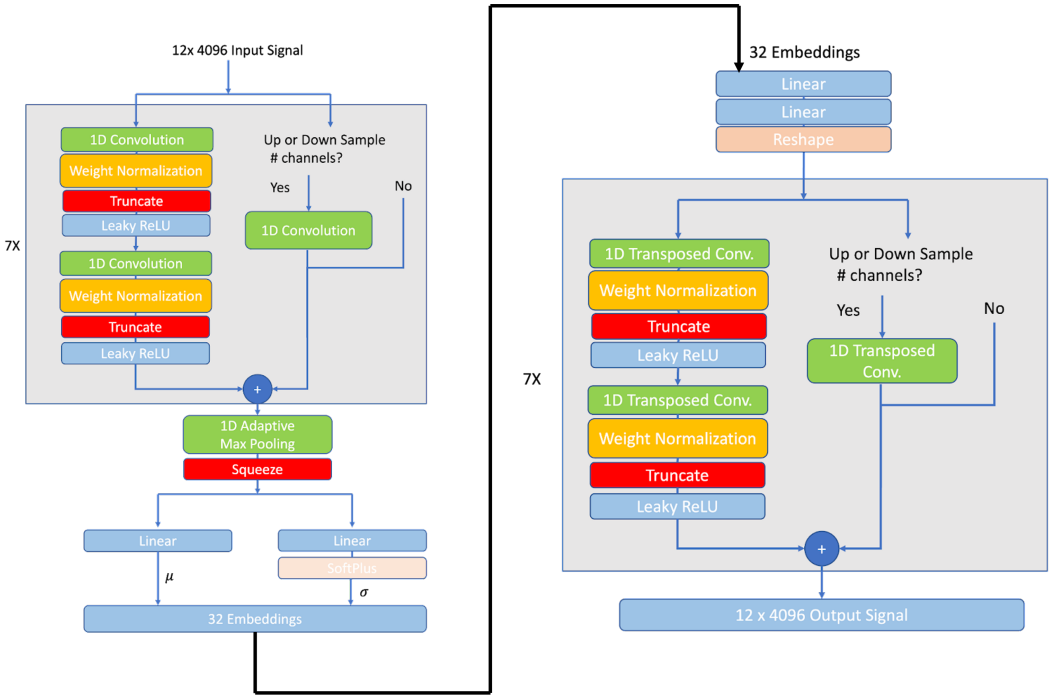

Figure 3.2: The architecture of TCN-based β-VAE to learn 12 leads ECG signals. Later on, this architecture was used to generate synthetic ECG signals.

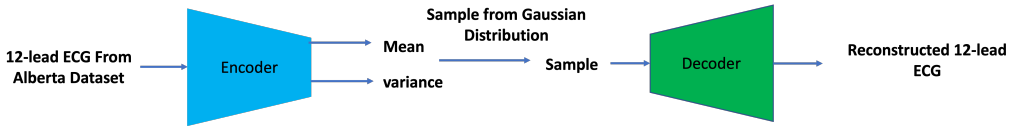

Figure 3.3: Schematic of the learning process of β-VAE.

ResNet Based β-VAE

Residual Networks (ResNets) are a foundational architecture in deep learning, initially developed for image processing tasks [19]. However, the application of ResNets extends beyond image processing and was successfully applied to times series [57] and sequential data [5]. These networks are characterized by their 'residual connections' which allow the network to skip one or more layers. These connections help in alleviating the vanishing gradient problem by enabling direct paths for gradient flow.

In our study, we use β-VAE architecture with residual connections, following the proposal by Jang et al. [24], and adapt it for our ECG dataset. This model design is created to learn single-lead ECG signals, and Figures 3.4 and 3.5 show a modified diagram of its architecture. For each of the 12 lead signals, a separate Resnet-based model was constructed. In Chapter 5, the extracted features from each ECG lead is used to estimate patient-specific ISD (Figure 3.1-c). The architecture consists of an encoder and a decoder. The encoder is formed from multiple residual encoder blocks, where each block consists of 2 blocks of 1-dimensional convolutional neural networks (1D CNN), Relu activation function and batch normalization. Each ECG instance is fed to the networks as a 4096×1 [number of data points in each signal $\times$ lead] numeric matrix. To train the model, we need to set two adjustable hyper-parameters of the β-VAE: embedding size and β. We chose the values 64 and 8, respectively, after conducting a grid search on a small subset of the training and validation set. During the learning process, the encoder takes in batches of single-lead ECGs and encodes it into 64 pairs [means, variances]. From these parameters, a sample is drawn from the Gaussian distribution, producing a 64-tuple that serves as input for the β-VAE decoder. The decoder's goal is to reconstruct the input ECG signal with low error. The decoder consists of multiple deconvolution layers that mirror the encoder blocks. After training the β-VAE model, the ECG signal is input into the encoder of the trained model (with frozen weights), which generates 64 pairs [means, variances]. Here, we use the means as ECG features. As the algorithm can only learn one lead ECG signal, each lead was trained separately, and the learned features were combined (with an embedding size of 768). In Chapter 5, we use these ECG features, along with age and sex, to train the ISD models. Here, For simplicity, we call this approach **RN-β-VAE-lead#**. (Note that we use the # sign to reflect the lead number that is used to train this model.)

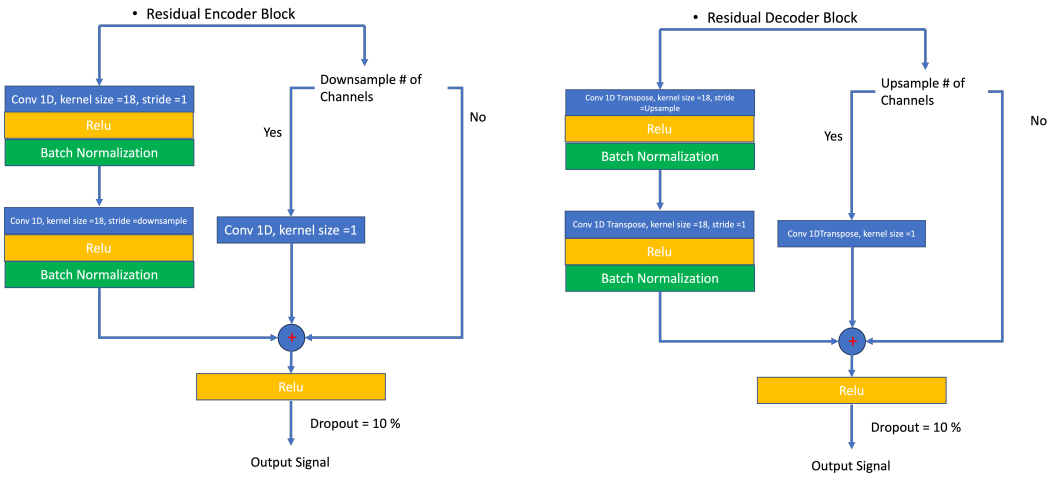

Figure 3.4: Building blocks of Residual Encoder and Residual Decoder used in Resnet based β-VAE.

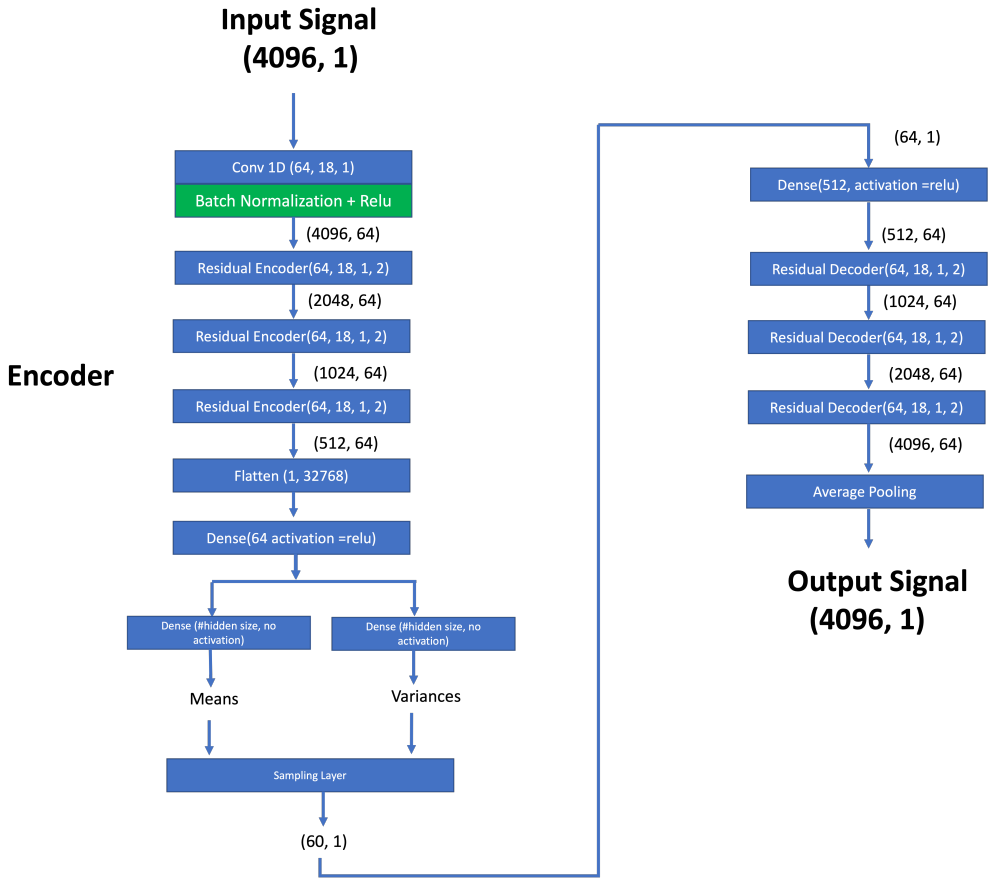

Figure 3.5: Model's Architecture of Resnet based β-VAE.

3.2.2 Supervised Models

Gradient Boosted Tree Ensembles (XGB) Model

Chapter 4 will first show the process of learning the TCN-Based β-VAE using the Alberta ECG dataset. Then, we will extract the embeddings for each ECG signal using the trained model. Afterwards, we will evaluate the quality of the learned embeddings by applying them to the task of the multi-label classification of 15 cardiovascular diagnoses- each for one versus all classifications of each label of Alberta ECG dataset. Additionally, we will use the sex (binary Male=1, Female=0) and age (continuous feature) associated with each ECG signal as additional features for this task. To this end, we will use the XGB model [44], where we will use hyperparameter values of max_depth = 3, and the number of estimators = 200. (Note that we did not perform any hyperparameter tuning to select values for these hyperparameters.). We limit the learning process of the models to a maximum of 200 epochs, and if there is no improvement in the loss of the validation set for ten consecutive epochs, we stop the learning process.

InceptionTime Model

In Chapter 4, for multilabel classification of ECG abnormalities of the CPSC 2018 dataset, we will use InceptionTime model [23], based on the TSAI [39] public library that implemented many state-of-the-art algorithms for time series tasks. To capture the role of generated ECGs on the performance of the classifier, we design and run multiple experiments that use various data augmentation techniques; see Table 3.2. First, we measure the performance of the classifier without any data augmentation method as a baseline experiment (CPSC_NA). Then, we augment the train+validation dataset with various augmentation; ECGs from Alberta dataset as a positive control for the contribution of augmented ECGs (AB_Orig_STE), VAE synthetic ECGs as a target experiment to evaluate the effect of ECG generated data (ABVAE_Gen_STE), and oversampled CPSC ECGs as a negative control (CPSC_OS_STE). (Note that we use random replacement sampling from the training set of CPSC 2018

dataset to generate the desired oversampling ECGs). These generated ECG signals were then added with a ratio of 90% into training and 10% into the validation set of the CPSC 2018 Dataset. For a fair comparison, we fixed the test set for all experiments. The scoring for our experiments is based on the F1 measure. For each 9 possible labels, the F1 score is defined as follows:

$$F1_i = \frac{2 \times N_{ii}}{N_{iX} + N_{Xi}} \tag{3.1}$$

for $i = 1, ..., 9$, where N_{ii} is the number of correctly classified cases for the i^{th} label, N_{iX} is the total number of cases predicted as the i^{th} label, and N_{Xi} is the total number of cases that are actually of the i^{th} label. During training, the InceptionTime model was run for 200 epochs with an early stopping mechanism. If there is no improvement in the average F1 score across all labels on the validation set for 50 consecutive epochs, training is stopped, and the model is saved for the inference stage on the test set.

Table 3.2: The experiments designed to capture the role of generated ECGs in the InceptionTime classifier performance.

Experiment Name	**Training Sample Size (# of ECGs)**	**Description**
CPSC_NA	6190	No data augmentation as baseline experiment
AB_Orig_STE	6190 + 1072 AB ECGs	1072 real ECGs with STE from Alberta dataset as a positive control experiment
ABVAE_Gen_STE	6190 + 1072 VAE generated From ABVAE	1072 AB VAE generated ECGs with STE abnormality as a target experiment
CPSC_OS_STE	6190 + 1072 oversampled ECGs from CPSC	1072 oversampled ECG from CPSC dataset as a negative control experiment.

ResNet-Based Model

In Chapter 5, we use another deep learning approach to obtain ECG features for estimating patient-specific ISD: the ResNet Based Model. We employ pre-trained model architectures developed in our group, as explained by Sun et al. [51], for multilabel classification of International Classification of Diseases, 10th revision (ICD-10) codes. (Note that total number of labels are 1414.) Details architecture/training process of their model is explained in their paper [51]. In short, their ResNet model consists of a convolution layer, 4 residual blocks, and a dense layer, following each convolution layer of the network with batch normalization [22], Relu activation function, and dropout [48] to reduce the chance of overfitting. Additionally, they used the 12-lead ECG signals (4096×12 [number of data point, number of ECG leads]) as input signals into the network. We will use the predicted probabilities generated by this model for raw ECG signals as supervised ECG features (embedding size of 1414)(Figure 3.1-b). These features along with age and sex were used to train the ISD models. Here, for simplicity, we call this approach **RN-ICD**.

End-to-End model

We will use raw labeled ECG signals to directly train ISD models. Each raw ECG signal consists of 4096 real values data points that we use as input to this model. To reduce the dimensionality of raw ECG signals, we can choose any number of layers, fully connected or convolutional, before entering the ISD models. Here, we will use the ResNet architecture developed by Sun, et al. [51]. Then, the output of those layers along with age and sex will be fed into the ISD modes (Figure 3.1-a). Here, for simplicity, we call this approach End-To-End Deep Learning (**E2E DL**).

3.2.3 Cox-Proportional Hazard (COX-PH) Model

The Cox Proportional Hazards Model [27] is one of the well-known statistical methods used for survival analysis. In Chapter 5, we will use the Cox Proportional Hazards Model to estimate patient-specific ISDs as a function of age,

sex, and ECG features obtained from various models.

3.2.4 Neural Multi-Task Logistic Regression (N-MTLR)

Neural-MTLR (N-MTLR) [12] is a modified version of Multi-Task Logistic Regression (MTLR) [63], which passes the data to multiple neural networks, either deep or shallow, before entering the MTLR model. We can consider MTLR as a series of logistic regression (LR) models, where each LR model estimates the survival probability at each time interval. To learn a MTLR model, we first divide the entire time horizon into m time bins. As explained by Yu et al [63], for each time bin, we can define the survival probability as follows:

$$P_{\boldsymbol{\theta_i}}(T \geq t_i|\mathbf{x}) = (1 + \exp(\boldsymbol{\theta_i} \cdot \mathbf{x} + b_i))^{-1}, \quad 1 \leq i \leq m \tag{3.2}$$

where T is the time, $\mathbf{x}$ is the individual's features, and the parameters vector $\boldsymbol{\theta_i}$ and the thresholds b_i are specific to a given time. The binary labels, $y_i = [T \geq t_i]$, can vary based on the value of the threshold t_i. We represent a patient's survival time, denoted as d, as a binary sequence $y = y(d) = (y_1, \ldots, y_m)$. In this sequence, each element y_i can either be 0 or 1, indicating the patient's survival status at time t_i. Here, y_i is set to 0 if death has not occurred by the time t_i when i.e. $t_i < d$. On the other hand, y_i is set to 1 when $t_i \geq d$. There are $m + 1$ valid sequences that take the form of $(0, 0, \ldots, 1, 1, \ldots, 1)$, which includes both the sequence consisting entirely of zeros and the sequence consisting entirely of ones. The probability of observing a specific survival status sequence of $Y = (y_1, \ldots, y_m)$ can be estimated as follows:

$$P_{\boldsymbol{\Theta}}(Y|\mathbf{x}) = \frac{\exp\left(\sum_{i=1}^{m} y_i(\boldsymbol{\theta}_i\mathbf{x} + b_i)\right)}{\sum_{k=0}^{m} \exp(f_{\boldsymbol{\Theta}}(\mathbf{x}, k))} \tag{3.3}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ and $f_{\boldsymbol{\Theta}}(\mathbf{x}, k) = \sum_{i=k+1}^{m}(\boldsymbol{\theta}_i\mathbf{x} + b_i)$ for $0 \leq k \leq m$ represents the score of the sequence when an event takes place within the time range $[t_k, t_{k+1})$. For more details on the optimization process, please refer to Yu, et al. [63].

Here, we use N-MTLR as a state-of-the-art algorithm to estimate ISD curves using ECG features obtained by various approaches described earlier. The ECG features, along with age and sex, serve as input features for the N-MTLR model, and the performance is evaluated using various metrics, including the concordance index (C-index), Marginal L1 loss [17], and the Integrated Brier Score [16]. Please refer to Section 3.4.2 for a description of these metrics. We choose the C-index as the deciding metric to select the best model and ECG features.

3.3 Data Generation Using Learned TCN-Based β-VAE Model

Generating new data with specific characteristics is a powerful feature of VAEs. For instance, in scenarios where the goal is to produce more examples with a specific label, VAEs can generate new samples that are similar to the ones present in the dataset, but not identical. In the case of ECG signals, VAEs can be used to generate synthetic signals resembling the input ECG signals while retaining the specific characteristics or abnormalities of the input signals.

After we have learned a set of 64 β-VAE parameters for each cardiovascular diagnosis (32 [mean, variance] pairs) from Alberta ECG Dataset, we will then use this learned model to generate new (realistic) synthetic ECG signals. During data generation (Figure 3.6), we froze the layers weights of both the encoder and decoder. Then, we feed selected 12-lead ECGs X with a specified abnormality into the encoder. Then, using the means and variances produced by the encoder, we draw a sample (Z) from the Gaussian distribution and feed it into the decoder. This generates a new 12-lead ECG (associated with the same abnormality as the one fed into the encoder), which we can then use to augment the CPSC 2018 dataset. (Note we give this instance Z the same age and sex as the original instance X.)

Using this framework, we can generate an unlimited number of ECGs from a single ECG data inputted into the encoder. However, we generate 1 ECG sample from each original ECG signal.

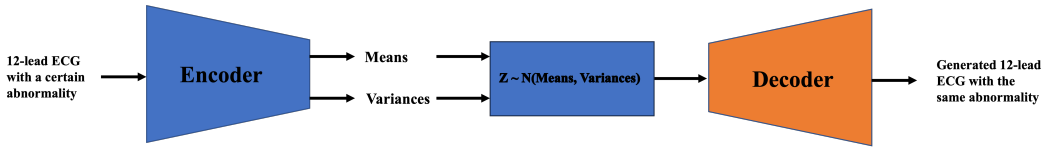

Figure 3.6: Schematic of ECG generation using trained β-VAE. The weights of both the encoder and decoder layers were fixed during data generation.

3.4 Evaluation Metrics

3.4.1 Evaluation of TCN-Based β-VAE Embeddings for Cardiovascular Diagnosis

To evaluate the quality of learned embeddings, we use the embeddings for the task of multi-label classification of cardiovascular diagnosis of Alberta ECG dataset. To generate the 32 ECG embeddings (Figure 3-1-d), we feed the 12-lead ECGs of Alberta ECG dataset into the encoder section of the trained TCN-Based β-VAE model, which produces 32 [mean, variance] pairs, which we use the mean to represent the signal. Afterwards, we run the gradient boosted tree ensembles (XGBoost) model on these instances using Alberta ECG dataset (from the train + validation set), along with age and sex, to learn 15 models – for one versus all classifications of each label. This involves creating a separate binary classifier for each label, which provides a probability score indicating the likelihood of the instance belonging to that label. Using a predetermined threshold for probability scores, we can predict whether the sample belongs to a specific label or not. (Note this means a single instance can be positive for several different diseases.) Then, for each label, we can create the confusion matrix, which consists of the following four cells, each representing the number of some specified instances:

- True Positive (TP) is the case where the model predicted *positive* and the actual value was *positive*.
- True Negative (TN) is the case where the model predicted *negative* and the actual value was *negative*.
- False Positive (FP) is the case where the model predicted *positive* but

the actual value was *negative*.

- False Negative (FN) are the cases where the model predicted *negative* but the actual value was *positive*.

Accuracy is a metric that is commonly used to evaluate the performance of a classifier. However, in cases where the dataset is imbalanced or when the cost is different for FP and FN, it can be misleading. Hence, we will use a threshold dependent metric called Youden's index [62], $J = \text{Sensitivity} + \text{Specificity} - 1$, which represents the difference between the TP rate (sensitivity) and the FP rate (1-specificity). A value of 1 shows perfect discrimination, while a value of 0 indicates a performance equal to random guessing, and negative values shows that the performance is worse than chance.

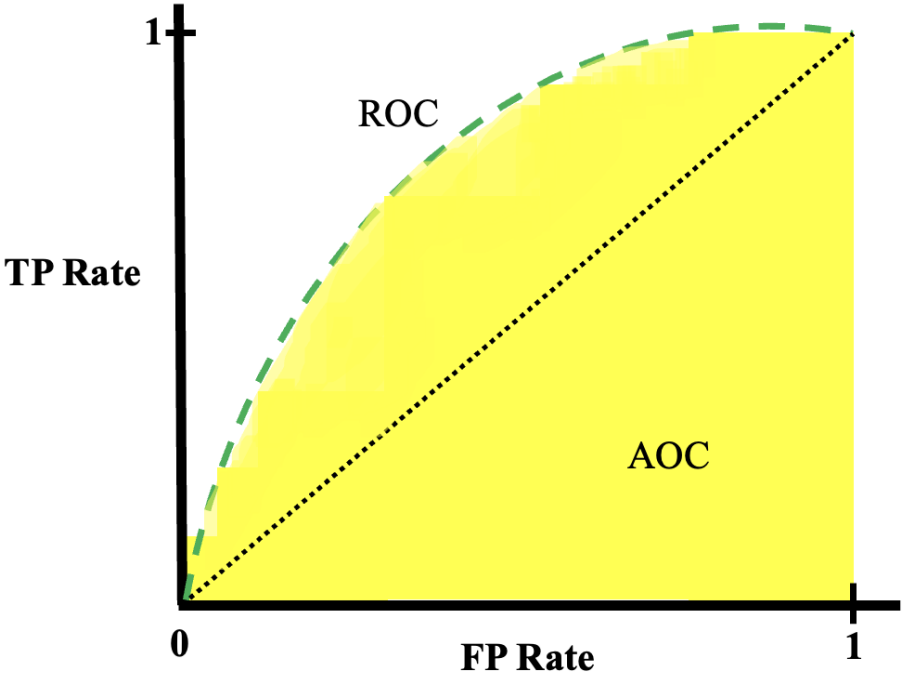

Figure 3.7: The plot illustrates the AUROC (Area Under the Receiver Operating Characteristic) metric. On the x-axis, we have the FP (False Positive) rate, and on the y-axis, the TP (True Positive) rate. The curve represents the ROC (Receiver Operating Characteristic), and the area enclosed by it is the AOC (Area Under the Curve).

To evaluate the model's performance, we used the two metrics - Area Under

the Receiver Operating Characteristic Curve (AUROC), which is a common metric for evaluation of classifiers, and F1 score. We are using these metrics because the AUROC provides a single measure of overall model performance that evaluates how well the model distinguishes between classes across all thresholds. To calculate AUROC, one must compare the TP rate (sensitivity) with the FP rate (1-specificity) for a range of probability thresholds (Figure 3.7). The resulting curve is then analyzed by calculating the area underneath it, which can range from 0 to 1. A value of 0.5 signifies that the model's performance is equivalent to random guessing (dotted line in Figure 3.6), whereas a score of 1 represents perfect discrimination between positive and negative classifications. The F1 score combines precision and recall to provide an overall measure of the model's accuracy. The F1 score is calculated by taking the harmonic mean of precision and recall values for the model on the test set, and can range from 0 to 1. The Precision, Recall and F1 score formula is shown below.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.5}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.6}$$

In the case of ECG abnormality detection, a high F1 score indicates that the model effectively balances precision and recall. A large F1 score, approaching 1, suggests high model performance in identifying true abnormalities (precision) while minimizing the risk of overlooking actual cases (recall). Conversely, a low F1 score indicates poor performance, where the model may either be missing too many true cases (low recall) or incorrectly identifying normal ECGs as abnormal (low precision).

3.4.2 Evaluation Metrics of Survival Prediction.

In Chapter 5, we estimate the ISD using ECG features and using each of Cox-PH and N-MTLR models. We used various metrics to compare the model's performances described in detail in Haider et al. [17]. Here, we briefly describe these metrics.

C-index

C-index (aka Concordance) is a well-known metric used to evaluate the performance of a risk model. The C-index measures how well the model can discriminate between individuals with different risk levels. Calculation of C-index starts by identifying the set of all comparable pairs. The metric then calculates the percentage of pairs that are correctly predicted. A pair is considered to be concordant when the individual with the shorter observed survival time also possesses a shorter predicted survival time, which is defined as the time corresponding to the median of their ISD curve, as per the ISD model. For example, if there are two uncensored individuals, A and B, and it is observed that patient B will survive longer than patient A, the model will calculate the median survival time for each patient, tmA and tmB. If tmA is less than tmB, the model's prediction that patient B will live longer is considered correct. Conversely, if tmA is greater than tmB, the prediction for this pair is deemed incorrect. The formula to calculate C-index is defined as following:

$$C\text{-index} = \frac{\sum_{i,j} \left(1_{T_j<T_i} \cdot 1_{\eta_j<\eta_i} \cdot \delta_j\right)}{\sum_{i,j} \left(1_{T_j<T_i} \cdot \delta_j\right)} \tag{3.7}$$

where η_i is the risk score of individual i, $\delta_i \in \{0, 1\}$ indicates if the i-th patient is dead (1) at that time T_i, or is censored (0). The range of the C-index varies from 0 to 1. The C-index value of 0.5 indicates the baseline, randomly assigning probabilities to instances would result in a 50% probability of correct ordering. A higher C-index value shows a better model performance.

Marginal L1-loss

To compute this metric, it is necessary to have the actual event time to compare the difference between the predicted and actual survival times. For uncensored patients, the actual event time (death) is known, but for censored patients, the survival time is estimated based on the expected survival time calculated using the Kaplan-Meier (KM) method. The difference between the predicted survival time and the actual survival time is then expressed as a marginal L1 loss. The marginal L1 loss quantifies the average deviation of the predicted survival times from the actual event times, with a lower value indicating more accurate predictions. To this end, for each censored individual, we will define a "Best-Guess" value, representing the individual's expected survival time given that s/he already survived until the censor time c.

$$BG(c) = c + \frac{\int_c^\infty S(t)\,dt}{S(c)} \tag{3.8}$$

where $S(.)$ is the survival function, which we estimate using KM generated from the training set. Using this $BG(c)$, we can calculate the $L1$-marginal loss as follows:

$$L1_{\text{margin}}(D, \hat{t}^{0.5}) = \frac{1}{\gamma}\left[\sum_{j \in D_{\text{uncensor}}} \left|d_j - \hat{t}_i^{0.5}\right| + \sum_{k \in D_{\text{censor}}} \alpha_k \left|BG(c_k) - \hat{t}_k^{0.5}\right|\right] \tag{3.9}$$

where $\gamma = |D_{\text{uncensor}}| + \sum_{k \in D_{\text{censor}}} \alpha_k$, and α_k is the weight in each estimate based on the Best-Guess for each individual, and d is the true event time for each uncensored patient. The term $\hat{t}_i^{0.5}$ represents the median survival time predicted by the ISD model for each individual. We set $\alpha_k = 1 - S(c_k)$ to place more weight on the late censor time instances. The reason for such a weight definition as explained by Haider et al. [17] is that individuals with early censor time give less information compared to those individuals with late censor time.

Integrated Brier Score

Brier Score [8] measures the mean squared error between the prediction made by the model and the actual event status (0 or 1) for a given time. If all data is uncensored, the Brier score at time t for a such dataset (D) is as follows:

$$BS_t(D, \hat{S}(t|\vec{x})) = \frac{1}{D} \sum_{(\vec{x}_i, d_i) \in D} \Big(I|d_i \leq t| - \hat{S}(t|\vec{x}_i) \Big)^2 \tag{3.10}$$

where $I|d_i \leq t|$ is an indicator function that is 1 if the actual event time d_i is less than or equal to t and 0 otherwise, and $\hat{S}(t|\vec{x}_i)$ is the predicted survival probability at time t for the covariates $\vec{x}_i$. We can extend the Brier score to a series of time points using Integrated Brier Score (IBS), which estimates the mean Brier score over the time interval.

$$IBS(\tau, D, \hat{S}(t|\vec{x})) = \frac{1}{\tau} \int_0^{\tau} BS_t(D, \hat{S}(t|\vec{x}))\, dt \tag{3.11}$$

Here, τ is the maximum event time of the combined dataset. If the model accurately predicts all time points, the score will be 0, and if the model always predicts 0.5, the score will be 0.25. So, a lower number indicates a better ISD model. The formula presented here assumes that we do not have any censored individual. To handle the censored individual, Graf et al. [16] suggest employing the *Inverse Probability of Censoring Weights* (IPCW) approach, where the instances subject to censoring are weighted equally to the uncensored instances. For more detailed description, please see Graf et al [16].

3.5 Model Comparison Using Bootstrapping

To determine if the results of the multilabel classification of ECG abnormalities of CPSC 2018 are statistically significant, we used the bootstrapping method. Bootstrapping is a statistical resampling technique used to estimate the properties of an estimator (such as its variance) by repeatedly sampling with replacement from the data set. It involves generating multiple bootstrap samples, each of which is the same size from the same test set. This process builds an empirical distribution of the statistic, allowing for the estimation of its variability, confidence intervals, and other properties. It provides

a non-parametric approach to statistical inference, enabling relatively robust estimation of statistical parameters and hypothesis testing without relying on specific distributional assumptions [37]. The steps that we take to compare models began by sampling the test set with bootstrapping. We created 10,000 samples from our test set, where each sample was formed by random replacement sampling. Then, for each sample set and label, we calculate the difference in F1 score between pairs of models. Afterwards, we calculate the mean difference in F1 score along with the 95% confidence intervals for these differences. Finally, we evaluated whether the observed differences were statistically significant or not. To decide whether the differences are statistically significant, we observed whether the 95% confidence interval of the difference in means included zero. If it did not, we considered the difference to be statistically significant. Our hypothesis testing in our case would be as follows:
Null Hypothesis: There is no significant difference in the F1 scores between the models for each label and model pair, implying that any observed difference was due to random variation.
Alternative Hypothesis: There is a significant difference in the F1 scores between the models for each label and model pair.

In Chapter 5, and for each training sample size and ECG features, we use 10 different random splitting training sets to train 10 models and plot the mean of the performance, with error bars reflecting the 95% confidence interval. We then presented the plots for three metrics of C-index, L1-Marginal and IBS.

Chapter 4

Generative Data by β-Variational Autoencoders Help Build Stronger Classifiers: ECG Use Case

A widespread tool currently used for the diagnosis of cardiovascular diseases is Electrocardiogram (ECG). However, detecting cardiac abnormalities through ECG is not easy and currently requires an expert. With the advancement of machine learning in healthcare, many researchers are now exploring ways to learn end-to-end diagnostic models using ECGs [21], [45], [61]. The open-source ECG data from the China Physiological Signal Challenge 2018 (CPSC 2018) has helped researchers develop various machine-learning models for ECG abnormalities/ diagnosis. However, the prediction performance of these models is not high for all labels. Adding more training ECG instances (either of real patients or artificially generated) of the labels might help the learning algorithm produce a more accurate model, and probably more accurate for the labels of those additional instances.

Since the ECG is one of the most common measurements and is routinely used during hospital admission, hospitals record ECG scans of many patients with various heart conditions/anomalies. However, due to the need to protect patients' privacy and confidentiality, these data often cannot be shared. However, using this private data set, we might be able to produce ECGs with certain abnormalities using a generative model, such as variational autoen-

coders (VAE), while retaining the privacy of health data. We can then add these generated ECGs to our current real labelled ECG training set to produce a model that potentially improve the performance on the test set. (Note that no synthetic ECGs are added to the test set.)

In recent years, some studies have used VAE to generate ECG instances, then use these intermediate learned embeddings of a trained VAE to learn models that could predict 1-year mortality or the type of ECG abnormality [52]. The advantage of these approaches over deep learning models is that one can explain the model's prediction. These explanations are example based and can be achieved in two steps. (1) Initially, an explainable artificial intelligence (XAI) method such as SHAPLEY [58] or LIME [46] can be used to find the important embeddings (learned mean of VAE) for certain predictions of the downstream task. (2) Subsequently, we can calculate the correlation between these embeddings and knowledge-based features using the dataset. For instance, suppose an XAI method shows that an embedding at index 3 is critical for certain predictions made by the model. In that case, examining the correlation map between this embedding and knowledge-based features may reveal a strong correlation with a specific feature, such as P wave duration. These explanations, when they resonate with the insights of domain experts, will improve the confidence in the model's prediction by the clinicians who are going to use it.

To diagnose an ECG abnormality, it is useful to consider both the morphologies of a single beat (such as R peaks, presence of P wave, etc.) and the rhythm (combination of multiple beats) as shown by Berkaya et al. [6]. In this regard, Jang et al. [24] used unsupervised convolutional VAE to encode input ECGs into 60 features through the reconstruction of lead II of ECG (both morphologies of single beat and the rhythm) collected from 1278 patients. van de Leur et al. [31] learned a VAE (from 1.1 million ECGs) to encode 12-lead ECG signals of a single beat into 21 learned features, then used these learned features for downstream tasks of detection of reduced ejection fraction, and 1-year mortality. They also correlated the learned 21 features with conventional electrocardiogram measurements generated by the ECG measurement

device during its collection to provide further explanation about VAE embeddings. However, the focus of both studies was on using the extracted ECG embedding for a downstream task. To the best of our knowledge, no study explored ways to generate synthetic ECGs using VAE and used them as a data augmentation method. These studies reached good arrhythmia classification performance using VAE-encoded features from ECGs. However, their models can generate either multiple beats (rhythm) of a single lead or a single beat of a 12-lead ECG signal. Both morphologies and rhythms of ECG signals are important for the diagnosis of ECG abnormalities as different abnormalities express their characteristics in different leads or they are related to rhythm rather than the morphology of a single beat [33]. In this Chapter, we use a generative model (TCN based β-VAE) to learn the rhythm of 12-lead ECG signals using a large dataset of 244,077 patients admitted to hospitals in Alberta, Canada between February 2007 and April 2020, where each is labeled with certain cardiovascular diagnoses identified with specified ICD-10 codes. As explained in Chapter 3, we then use this trained model to generate new ECGs, each with one or more of these specified diagnoses. We then identified ways to use this Synthetic ECGs to improve the performance of multi-label classification, using publicly available 12-lead ECG CPSC dataset with various abnormalities.

Figure 4.1 displays the overall methodology being used in this chapter. In short, we are initially training the TCN-based β-VAE model using the Alberta ECG dataset. Afterward, we are evaluating the quality of β-VAE's learned embeddings using the Alberta ECG Dataset. In this evaluation, we utilize the ECG embedding along with age and sex for the multilabel classification of 15 cardiovascular diagnoses from the Alberta ECG dataset. Then, we will evaluate different data augmentation methods based on the downstream prediction error of the classifiers learned using that data, for the task of multi-label classification of ECG abnormalities of CPSC 2018. Here, we consider using ECGs generated by a model learned from Alberta ECGs, data addition of real Alberta ECGs, and oversampling of ECGs of CPSC 2018 dataset.

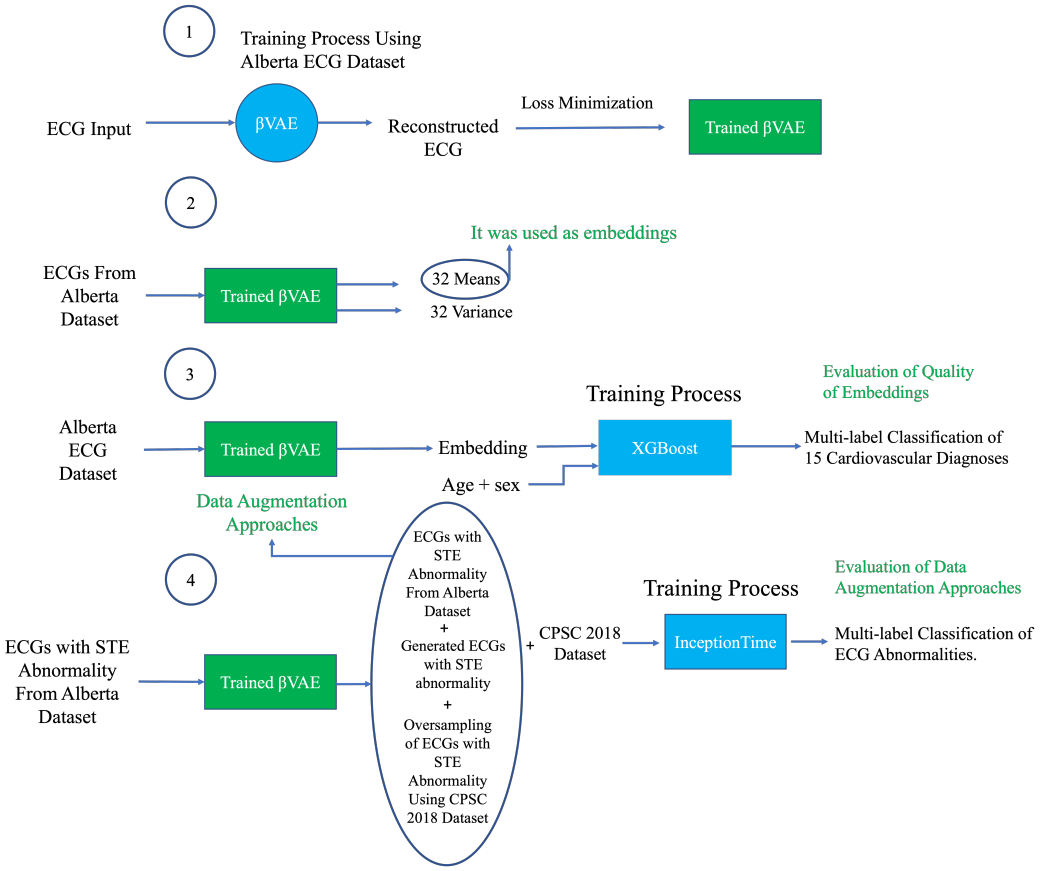

Figure 4.1: Overall methodology used in this Chapter: (1) β-VAE was trained using Alberta dataset. (2) For each ECG of the Alberta ECG dataset, we selected the 32 embedding obtained from the trained encoder of β-VAE (3) Quality of embeddings was evaluated using the multi-label classification of 15 cardiovascular diagnoses. (4) We compared various data augmentation methods (data generation, data addition, and oversampling) using CPSC 2018 dataset for the task of multi-label classification of ECG abnormalities.

4.1 Results

In the following, we first provide the result of β-VAE training and the quality of its learned embeddings on Alberta ECG dataset. Then, we generate different levels of ECGs with certain abnormalities and evaluate the role of these synthetic ECGs on the performance of a classifier for the task of multi-label classification using CPSC 2018.

4.1.1 Performance of Multi-Label Classification of Cardiovascular Diagnoses of Alberta ECG Dataset

The primary objective of our study was to explore the capability of TCN-based β-VAE in reconstructing Alberta 12-lead ECG signals and to assess the ability of the learned embeddings to capture meaningful information for downstream tasks. Figure 4.2 shows an example of ECG signal reconstructed by the β-VAE model. We also calculate the Pearson correlation coefficient of these 32 embeddings with 22 knowledge-based ECG features (Figure 4.3). These correlations can provide an explanation for the characteristics of learned embeddings and their relation with well-defined knowledge-based ECG features. Due to the unsupervised nature of β-VAE, the model might learn the characteristics of some ECG labels more than other labels depending on the number of training data. To evaluate the quality of learned ECG features for different labels, we used these extracted features with the addition of age and sex for the task of multilabel classification of cardiovascular diagnoses. Also, we used the knowledge-based ECG features along with age and sex for the same task. We used XGBoost to create 15 independent models, one for each type of diagnosis (Table 4.1). The performance of β-VAE features was lower than knowledge-based ECG features. It is important to note that the focus of this comparison is not to establish the superiority of one method over the other but rather to demonstrate the potential of the β-VAE model in learning useful information (for a downstream task) and in its ability to generate synthetic ECGs. We acknowledge that a statistical significance test comparing the two methods was beyond the scope of this study, primarily due to computational constraints. However, the large size of our test set (640,527 ECG signals from 97,631 patients) provides a level of reliability to our findings.

The performance of β-VAE features varied among labels. If we select AUROC = 0.70 as a threshold for reasonable learning performance, 9 labels – ST Elevation Myocardial Infarction (STEMI), Heart Failure, Unstable Angina, Atrial Fibrillation, Ventricular Tachycardia, Atrioventricular Block, Pulmonary Hypertension, Hypertrophic Cardiomyopathy, and Hypertrophic

Cardiomyopathy – have the performance above this threshold, suggesting that based on this learned TCN-based β-VAE, these labels might be more suitable candidates for data generation as compared to other 6 labels.

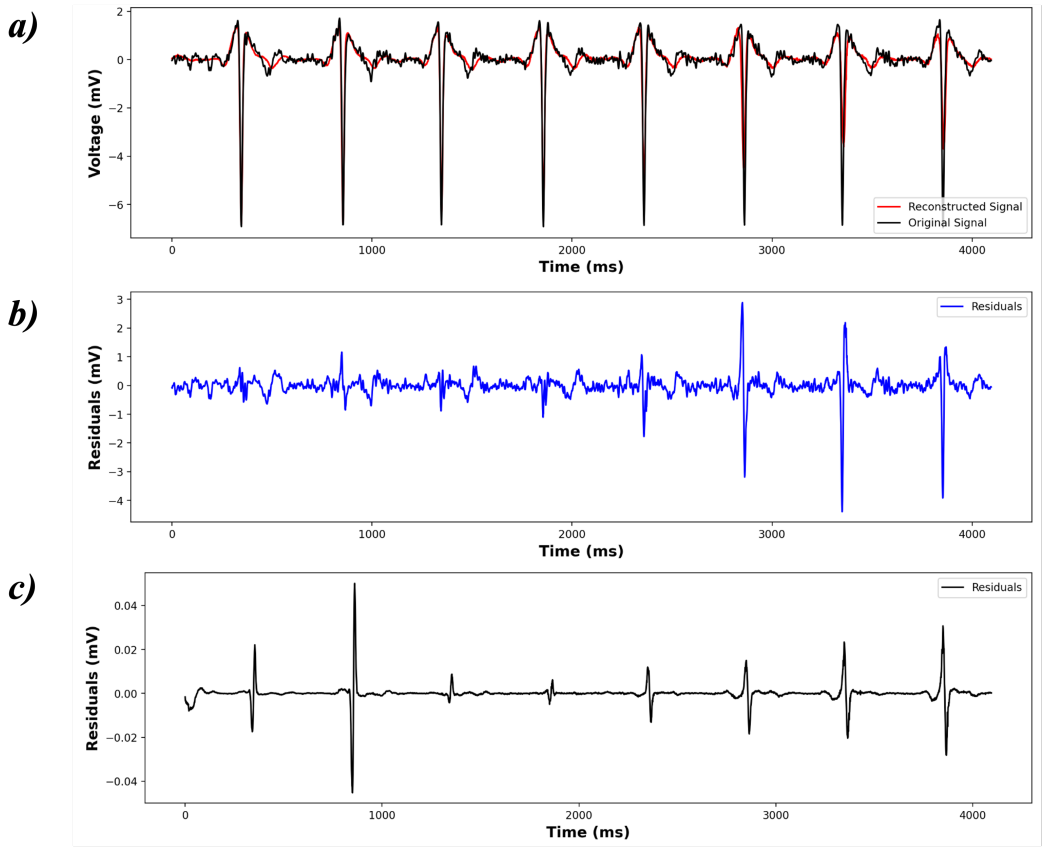

Figure 4.2: a) An example of reconstructed signal for lead 4 of an ECG signal. The black color shows the original ECG signal and the red color shows the reconstructed ECG signal. b) The residual between the reconstructed signal and the original signal. c) the residual between the two ECGs that were both generated from the same original ECG signal.

4.1.2 Performance of Multi-Label Classification of ECG Abnormalities for CPSC 2018

Since CPSC 2018 ECG dataset is collected from 11 hospitals, the distribution of the CPSC 2018 ECG dataset might be different than the distribution of the ECG dataset collected from Alberta Hospitals (14 hospitals). With the addition of synthetic ECGs from Alberta Dataset into CPSC 2018 Dataset, we might be able to improve the performance of the classifier.

Table 4.1: The multi-label classification of 15 cardiovascular diagnoses using the Alberta ECG dataset using 32 TCN Based β-VAE embeddings versus 22 ECG Global Measurments(GMs) from Philips Machine. All features/embeddings include age and sex as additional features.

Label	32 Embeddings		22 ECG GMs	
	AUROC	F1	AUROC	F1
NSTEMI	0.65	0.22	0.77	0.33
STEMI	0.74	0.25	0.88	0.49
Heart Failure	0.77	0.33	0.83	0.35
Unstable Angina	0.73	0.14	0.76	0.18
Atrial Fibrillation	0.75	0.31	0.72	0.18
Ventricular Tachycardia	0.72	0.09	0.77	0.12
Cardiac Arrest	0.64	0.06	0.74	0.09
Supraventricular Tachycardia	0.62	0.10	0.60	0.08
Atrioventricular Block	0.84	0.23	0.89	0.31
Pulmonary Embolism	0.61	0.04	0.69	0.11
Aortic Stenosis	0.68	0.05	0.80	0.09
Pulmonary Hypertension	0.70	0.05	0.77	0.11
Hypertrophic Cardiomyopathy	0.70	0.03	0.86	0.11
Mitral Valve Prolapse	0.62	0.02	0.72	0.04
Mitral Valve Stenosis	0.61	0.01	0.76	0.02

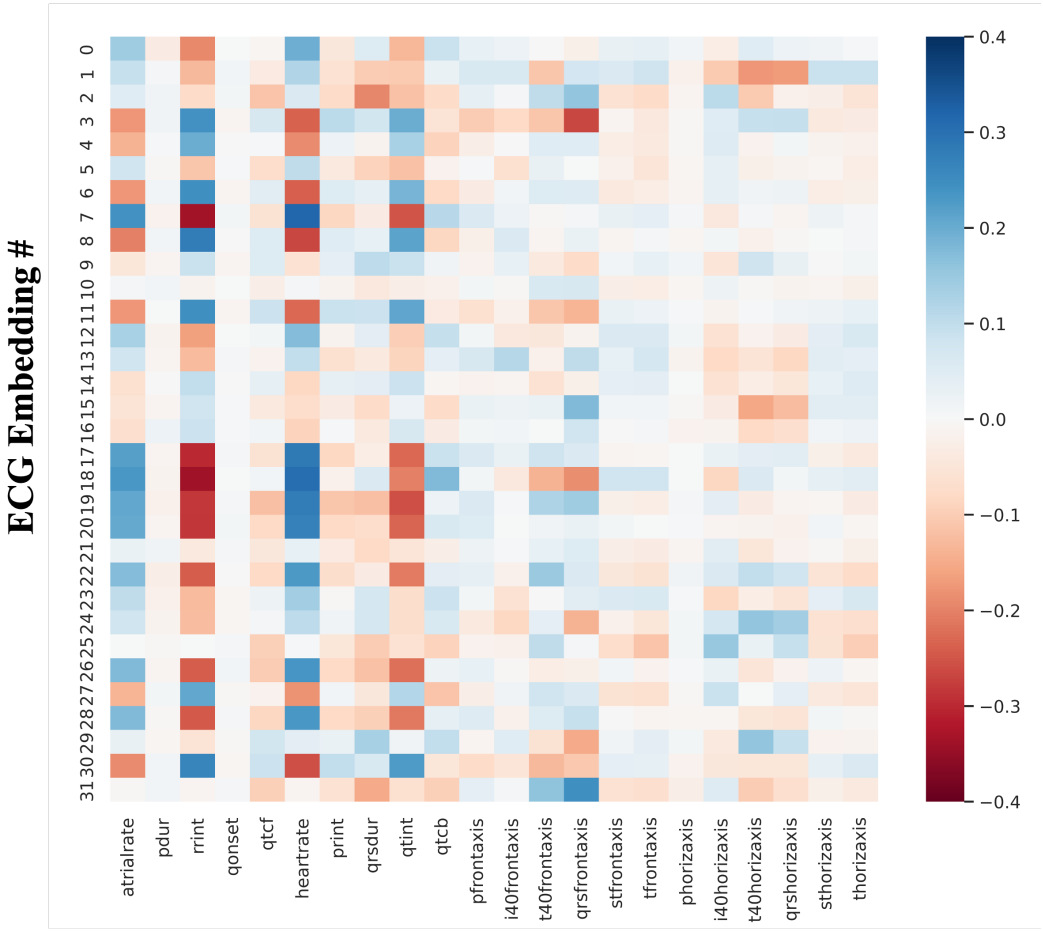

Figure 4.3: Heatmap illustrating the Pearson correlation coefficients between 32 learned embeddings and 22 knowledge-based ECG features in the test set of the Alberta ECG dataset. This visualization highlights significant positive and negative correlations, offering insights into the characteristics and interpretability of the embeddings. For the definition of knowledge-based features, refer to Table 2.1.

We used the InceptionTime model, described in Section 3.2.2, to classify ECG abnormalities of CPSC 2018; Figure 4.4 shows the models' performance on the test set. As a first data augmentation experiment (CPSC_NA), we selected the label with the lowest F1 score (STE), then selected new raw instances from the AB dataset. In particular, we used AB instances whose STEMI label was a negative diagnosis for all other labels (72 cases had this condition), and 1000 samples of STEMI that had a negative diagnosis for at

least the labels of the CPSC 2018 ECG dataset (AB_Orig_STE). Note that our dataset labels are limited to only 15, and the selected ECGs might have other possible abnormalities not covered by our set of labels.

The results presented in Tables 4.2 to 4.6 are derived from a pairwise statistical analysis conducted between different pairs of models. This analysis assesses the differences in the F1 performance between each pair of models across all ECG labels. To conduct this analysis, we first bootstrapped the test set 10,000 times. For each bootstrap sample, we measured the F1 performance of each model across all labels. We then calculated the difference in F1 scores between all possible model pairs for each label across all 10,000 bootstrap samples. This process yielded 10,000 F1 score differences for each label and model pair. Subsequently, we calculated the mean F1 score of these differences along with their 95% confidence intervals. The significance of these differences was determined based on whether the 95% confidence intervals included zero. If the interval did not include zero, the difference in performance for that specific label between the two models was considered statistically significant. For example, consider the results in Table 4.2, which compared the pairwise differences in F1 scores between the AB_Orig_STE and CPSC_NA models for each ECG label. We observe that for the STE label, the mean F1 score is 0.089, and the confidence intervals associated with this label do not include zero, indicating that the difference in F1 performance between the AB_Orig_STE and CPSC_NA models is statistically significant.

The results show that the addition of synthetic data (generation or augmentation) increased the F1 score performance of STE compared to the models trained on just the original dataset. The addition of raw AB ECGs of patients with STE (AB_Orig_STE) labels had the highest performance (0.0890[0.0597-0.1185] [1] for STE diagnosis compared with ABVAE_Gen_STE (0.0463[0.0267-0.0659]) or CPSC_OS_STE (0.0447[0.0135-0.0760]) approaches. However, data addition had a mixed effect on the model's performance of other labels. For the AB_Orig_STE experiment, we observed statistically significant changes in SR,

[1] mean pairwise difference in F1 scores followed by 95% confidence interval of the mean pairwise difference

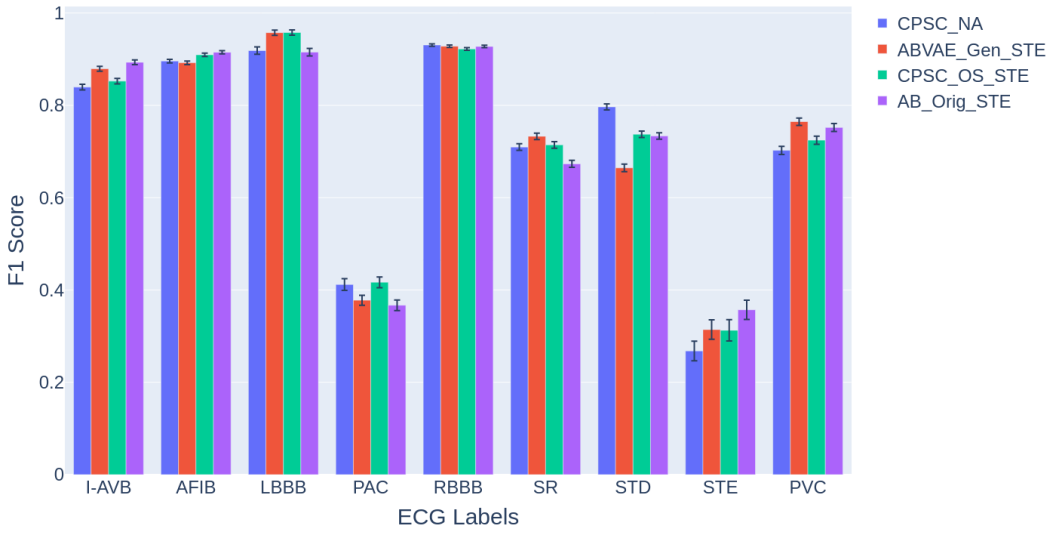

Figure 4.4: Model performance of classification of ECG abnormalities for CPSC 2018 dataset. The error bar shows the the upper and lower 95% confidence intervals. The x-axis represents ECG labels, including the 'no abnormality' label of SR (Sinus Rhythm), and 8 abnormalities: AFIB (Atrial Fibrillation), IAVB (First-degree Atrioventricular Block), LBBB (Left Bundle Branch Block), RBBB (Right Bundle Branch Block), PAC (Premature Atrial Contraction), PVC (Premature Ventricular Contraction), STD (ST-segment Depression), and STE (ST-segment Elevated). The legend corresponds to the experimental designs defined in Table 3.2. 'CPSC_NA' denotes the baseline experiment with no data augmentation; 'ABVAE_Gen_STE' represents the experiment with 1072 Alberta VAE generated ECGs with STE abnormality; 'CPSC_OS_STE' refers to the negative control experiment with 1072 oversampled ECGs with STE abnormality from the CPSC dataset; 'AB_Orig_STE' indicates the positive control experiment with 1072 real ECGs with STE abnormality from the Alberta dataset.

AFIB, 1-AVB, PAC, PVC, STD and STE, an statistically insignificant difference in LBBB and RBBB. AB_Orig_STE data has better performance in AFIB, 1-AVB, PVC, STE and worse in SR, PAC, and STD. For the AB_Gen_VAE, we observed statistically significant changes in the performance of SR, AFIB, 1-AVB, LBBB, PVC, STD, and STE labels. An insignificant difference in PAC and RBBB, a significant decrease in the performance for the STD label, and a slightly worse performance on AFIB labels. For the CPSC_OS_STE experiment, we observed a significant difference in performance for AFIB, 1-AVB,

Table 4.2: The pairwise differences between AB_Orig_STE and CPSC_NA models. 'AB_Orig_STE' indicates the positive control experiment with 1072 real ECGs with STE abnormality from the Alberta dataset. 'CPSC_NA' denotes the baseline experiment with no data augmentation. This table presents the results of a statistical analysis comparing the performance of two models, AB_Orig_STE and CPSC_NA, in classifying various ECG abnormalities. Each row corresponds to a different ECG label (e.g., SR, AFIB). The 'mean_F1' column shows the average difference in F1 scores between the two models across 10,000 bootstrap samples. A positive mean indicates that the AB_Orig_STE model generally performed better than the CPSC_NA model for that label, while a negative mean suggests the opposite. The 'CI_upper' and 'CI_lower' columns provide the upper and lower bounds of the 95% confidence interval for these mean differences. If this interval does not include zero, it implies that the difference in performance between the two models for that label is statistically significant.

Label	**mean_F1**	**CI_upper**	**CI_lower**
SR	-0.0363	-0.0277	-0.0469
AFIB	0.0191	0.0243	0.0138
1-AVB	0.0537	0.0618	0.0457
LBBB	-0.0033	0.008	-0.014
RBBB	-0.0030	0.0008	-0.0068
PAC	-0.0448	-0.0277	-0.0620
PVC	0.0497	0.0619	0.0376
STD	-0.0631	-0.0536	-0.0726
STE	0.0890	0.1185	0.0597

LBBB,SR, PVC, STD, and STE. Performance was not significant for RBBB and PAC. Oversampling significantly decreased the performance for STD, but did improve performance for AFIB, 1-AVB, LBBB, SR, PVC, and STE. Most performance increases were small but LBBB had significant improvement of 0.03 to 0.05 in F-1. These results suggest that adding either β-VAE, or real ECG signals of the AB dataset, to the training dataset led to models that had an overall better performance improvement compared with oversampling of STE ECG signals from the CPSC 2018 ECG dataset.

In addition to the aforementioned statistical tests, we also utilized Z-tests to evaluate whether the differences in F1 scores across model pairs and labels are significant. The Z-test is a statistical method designed to ascertain if there are notable differences between the means of two groups. This method

Table 4.3: The pairwise differences between ABVAE_Gen_STE and CPSC_NA models. 'ABVAE_Gen_STE' represents the experiment with 1072 Alberta VAE generated ECGs with STE abnormality, while 'CPSC_NA' denotes the baseline experiment with no data augmentation. This table compares the performance of these two models in classifying various ECG abnormalities. Each row corresponds to a different ECG label (e.g., SR, AFIB). The 'mean_F1' column shows the average difference in F1 scores between the two models across 10,000 bootstrap samples. A positive mean indicates that the ABVAE_Gen_STE model generally performed better than the CPSC_NA model for that label, while a negative mean suggests the opposite. The 'CI_upper' and 'CI_lower' columns show the upper and lower bounds of the 95% confidence interval for these mean differences. A confidence interval not including zero signifies a statistically significant difference in performance between the two models for that label.

Label	**mean_F1**	**CI_upper**	**CI_lower**
SR	0.0233	0.0296	0.0169
AFIB	-0.0037	-0.0006	-0.0069
1-AVB	0.0398	0.0449	0.0348
LBBB	0.0388	0.0442	0.0334
RBBB	-0.0025	-0.00001	-0.0005
PAC	-0.0340	-0.0296	0.0169
PVC	0.0623	0.0698	0.0548
STD	-0.1322	-0.1246	-0.1398
STE	0.0463	0.0659	0.0267

is particularly relevant for large sample sizes and assumes that the data is normally distributed. For each model pair and label, we calculated the Z-statistic, which quantifies the difference in sample means in relation to the variability within the data. The significance of this Z-test was assessed using a p-value threshold of 0.05. A p-value less than this threshold indicates that the observed difference is unlikely due to chance, leading us to reject the null hypothesis and conclude a statistically significant difference between the models. Upon analyzing the z-test results, all comparisons between model pairs across various labels revealed statistically significant differences, except for the labels of LBBB and STE when comparing the ABVAE_Gen_STE and CPSC_OS_STE models.

Table 4.4: The pairwise differences between CPSC_OS_STE and CPSC_NA models. 'CPSC_OS_STE' refers to the negative control experiment with 1072 oversampled ECGs with STE abnormality from the CPSC dataset, while 'CPSC_NA' denotes the baseline experiment with no data augmentation. This table compares the performance of these two models in classifying various ECG abnormalities. Each row corresponds to a different ECG label (e.g., SR, AFIB). The 'mean_F1' column shows the average difference in F1 scores between the two models across 10,000 bootstrap samples. A positive mean indicates that the CPSC_OS model generally performed better than the CPSC_NA model for that label, while a negative mean suggests the opposite. The 'CI_upper' and 'CI_lower' columns show the upper and lower bounds of the 95% confidence interval for these mean differences. A confidence interval not including zero signifies a statistically significant difference in performance between the two models for that label.

Label	mean_F1	CI_upper	CI_lower
SR	0.0046	0.0145	-0.0053
AFIB	0.0135	0.0188	0.0081
1-AVB	0.0130	0.0216	0.0044
LBBB	0.0390	0.0488	0.0293
RBBB	-0.0085	-0.0047	-0.0122
PAC	0.0049	0.0222	-0.0123
PVC	0.0221	0.0344	0.0098
STD	-0.0593	-0.0497	-0.0689
STE	0.0447	0.0760	0.0135

4.2 Discussion

We used β-VAE to develop a generative model of the rhythm of 12-lead ECG signals. To the best of our knowledge, this is the first study that was able to learn the rhythm of 12 lead signals using β-VAE. Using β-VAE, other studies were able to learn either the rhythm of 1-lead ECG signals [24] or 1 beat of 12-lead signals [31]. (Note they also used the learned embeddings for downstream tasks.) Here, using generated ECG data from learned β-VAE based on a large ECG dataset of Alberta Hospitals, we investigate the role of synthetic data to help learn models that can classify ECG abnormalities of CPSC 2018 ECG dataset. We focused on the comparison of the model's performance under different numbers of Alberta β-VAE generated ECGs, over-sampling of ECGs, and addition of new ECGs obtained from the Alberta Hospitals Dataset. Figure 4.5 shows the mean F1 score (%) differences between various

Table 4.5: The pairwise differences between ABVAE_Gen_STE and CPSC_OS_STE models. 'ABVAE_Gen_STE' represents the experiment with 1072 Alberta VAE generated ECGs with STE abnormality, while 'CPSC_OS_STE' refers to the negative control experiment with 1072 oversampled ECGs with STE abnormality from the CPSC dataset. This table compares the performance of these two models in classifying various ECG abnormalities. Each row corresponds to a different ECG label (e.g., SR, AFIB). The 'mean_F1' column shows the average difference in F1 scores between the two models across 10,000 bootstrap samples. A positive mean indicates that the ABVAE_Gen_STE model generally performed better than the CPSC_OS_STE model for that label, while a negative mean suggests the opposite. The 'CI_upper' and 'CI_lower' columns show the upper and lower bounds of the 95% confidence interval for these mean differences. A confidence interval not including zero signifies a statistically significant difference in performance between the two models for that label.

Label	**mean_F1**	**CI_upper**	**CI_lower**
SR	0.0187	0.0286	0.0088
AFIB	-0.0172	-0.0118	-0.0226
1-AVB	0.0268	0.0349	0.0188
LBBB	-0.0003	0.0079	-0.0084
RBBB	0.0060	0.0098	0.0022
PAC	-0.0390	-0.0231	-0.0549
PVC	0.0402	0.0519	0.0285
STD	-0.0729	-0.0621	-0.0836
STE	0.0016	0.0328	-0.0296

data augmentation approaches and the original, non-augmented CPSC model (CPSC_NA) for each label. We found that Alberta β-VAE generated ECGs with STE abnormality not only were able to improve the model's performance (F1 score) on the STE label of the test set but also improve the model's performance on 4 other labels. The performance of oversampling the STE label also improved the model's performance of the STE label, but its positive effect on the performance of other labels was less than β-VAE generated data. For the STE label, among Alberta dataset β-VAE generated ECG data and the addition of Alberta original ECG data, the Alberta original ECGs improved the model's performance of the STE label by $\sim$9 %, while the Alberta β-VAE generated ECGs improved it by $\sim$5 %. We assume this lower performance (of Alberta β-VAE generated ECGs compared with the AB original ECG

Table 4.6: The pairwise differences between AB_Orig_STE and CPSC_OS_STE models. 'AB_Orig_STE' indicates the positive control experiment with 1072 real ECGs with STE abnormality from the Alberta dataset, while 'CPSC_OS_STE' refers to the negative control experiment with 1072 oversampled ECGs with STE abnormality from the CPSC dataset. This table compares the performance of these two models in classifying various ECG abnormalities. Each row corresponds to a different ECG label (e.g., SR, AFIB). The 'mean_F1' column shows the average difference in F1 scores between the two models across 10,000 bootstrap samples. A positive mean indicates that the AB_Orig_STE model generally performed better than the CPSC_OS_STE model for that label, while a negative mean suggests the opposite. The 'CI_upper' and 'CI_lower' columns show the upper and lower bounds of the 95% confidence interval for these mean differences. A confidence interval not including zero signifies a statistically significant difference in performance between the two models for that label.

Label	**mean_F1**	**CI_upper**	**CI_lower**
SR	-0.0409	-0.0306	-0.0513
AFIB	0.0056	0.0106	0.0005
1-AVB	0.0407	0.0487	0.0328
LBBB	-0.0423	-0.0324	-0.0522
RBBB	0.0055	0.0093	0.0016
PAC	-0.0498	-0.0335	-0.0661
PVC	0.0276	0.0397	0.0156
STD	-0.0038	0.0062	-0.0138
STE	0.0444	0.0752	0.0135

data) is because the reconstructed ECGs were not perfect and there was some information loss. However, this reduction might be acceptable, as it means the ECGs used do not compromise the patients' privacy.

The beneficial effect of ECG data generation on the model's performance was previously introduced by other studies, which used generative adversarial networks (GAN) to generate synthetic ECG data. Wang et al. [56] used a modified version of GAN called auxiliary classifier generative adversarial network (ACGAN) to generate synthetic data. Their method requires first identifying the R peaks of the signal and concatenation of 5 generated heartbeats as a sample (12×1500, where the first number shows the number of lead, and the second number represents the number of data points). They used the CPSC 2018 dataset, where they segmented the original ECG data

into shorter lengths that resulted in 13754 samples rather than the original 6877 samples. Then, they selected 50 instances from each label as a test set. These generated ECG data improved the performance of the classifier in the test set for all labels, compared with models that were trained with no data-generated ECG. Since they segmented the original dataset into short lengths, we cannot directly compare our classifier performance with this study. Others also used GAN-based methods to generate synthetic ECG data and observe an improvement of data generation over their baseline model using other ECG datasets [34] [64].

There are some limitations associated with our study. While our β-VAE model was able to learn the 12-lead signals, the predictive ability of these learned embeddings was lower than 22 ECG Global Measurements. The focus needs to be shifted to finding algorithms that can better encode ECG domains. To enhance the encoding of ECGs, one approach could be the development of a multi-output architecture. In such a design, one output would focus on the unsupervised task of reconstructing the original signal, while another output would utilize the intermediate representation for supervised multi-label classification. This multi-output approach aims to enhance the embeddings by ensuring they are informative for both signal reconstruction and diagnostic classification. Successfully implementing this architecture could lead to embeddings with richer representations, potentially improving their effectiveness in various downstream tasks such as survival prediction.

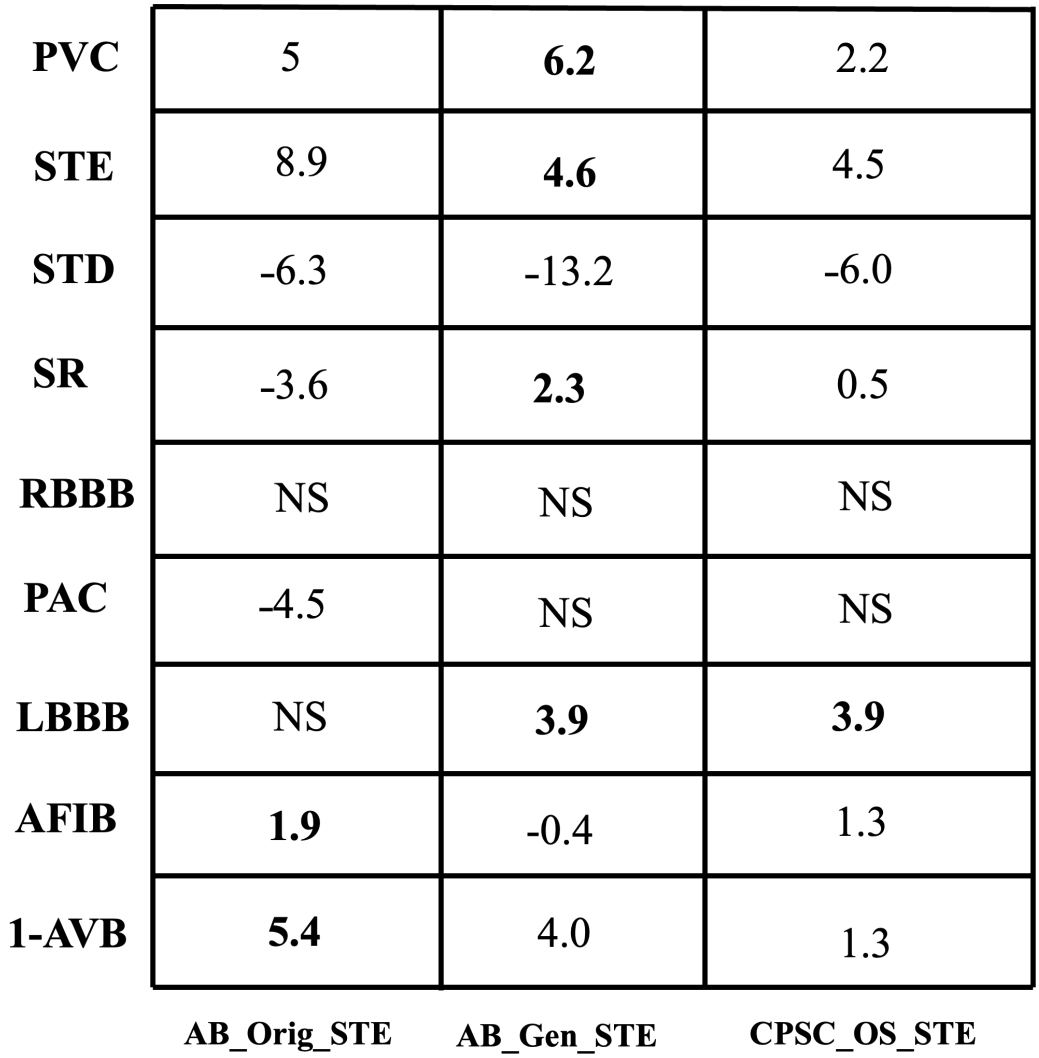

Figure 4.5: Mean F1 score (%) differences between various data augmentation approaches and the original real CPSC model for each ECG label. The x-axis categorizes the models compared against the non-augmented baseline model (CPSC_NA), while the y-axis lists the ECG labels. The numbers displayed indicate the magnitude of performance improvement or degradation for each label. Each row in the figure corresponds to a different ECG label, such as SR or AFIB. The numbers represent the average difference in F1 scores between the augmentation approach and the CPSC_NA model across 10,000 bootstrap samples. A positive mean F1 indicates better performance than the CPSC_NA model for that label, while a negative mean suggests lower performance. The statistical significance of these differences was evaluated based on the 95% confidence intervals. If a confidence interval does not include zero, it indicates a statistically significant difference in performance between the augmentation approach and the CPSC_NA model for that label. Results marked with 'NS' indicate that the performance difference was not statistically significant.)

Chapter 5

Supervised ECG features outperform knowledge-based and unsupervised features in individualized survival prediction

Heart abnormalities are one of the leading mortality causes in the world. In 2020, 19.05 million individuals died globally due to heart disease [54]. By identifying patients who are at a higher risk, we anticipate that we will be able to reduce the number of fatalities and related healthcare expenses. Additionally, this approach may help to direct limited resources to patients who have a greater chance of being treated. Electrocardiograms (ECGs) are a valuable and routinely collected measurement of heart health and have been successfully used along with age and sex to predict 1-year mortality with good results [52], suggesting ECGs contain the information needed for mortality prediction [43] Traditional risk assessment methods, such as the Cox Proportional Hazard model [10], offer time-independent risk scores. Models like those proposed by Gail et al. [13] provide single-time survival probability (i.e, the probability that a woman will develop breast cancer within 5 years based on its characteristics), and Kaplan-Meier method [26] estimates a population-level survival curve, which, while valuable, gives an average survival probability for a broad group of individuals. However, neither can offer individualized, time-dependent sur-

vival distributions. As highlighted by Haidar et al. [17], there is a need for models that can estimate individual survival distributions (ISDs), since these distributions provide more information for explaining an individual's survival compared to single point estimates. Researchers developed models such as the Kalbfleisch-Prentice extensions of the Cox (Cox-KP) [25], the elastic net Cox (Coxen-KP) [60], and Multi-task Logistic Regression (MTLR) [63] to estimate Individual Survival Distribution (ISD) [17].

To learn a model that can estimate patient's ISD using that patient's ECG, we can use two general approaches; (1) an end-to-end ISD model (deep or shallow) that uses raw ECG signals as input, and (2) High-level features/embeddings of the ECG signal, derived from intermediate supervised tasks, are utilized as input for learning an ISD model. In the latter approach, the ECGs are encoded into a lower dimension while retaining sufficient important information about the ECGs. This can be achieved through supervised, semi-supervised, unsupervised machine learning, or knowledge-based methods. In the case of supervised feature extractor models, the algorithms are learned for a particular task, such as multi-label classification. These algorithms produce ECG encodings (features), typically using deep learning models like Inception [53] or ResNet [19]. These encodings can help estimate patient-specific ISD. However, it is not guaranteed that these features, optimized for these various tasks, will produce an more accurate ISD. Unsupervised machine learning techniques such as autoencoder (AE) or variational autoencoder (VAE) can encode ECGs into lower-dimensional features. However, there is no guarantee of good performance in producing more accurate ISD, similar to the limitations with supervised extracted features obtained from a different supervised task.

An alternative approach for encoding ECGs is through time series analysis, which produced global features during ECG data collection. These methods use time series techniques to convert ECGs into features, however, the features that are used are limited (i.e. QRS duration and PR interval). Models trained on these knowledge-based features are generally less accurate than supervised or unsupervised methods, but the models are easier to interpret as features have physical meaning. In this chapter, we will evaluate the effectiveness of

ECG features obtained by various techniques, based on the accuracy of the downstream patient-specific ISD.

5.1 Objectives and Methods

In this Chapter, we aim to achieve two goals (Figure 5.1-a). First, we compare the performance of ISD models (COX-PH versus N-MTLR) by utilizing ECG features obtained through various methods as discussed in Chapter 3. Second, we compare the performance of ISD models trained on representative ECG features (supervised, unsupervised, and knowledge-based) using the better-performing ISD model (hint: N-MTLR) as a function of development sample size. To accomplish the second objective, we select 7 different training sample sizes: 100, 500, 1000, 5000, 10000, 100000, and 50000. The diagram in Figure 5.1-b shows the sample sizes utilized for each of these objectives. For each training sample size and ECG features, we use 10 different random splitting training sets to train 10 models and plot the mean of the performance, with error bars reflecting the 95% confidence interval around this mean. Here, we provide a brief overview of the methods used for ECG feature extraction. For a more detailed exploration of the methodologies readers are referred to Chapter 3:

- **TCN-β-VAE**: This approach uses a Temporal Convolutional Network-based β-Variational Autoencoder for unsupervised feature extraction from the Alberta ECG dataset (Figure 5.1-d), resulting in an ECG feature size of 32.

- **RN-β-VAE**: This method employs a ResNet-based β-VAE architecture to extract features from each ECG lead. Each lead is trained separately, and the learned features are then combined, resulting in an ECG feature size of 768 (Figure 5.1-c).

- **RN-β-VAE-lead#**: A method involving a ResNet-based β-VAE architecture to extract features from each ECG lead (Figure 5.1-c), resulting in an ECG feature size of 64 for each lead.

- **RN-ICD**: This utilizes a ResNet-based model pre-trained for multilabel classification of ICD-10 codes to generate supervised ECG features (Figure 5.1-b), resulting in an ECG feature size of 1414.

- **End-To-End Deep Learning (E2EDL)**: Directly trains ISD models using raw labeled 12-lead ECG signals with a ResNet architecture (Figure 5.1-a).

- **Knowledge-based ECG Features**: Extracts 22 knowledge-based ECG features for each of the 12 leads using the Philips IntelliSpace ECG Machine. This includes well-known features such as QRS duration and RR interval (Figure 5.1-e).

To evaluate the performance of ISD models, we use three metrics: the Concordance index (C-index), Marginal L1-loss, and Integrated Brier Score (IBS). These metrics are described in Section 3.4.2 of Chapter 3. The higher value of the C-index and lower value of Marginal L1 loss and IBS show a better model performance.

5.2 Results

We evaluated the effectiveness of different ECG features in estimating ISD using two models. Tables 5.1 and 5.2 show the results of the ISD estimated by COX-PH and N-MTLR models using all training data (as per the first objective), respectively. The ECG features that performed the best for each performance metric are highlighted in bold. To set a baseline, we calculated the median survival time for both uncensored and all patients using the Kaplan-Meier (KM) method, and the Marginal L1 loss when the model predicted the median survival time for all patients. If the ECG features have the sufficinet information to estimate ISD, the model trained on these Features should have a smaller Marginal L1 loss than the baseline models. For the COX-PH model (Table 5.1), the performance of all unsupervised features (RN-β-VAE and TCN-β-VAE) and knowledge-based features are close to the performance of

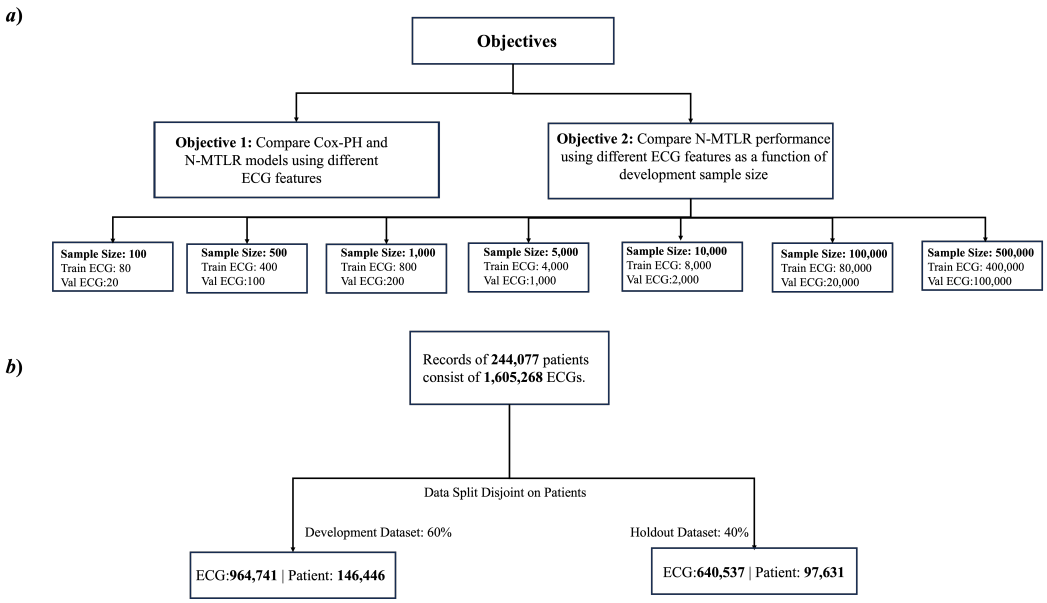

Figure 5.1: Flowchart of study design: a) Objectives and experiment design of this study. For both objectives, we use the same hold-out dataset for their evaluations. b) Data split of the development dataset (90% for training and 10% for validation) and hold-out dataset. During training, the ISD models were run for 512 epochs with an early stopping mechanism implemented. Training will be stopped if there are no improvements in the validation set's loss for 5 consecutive epochs.

the baseline. RN-ICD showed a significantly better performance in all metrics than the baseline. For the N-MTLR model (Table 5.2), the results show that all ECG features have significantly better performance than the baseline. Among ECG features, supervised deep learning features (RN-ICD and E2E DL) outperformed unsupervised features (RN-β-VAE and TCN-β-VAE) and knowledge-based features in terms of C-index and IBS. E2E DL showed a slightly better Marginal L1 loss among all models. Among all feature sets, the RN-ICD features stood out, being the best in two metrics, C-index (0.8058) and IBS (0.1360), and comparable performance in terms of Marginal L1 loss. The difference in performance between unsupervised features of RN-β-VAE and TCN-β-VAE was negligible. It is worth mentioning that the performance metrics for each lead individually (RN-β-VAE-lead) are lower than the 12 lead signals (TCN-β-VAE) – see Table 5.3. For all metrics, the knowledge-based ECG features had slightly better performance than RN-β-VAE but had per-

formance similar to unsupervised features of TCN-β-VAE.

To achieve the second objective of our study, which is to analyze the impact of the development dataset sample size on the performance of the ISD model, we will use the N-MTLR model as it performed better compared to the COX-PH model. The selected sample sizes are 100, 500, 1000, 10,000, 50,000, 100,000, and 500,000. We chose better-performing features from different categories, including supervised (RN-ICD and E2E DL), unsupervised (TCN-β-VAE), and knowledge-based ECG features. Furthermore, we will use age and sex features as a baseline to compare the model's performance when no ECG features are used. This baseline will serve as a reference point to assess the contribution of ECG features to the learning process and enhancement of the ISD model's performance. Figures 5.2, 5.3, 5.4 show the result of these experiments. For all metrics, as we added more training samples, the model's performance improved, as expected. The supervised ECG features of RN-ICD outperformed other ECG features for all metrics and sample sizes. The C-index of RN-ICD showed a clear advantage of using ECG features, even with as few as 500 training instances compared to baseline age and sex features. For the E2EDL, up to 10,000 training sample size, C-index was inferior than age and sex, and started getting better up to the maximum training sample size of 500,000. However, for all other metrics, the performance of E2EDL was lower than other models. For knowledge-based and TCN-β-VAE, however, a training size of 5,000 and 10,000 respectively was required to achieve higher performance than the baseline. Improvement in performance was minimal after 50,000 training samples for all metrics and all ECG features. Additionally, knowledge-based ECG features showed slightly better performance than TCN-β-VAE features for all training sample sizes.

5.3 Discussion

Using COX-PH and N-MTLR models and a large ECG dataset of 244,077 patients, we investigated the performance of ISD models using raw ECG signals, and ECG features obtained from supervised and unsupervised learning as well

Table 5.1: Survival Prediction performance using various generated embedding approaches from ECG signals to predict time until death using the COX-PH model. Higher values indicate better performance for the C-index, while lower values are preferable for both Marginal L1 loss and the Integrated Brier Score (IBS). Due to the large size of the test set (641,000 ECGs), multiple experiments were not conducted due to computational constraints. However, the range of variability in the results can be inferred to be similar to that observed in the largest training size example from the sample size effect study, which is based on the results of 10 independent experiments.

ECG Feature Approach	**Feature Size**	**Marginal L1 loss (days)**	**C-index**	**IBS**
E2E DL	12×4096	2622.83	0.50	0.22
RN-ICD	1414	**1984.14**	**0.77**	**0.15**
RN-β-VAE	768	2653.46	0.50	0.24
TCN-β-VAE	32	2607.17	0.51	0.23
ECG Measurement	22	2672.32	0.50	0.22
Median Survival time from all patients = 3420	-	2749.90	-	-
Median Survival time from all uncensored patients = 496	-	2615.52	-	-

as knowledge-based features. The results for both models showed that ECG features obtained from the supervised ECG extractor method have higher performance than using raw ECG signals as well as unsupervised and knowledge-based ECG features. However, except for Marginal L1 loss and using RN-ICD, which had a better performance using the COX-PH model, the performance of COX-PH was inferior to N-MTLR for all metrics(C-index, Integrated Brier Score, and Marginal L1 loss) possibly because the former assumes a constant hazard ratio and a linear relationship between the features and the log hazard, which is unrealistic. The N-MTLR model is more appropriate as it is not make such assumptions.

The supervised ECG features (RN-ICD) achieved superior performance when compared to other ECG features. Considering the direct relationship between morbidity and mortality, it is clinically sensible to incorporate di-

Table 5.2: Survival Prediction performance using various generated embedding approaches from ECG signals to predict time until death using the N-MTLR model. Higher values indicate better performance for the C-index, while lower values are preferable for both Marginal L1 loss and the Integrated Brier Score (IBS). Due to the large size of the test set (641,000 ECGs), multiple experiments were not conducted due to computational constraints. However, the range of variability in the results can be inferred to be similar to that observed in the largest training size example from the sample size effect study, which is based on the results of 10 independent experiments.

ECG Feature Approach	**Feature Size**	**Marginal L1 loss (days)**	**C-index**	**IBS**
E2E DL	12×4096	**2021.94**	0.75	0.16
RN-ICD	1414	2152.02	**0.81**	**0.14**
RN-β-VAE	768	2145.15	0.70	0.17
TCN-β-VAE	32	2106.28	0.72	0.17
ECG Measurement	22	2121.62	0.73	0.17
Median Survival time from all patients = 3420	-	2749.90	-	-
Median Survival time from all uncensored patients = 496	-	2615.52	-	-

agnostic predictions as features for training accurate survival models. This suggests that ECG features obtained from an intermediate supervised task are a better candidate as ECG features for training ISD models. This finding is aligned with the study of Popescu et al. [41], which developed a deep learning algorithm that leveraged patient covariates, including some ECG global features, and 3D cardiac magnetic resonance images, to predict ISDs for the task of sudden cardiac death in patients with ischemic heart disease. The model achieved C-index and IBS of 0.83 and 0.12, respectively, for their internal validation set. Other studies in the literature have primarily focused on predicting single-time mortality (such as 1-year mortality prediction using ECG signals) [52].

The results indicate that the performance of unsupervised ECG features (RN-β-VAE and TCN-β-VAE) and knowledge-based features are similar, sug-

Table 5.3: Survival Prediction performance using various ECG lead's embedding to predict time until death using the N-MTLR model. Higher values indicate better performance for the C-index, while lower values are preferable for both Marginal L1 loss and the Integrated Brier Score (IBS). Due to the large size of the test set (641,000 ECGs), multiple experiments were not conducted due to computational constraints. However, the range of variability in the results can be inferred to be similar to that observed in the largest training size example from the sample size effect study, which is based on the results of 10 independent experiments.

Embedding Approach	**Feature Size**	**Marg. L1 loss (days)**	**C-index**	**IBS**
RN-β-VAE-lead#1	64	2176.09	0.70	0.17
RN-β-VAE-lead#2	64	2179.39	0.70	0.18
RN-β-VAE-lead#3	64	2179.49	0.70	0.17
RN-β-VAE-lead# aVR	64	2173.28	0.70	0.17
RN-β-VAE-lead# aVL	64	2174.60	0.71	0.17
RN-β-VAE-lead# aVF	64	2173.15	0.71	0.17
RN-β-VAE-lead# V1	64	2170.17	0.70	0.17
RN-β-VAE-lead# V2	64	2165.49	0.70	0.17
RN-β-VAE-lead# V3	64	2167.94	0.71	0.17
RN-β-VAE-lead# V4	64	2194.94	0.70	0.18
RN-β-VAE-lead# V5	64	2160.49	0.71	0.17
RN-β-VAE-lead# V6	64	2181.72	0.69	0.18

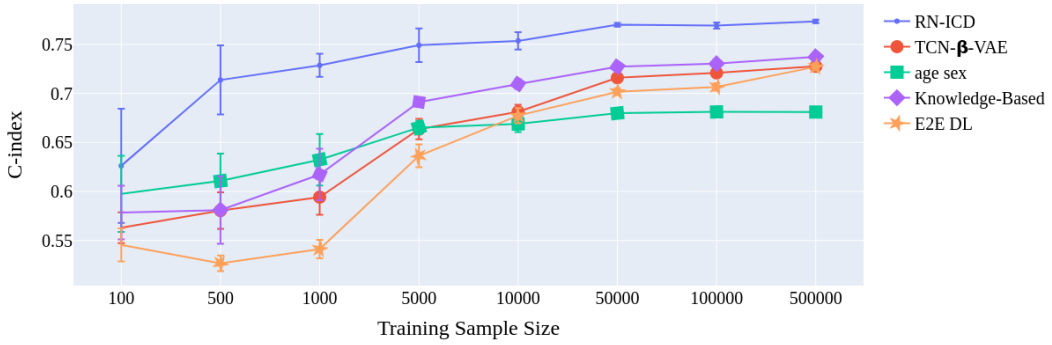

Figure 5.2: C-index of N-MTLR model as a function of training sample size using various supervised, unsupervised, and knowledge-based ECG features. The points represent the mean value over the 10 experiments, and the bars represent confidence intervals.

gesting there is no clear advantage in using knowledge-based features over unsupervised, even though knowledge-based is expected to be more informative. However, the trained unsupervised architecture can be used to generate synthetic ECGs that might be beneficial for other tasks. Also, there is no significant difference between the performance metrics of features obtained from different single leads (RN-β-VAE-lead# features), suggesting that any of the leads can be used to train the ISD model with no significant compromise on the performance metrics.

Supervised ECG features outperformed other ECG-obtained features at a smaller sample size when considering training sample size. Only a training sample size of 500 was required for supervised ECG features to achieve better performance than using age and sex alone. The performance of E2EDL was lower than all other models and started improving with a larger training sample size. For unsupervised and knowledge-based features, a training sample size of 5000 was needed to achieve higher performance than using only age and sex features. Additionally, we did not observe any significant improvement in the ISD model's performance using training sizes beyond 50,000 samples for all ECG features.

Here, unsupervised ECG features and knowledge-based features had a com-

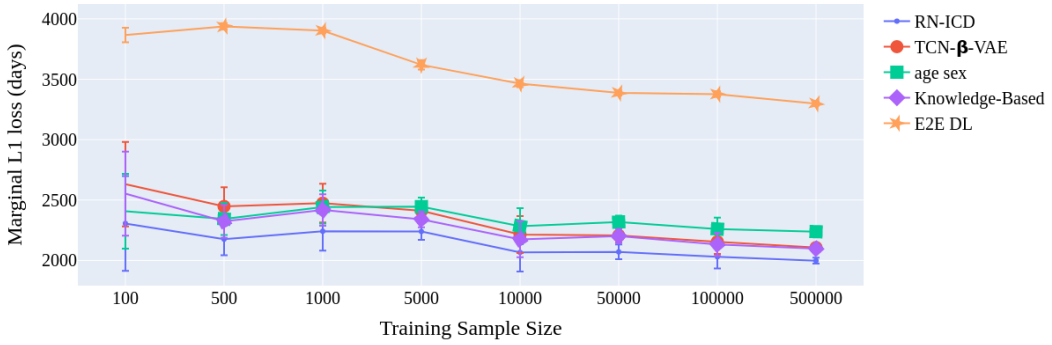

Figure 5.3: Marginal L1 loss of N-MTLR model as a function of training sample size using various supervised, unsupervised, and knowledge-based ECG features. The points represent the mean value over the 10 experiments, and the bars represent confidence intervals.

parable performance for training ISD models. However, more recently developed unsupervised algorithms and/or semi-supervised training of such models (including multi-task learning during the VAE training) could lead to unsupervised ECG features that might outperform knowledge-based features to estimate patient-specific ISD. Note that our ISD models, trained on supervised ECG features, demonstrated superior performance. We therefore expect that this hybrid approach will enrich the embeddings, making them more effective in estimating ISDs.

5.3.1 Limitations

Our results should be considered in light of certain limitations. First, our study has explored only a specific set of feature extraction and embedding methods, as well as ISD methods. Additionally, we have utilized a selected set of labels for supervising the supervised feature extractor, in our case, medical diagnoses, given their direct implications on mortality. This was made possible due to our unique dataset, which includes a population-scale linkage between over 1 million digitized ECGs and more than 1000 wide-ranging ICD clinical diagnoses. However, it is important to note that these ECGs were generated by machines from the same manufacturer, which might limit the generalizabil-

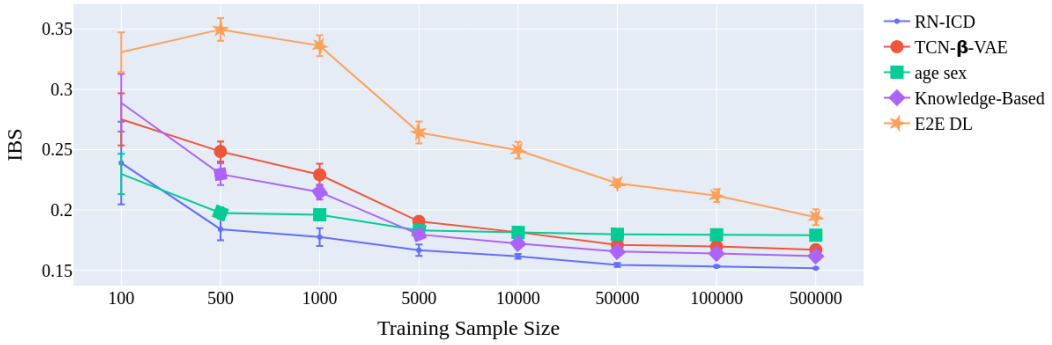

Figure 5.4: IBS of N-MTLR model as a function of training sample size using various supervised, unsupervised, and knowledge-based ECG features. The points represent the mean value over the 10 experiments, and the bars represent confidence intervals.

ity of our findings to ECGs from other systems. Furthermore, our prognostic models may be influenced by the inclusion of deaths unrelated to clinical factors, such as those resulting from traffic accidents or homicides. When our paper was submitted, we could not find any publicly available ECG datasets containing mortality information (or other temporal events) for use as labels in ISD tasks. Prominent ECG datasets, such as PhysioNet [14], MIT-BIH [38], and PTB [7], do not include death-related data linked to ECGs. However, it is possible that more comprehensive clinical datasets that include ECGs, like MIMIC [15], may become available in the future. These datasets could serve as benchmark data for ECG-based ISD tasks and external validation.

Chapter 6

Conclusions and Future Perspectives

Electrocardiogram (ECG) signals are widely used in clinical settings and contain informative measurement of a heart's health. In this thesis, we used the Alberta Hospital ECG Dataset, consisting of more than 1.6 million ECG collected from 244,077 patients, for two objectives: (1) To explore ways to build a stronger classifier using ECG signals to classify ECG abnormalities. (2) To extract high-level features from ECG traces, through various approaches including supervised with clinical diagnoses, unsupervised approaches, or knowledge-based ECG features and their efficacy in estimating patient-specific ISD. **Chapter 4** explored the first objectives, where we use unsupervised β-VAE algorithms to generate synthetic ECG signals, based on Alberta ECG dataset with specific abnormalities. We then added these synthetic ECG signals into the public dataset of China Physiological Signal Challenge 2018 Dataset (CPSC 2018). We found that a learner trained on this extended dataset performed better than one trained on only the original data for ECG classification. It is important to note some potential differences between the Alberta ECG Dataset and the CPSC 2018 Dataset. Firstly, the demographic characteristics of the two datasets may differ, as they were collected from distinct geographical locations (Canada and China, respectively), potentially reflecting variations in patient ethnicity, lifestyle, and environmental factors. Additionally, the Alberta ECG Dataset covers a broader range of cardiac conditions (15 possible labels) compared to the 9 conditions in the CPSC 2018 Dataset, indicating a

possible difference in clinical characteristics of the patient populations. Furthermore, while both datasets utilized 12-lead ECGs, variations in the specific data collection protocols and equipment could influence the ECG signals.

Chapter 5 focused on the second objective, where we first extracted ECG features using different methods: (1) supervised with clinical diagnoses, (2) unsupervised approaches, or (3) knowledge based ECG features. Using these ECG features, along with age and sex, models were trained to estimate patient-specific individual survival distributions (ISD) to predict time to death. The results showed that supervised learning approaches produced ECG features that can estimate patient-specific ISD curves better than ECG features obtained from unsupervised and knowledge-based methods. Supervised ECG features required fewer training instances (as low as 500) to learn ISD models that performed better than models that only used age and sex. The results reported here may assist researchers to build stronger classifiers for ECG use cases as well as assisting in selection of the most appropriate method for extracting high-level features from ECG signals to estimate patient-specific ISD curves.

6.1 Future Directions

In this thesis, we showed the beneficial effect of synthetic ECGs. But there are still some interesting research directions that can be explored. We can divide these future research into two main categories: exploration and improvement of synthetic ECG data, and ECG data privacy. In this thesis, we utilized β-VAE algorithms to generate synthetic ECG data. However, we did not explore other generative models like Generative Adversarial Networks (GANs). A comparative study between β-VAE and GANs could yield valuable insights into which method is more effective for data augmentation in the classification of ECG abnormalities. In addition of other data generation approaches, the privacy aspect of synthetic ECGs is another critical area that needs further exploration. We need to ensure that synthetic ECGs cannot be traced back to individual patients. We could use methods like differential privacy to add

statistical noise to the synthetic ECGs, thereby making it difficult to identify the original source. However, we need to make sure that the added noise does not significantly reduce the usefulness of synthetic ECGs for downstream tasks, such as the classification of ECG abnormalities. Also, we need to validate that the synthetic ECGs retain sufficient information about the original ECGs, such as QRS duration distribution and heart rate.

References

[1] (), [Online]. Available: https://ecgwaves.com/topic/ekg-ecg-leads-electrodes-systems-limb-chest-precordial/#:~:text=on%20the%20ECG.-,Derivation%20of%20the%20ECG%20leads,as%20reference%20(negative)%20electrode..

[2] (), [Online]. Available: https://www.mayoclinic.org/diseases-conditions/heart-failure/symptoms-causes/syc-20373142#dialogId12064530.

[3] (), [Online]. Available: P.%20M.%20Systems,%20Philips%20dxl%20ecg%20algorithm%20physician%E2%80%99s%20guide%20for%20ph100b,%20edition%202,%20April,%202009,%202009..

[4] E. Adib, A. S. Fernandez, F. Afghah, and J. J. Prevost, "Synthetic ecg signal generation using probabilistic diffusion models," *IEEE Access*, 2023.

[5] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[6] S. K. Berkaya, A. K. Uysal, E. S. Gunal, S. Ergin, S. Gunal, and M. B. Gulmezoglu, "A survey on ecg analysis," *Biomedical Signal Processing and Control*, vol. 43, pp. 216–235, 2018.

[7] R. Bousseljot, D. Kreiseler, and A. Schnabel, "Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet," 1995.

[8] H. Byers, H. Landsberg, H. Wexler, *et al.*, "Verification of weather forecasts," *Compendium of Meteorology: Prepared under the Direction of the Committee on the Compendium of Meteorology*, pp. 841–848, 1951.

[9] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package)," *Neurocomputing*, vol. 307, pp. 72–77, 2018.

[10] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[11] B. Dahlöf, "Cardiovascular disease risk factors: Epidemiology and risk assessment," *The American journal of cardiology*, vol. 105, no. 1, 3A–9A, 2010.

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

[13] M. H. Gail, L. A. Brinton, D. P. Byar, *et al.*, "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually," *JNCI: Journal of the National Cancer Institute*, vol. 81, no. 24, pp. 1879–1886, 1989.

[14] A. L. Goldberger, L. A. Amaral, L. Glass, *et al.*, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, e215–e220, 2000.

[15] B. Gow, T. Pollard, L. A. Nathanson, *et al.*, "Mimic-iv-ecg-diagnostic electrocardiogram matched subset," 2023.

[16] 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.

[17] H. Haider, B. Hoehn, S. Davis, and R. Greiner, "Effective ways to build and evaluate individual survival distributions," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 3289–3351, 2020.

[18] M. Hammad, A. Maher, K. Wang, F. Jiang, and M. Amrani, "Detection of abnormal heart conditions based on characteristics of ecg signals," *Measurement*, vol. 125, pp. 634–644, 2018.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] I. Higgins, L. Matthey, A. Pal, *et al.*, "Beta-vae: Learning basic visual concepts with a constrained variational framework," in *International conference on learning representations*, 2017.

[21] S. Hong, Y. Zhou, J. Shang, C. Xiao, and J. Sun, "Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review," *Computers in biology and medicine*, vol. 122, p. 103 801, 2020.

[22] 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.

[23] H. Ismail Fawaz, B. Lucas, G. Forestier, *et al.*, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.

[24] J.-H. Jang, T. Y. Kim, H.-S. Lim, and D. Yoon, "Unsupervised feature learning for electrocardiogram data using the convolutional variational autoencoder," *PloS one*, vol. 16, no. 12, e0260612, 2021.

[25] J. D. Kalbfleisch and R. L. Prentice, *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

[26] 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.

[27] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC medical research methodology*, vol. 18, no. 1, pp. 1–12, 2018.

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

[29] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[30] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165.

[31] R. R. van de Leur, M. N. Bos, K. Taha, *et al.*, "Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders," *European Heart Journal-Digital Health*, vol. 3, no. 3, pp. 390–404, 2022.

[32] F. Liu, C. Liu, L. Zhao, *et al.*, "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection," *Journal of Medical Imaging and Health Informatics*, vol. 8, no. 7, pp. 1368–1373, 2018.

[33] X. Liu, H. Wang, Z. Li, and L. Qin, "Deep learning in ecg diagnosis: A review," *Knowledge-Based Systems*, vol. 227, p. 107 187, 2021.

[34] S. Ma, J. Cui, C.-L. Chen, X. Chen, and Y. Ma, "An effective data enhancement method for classification of ecg arrhythmia," *Measurement*, vol. 203, p. 111 978, 2022.

[35] D. Makowski, T. Pham, Z. J. Lau, *et al.*, "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, Feb. 2021. DOI: 10.3758/s13428-020-01516-y. [Online]. Available: https://doi.org/10.3758%2Fs13428-020-01516-y.

[36] M. Merone, P. Soda, M. Sansone, and C. Sansone, "Ecg databases for biometric systems: A systematic review," *Expert Systems with Applications*, vol. 67, pp. 189–202, 2017.

[37] S. F. Mokhtar, Z. M. Yusof, and H. Sapiri, "Confidence intervals by bootstrapping approach: A significance review," *Malaysian Journal of Fundamental and Applied Sciences*, vol. 19, no. 1, pp. 30–42, 2023.

[38] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE engineering in medicine and biology magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[39] I. Oguiza, *Tsai - a state-of-the-art deep learning library for time series and sequential data*, Github, 2022. [Online]. Available: `https://github.com/timeseriesAI/tsai`.

[40] A. v. d. Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[41] D. M. Popescu, J. K. Shade, C. Lai, *et al.*, "Arrhythmic sudden death survival prediction using deep learning analysis of scarring in the heart," *Nature Cardiovascular Research*, vol. 1, no. 4, pp. 334–343, 2022.

[42] A. J. Prakash, K. K. Patro, M. Hammad, R. Tadeusiewicz, and P. Pławiak, "Baed: A secured biometric authentication system using ecg signal based on deep learning techniques," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 4, pp. 1081–1093, 2022.

[43] S. Raghunath, A. E. Ulloa Cerna, L. Jing, *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.

[44] 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, pp. 651–662, 2016.

[45] A. H. Ribeiro, M. H. Ribeiro, G. M. Paixão, *et al.*, "Automatic diagnosis of the 12-lead ecg using a deep neural network," *Nature communications*, vol. 11, no. 1, p. 1760, 2020.

[46] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[47] S. Somani, A. J. Russak, F. Richter, *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.

[48] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[49] J. Sun, X. Wang, N. Xiong, and J. Shao, "Learning sparse representation with variational auto-encoder for anomaly detection," *IEEE Access*, vol. 6, pp. 33 353–33 361, 2018.

[50] W. Sun, "Learning models for diagnosis and prognosis from electrocardiogram data," 2023.

[51] W. Sun, S. V. Kalmady, N. Sepehrvan, *et al.*, "Improving ecg-based covid-19 diagnosis and mortality predictions using pre-pandemic medical records at population-scale," *arXiv preprint arXiv:2211.10431*, 2022.

[52] W. Sun, S. V. Kalmady, N. Sepehrvand, *et al.*, "Towards artificial intelligence-based learning health system for population-level mortality prediction using electrocardiograms," *NPJ Digital Medicine*, vol. 6, no. 1, p. 21, 2023.

[53] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[54] C. W. Tsao, A. W. Aday, Z. I. Almarzooq, *et al.*, "Heart disease and stroke statistics—2023 update: A report from the american heart association," *Circulation*, vol. 147, no. 8, e93–e621, 2023.

[55] E.-S. Väliaho, J. A. Lipponen, P. Kuoppa, *et al.*, "Continuous 24-h photoplethysmogram monitoring enables detection of atrial fibrillation," *Frontiers in Physiology*, vol. 12, p. 778 775, 2022.

[56] P. Wang, B. Hou, S. Shao, and R. Yan, "Ecg arrhythmias detection using auxiliary classifier generative adversarial network and residual network," *Ieee Access*, vol. 7, pp. 100 910–100 922, 2019.

[57] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *2017 International joint conference on neural networks (IJCNN)*, IEEE, 2017, pp. 1578–1585.

[58] E. Winter, "The shapley value," *Handbook of game theory with economic applications*, vol. 3, pp. 2025–2054, 2002.

[59] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[60] Y. Yang and H. Zou, "A cocktail algorithm for solving the elastic net penalized cox's regression in high dimensions," *Statistics and its Interface*, vol. 6, no. 2, pp. 167–173, 2013.

[61] Ö. Yıldırım, P. Pławiak, R.-S. Tan, and U. R. Acharya, "Arrhythmia detection using deep convolutional neural network with long duration ecg signals," *Computers in biology and medicine*, vol. 102, pp. 411–420, 2018.

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

[63] 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.

[64] Y.-H. Zhang and S. Babaeizadeh, "Synthesis of standard 12-lead electrocardiograms using two-dimensional generative adversarial networks," *Journal of Electrocardiology*, vol. 69, pp. 6–14, 2021.

[65] H. Zheng, Z. Yang, W. Liu, J. Liang, and Y. Li, "Improving deep neural networks using softplus units," in *2015 International joint conference on neural networks (IJCNN)*, IEEE, 2015, pp. 1–4.