Supervised Electrocardiogram(ECG) Features Outperform Knowledge-based And Unsupervised Features In Individualized Survival Prediction

Yousef Nademi[*] NADEMI@UALBERTA.CA
Sunil V Kalmady[†] KALMADY@UALBERTA.CA
Weijie Sun[‡] WEIJIE2@UALBERTA.CA
Shi-ang Qi[§] SHIANG@UALBERTA.CA
Abram Hindle[¶] HINDLE1@UALBERTA.CA
Padma Kaul[‖] PKAUL@UALBERTA.CA
Russell Greiner[**] RGREINER@UALBERTA.CA
[1]*Department of Computing Science, University of Alberta, Edmonton, Canada*
[2]*Canadian VIGOUR Centre, Department of Medicine, University of Alberta, Edmonton, Canada*
[3]*Department of Medicine, University of Alberta, Edmonton, Canada*
[4]*Alberta Machine Intelligence Institute, Edmonton, Canada*

Abstract

An electrocardiogram (ECG) provides crucial information about an individual's health status. Researchers utilize ECG data to develop learners for a variety of tasks, ranging from diagnosing ECG abnormalities to estimating time to death – here modeled as individual survival distributions (ISDs). The way the ECG is represented is important for creating an effective learner. While many traditional ECG-based prediction models rely on hand-crafted features, such as heart rate, this study aims to achieve a better representation. The effectiveness of various ECG based feature extraction methods for prediction of ISDs, either supervised or unsupervised, have not been explored previously. The study uses a large ECG dataset from 244,077 patients with over 1.6 million 12-lead ECGs, each labeled with the patient's disease – one or more International Classification of Diseases (ICD) codes. We explored extracting high-level features from ECG traces using various approaches, then trained models that used these ECG features (along with age and sex), across a range of training sizes, to estimate patient-specific ISDs. The results showed that the supervised feature extractor method produced ECG features that can estimate ISD curves better than ECG features obtained from unsupervised or knowledge-based methods. Supervised ECG features required fewer training instances (as low as 500) to learn ISD models that performed better than the baseline model that only used age and sex. On the other hand, unsupervised and knowledge-based ECG features required over 5,000 training samples to produce ISD models that performed better than the baseline. The study's findings may assist researchers in selecting the most appropriate approach for extracting high-level features from ECG signals to estimate patient-specific ISD curves.

Keywords: Electrocardiogram, Individual Survival Distributions, Variational AutoEncoder

1. Introduction

Heart abnormalities are one of the leading causes of mortality in the world. In 2020, 19.05 million individuals died globally due to heart disease (Tsao et al., 2023). Electrocardiograms (ECGs) are easy to collect measurements, possessing valuable information pertinent to heart health and mortality. Each individual's ECG provide a unique information for the individual's health trajectory and potentially his/her

[*] 1
[†] 1,2
[‡] 1
[§] 1
[¶] 1
[‖] 2,3
[**] 1,4

survival time. Accurate estimation of Individual Survival Distributions (ISDs) has the potential to significantly reduce both mortality and healthcare expenses. Factors like age and health status significantly influence these distributions for each patient (Haider et al., 2020). Additionally, through the estimation of ISDs, healthcare professionals can offer personalized predictions that can guide medical decisions, interventions, and resource allocation.

2. Related Work

Traditional risk assessment tools, such as the Cox Proportional Hazard model (Kalbfleisch and Prentice, 2011), yield time-independent risk scores without distributions across multiple time points. Alternatively, models like those proposed by Gail et al. (1989) give single-time survival probabilities (i.e. the probability that a woman will develop breast cancer within 5 years based on its characteristics), while the Kaplan and Meier (1958) curve, which, while valuable, gives an average survival probability for a broad group of individuals. However, none of these can offer individualized survival distributions – a probability curve for all future time points for a specific patient. Recognizing this gap, researchers developed models such as the random survival forest (RSF) (Iswaran et al., 2008) Kalbfleisch-Prentice extensions of the Cox (Cox-KP) (Kalbfleisch and Prentice, 2011), the elastic net Cox (Coxen-KP) (Yang and Zou, 2013), Multi-task Logistic Regression (MTLR) (Yu et al., 2011), Neural-MTLR (N-MTLR) (Fotso, 2018), DeepHit (Lee et al., 2018) , SODEN (Tang et al., 2022), and BNN-ISD (Qi et al., 2023) to estimate ISD

ECGs, when combined with parameters like age and sex, have been successfully used to predict 1-year mortality (Sun et al., 2023) suggesting ECGs contain the information needed for mortality prediction (Raghunath et al., 2020). To learn a model that can estimate patient's ISD using that multidimensional bio-signal data such as 12 lead ECGs, we can use two general approaches: (1) using an end-to-end neural network based ISD learning system (deep or shallow) that takes raw ECG signals, each labeled for survival time, as input, and (2) first learning some high-level features/embeddings of ECG signal, produced from intermediate tasks, then using that encoding as input for learning an ISD model. In the latter approach, the goal is to encode the ECGs into a lower dimension while retaining sufficient information to produce an effective survival model. This can be achieved through supervised, semi-supervised, unsupervised machine learning, or knowledge-based methods (defined in Section 3.3.6). In the case of supervised feature extractor models, the algorithms are learned for a particular task other than estimating ISD, but supervised by clinically relevant labels. such as multi-label classification. These algorithms produce ECG encodings (Features for downstream prediction task), typically using deep learning models like Inception (Szegedy et al., 2017) or ResNet (He et al., 2016). However, it is not guaranteed that these features, optimized for these various tasks, will produce accurate ISDs. Unsupervised machine learning techniques, such as autoencoder (AE) (Hinton and Zemel, 1993) or variational autoencoder (VAE) (Higgins et al., 2016), can also encode ECGs into lower-dimensional features, as evidenced by studies like Kuznetsov et al. (2020) and Jang et al. (2021). However, there is no guarantee that this will lead to an accurate ISD model, similar to the limitations with features extracted by supervision. An alternative approach for encoding ECGs is through clinical knowledge based features, which produced global features during ECG data collection. They use clinical morphology of waveforms 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 clinical physical meaning.

Therefore, this study explores the various feature extraction methods for ECGs, and compares their effectiveness in predicting ISD across the range of training sample sizes. The analysis conducted across different sample sizes can be a valuable reference for other researchers, helping them identify optimal starting points and choose the most suitable methods for their research goals.

3. Method

The ECG data for this study received approval from the Health Research Ethics Board at the University of Alberta.

3.1. Dataset

Our dataset (Sun et al., 2023) consists of 12-lead ECG signals from 244,077 patients, collected using

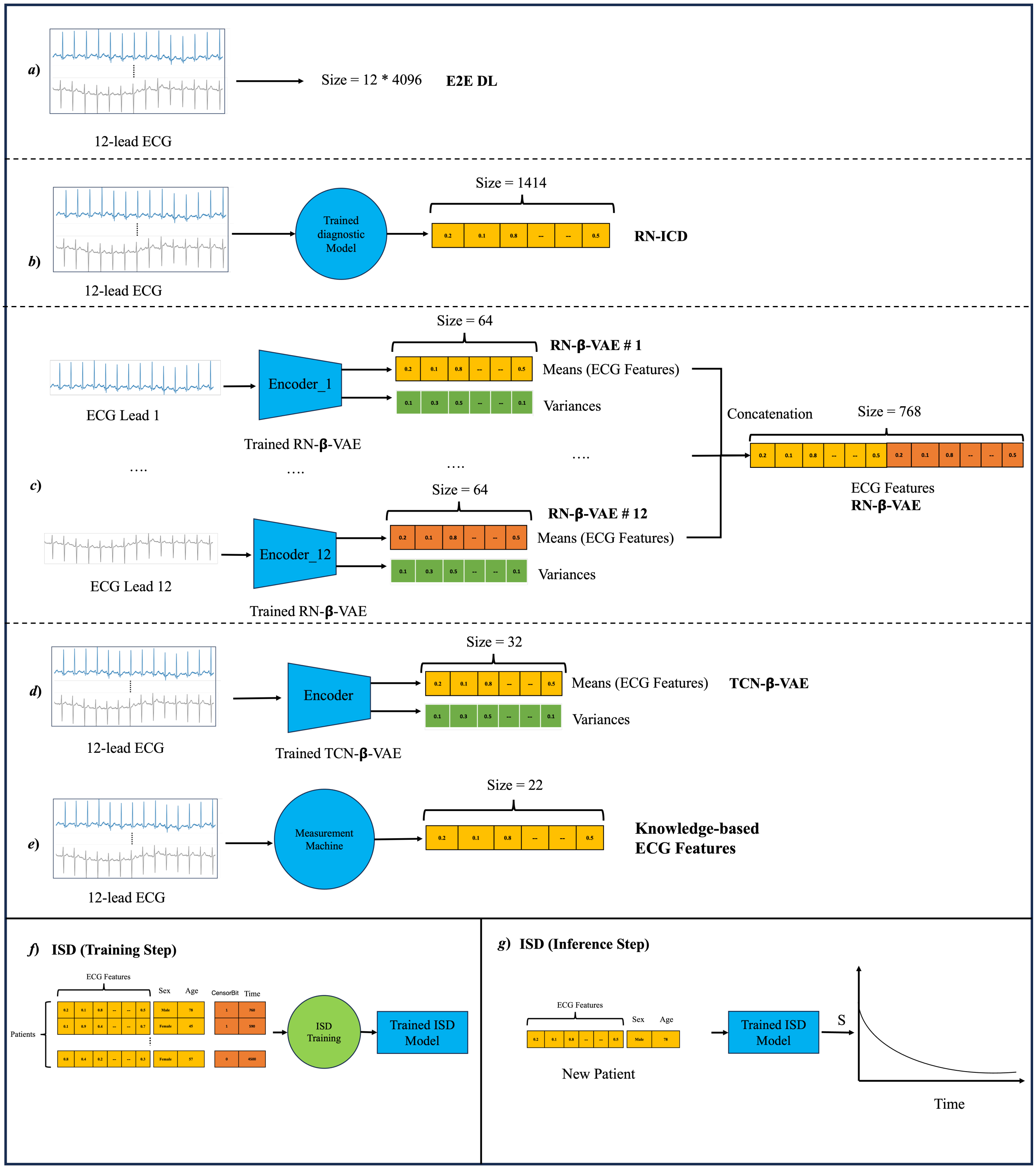

Figure 1: a-e) ECG feature extraction methods used to produce intermediate encoding. f)Using ECG features (from (a) to (e) above), as well as Sex and Age, and the label (CensorBit and Time) to train ISD models. g) In the ISD inference step, we use the trained ISD model to estimate the ISD for a new patient, based on his/her ECG features along with age and sex.

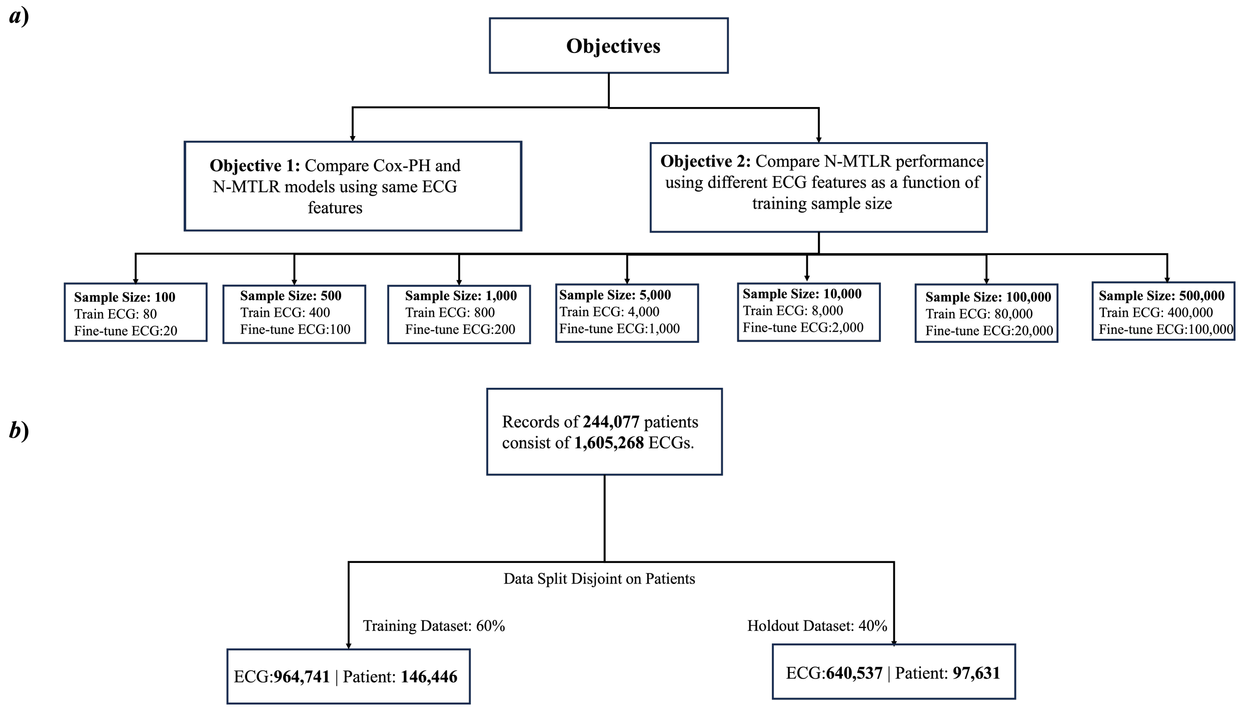
a)
Objectives
Objective 1: Compare Cox-PH and N-MTLR models using same ECG features
Objective 2: Compare N-MTLR performance using different ECG features as a function of training sample size
Sample Size: 100 Train ECG: 80 Fine-tune ECG:20
Sample Size: 500 Train ECG: 400 Fine-tune ECG:100
Sample Size: 1,000 Train ECG: 800 Fine-tune ECG:200
Sample Size: 5,000 Train ECG: 4,000 Fine-tune ECG:1,000
Sample Size: 10,000 Train ECG: 8,000 Fine-tune ECG:2,000
Sample Size: 100,000 Train ECG: 80,000 Fine-tune ECG:20,000
Sample Size: 500,000 Train ECG: 400,000 Fine-tune ECG:100,000
b)
Records of 244,077 patients consist of 1,605,268 ECGs.
Data Split Disjoint on Patients
Training Dataset: 60%
Holdout Dataset: 40%
ECG:964,741 | Patient: 146,446
ECG:640,537 | Patient: 97,631

Figure 2: 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 training dataset (90% for training and 10% for fine-tuning) and hold-out dataset for both objectives and all ECG feature extractor methods.

a Philips IntelliSpace ECG machine at a sampling frequency of 500 Hz. For training the ISD models, we normalized the ECG features using the z-score normalization method. To train the models, both models to produce ECG features and models to estimate ISDs, we split the dataset into a training set (60% = 964,741 ECG signals from 146,466 patients) and a holdout set (40% = 640,527 ECG signals from 97,631 patients) while making sure that ECGs from the same patient is not included in both training and holdout datasets. (Figure 2).

3.2. Prediction task

We want to estimate patient-specific ISD for the event of interest (death) using ECG signals. We will use the training set along with the patient's age and sex, a specific partitioning of time into bins, and target event (either still alive or dropped out of the study as 0, or death as 1 for each time bin) to train ISD models. We used GPUs (Tesla V100-SXM2, 32GB of RAM) and Pytorch 1.11 to train all our models.

3.3. ECG features for estimating ISD

Figure 1 shows a summary of ECG features that we use to train ISD models.

3.3.1. End-To-End Deep Learning

The first approach (Figure 1-a) uses raw ECG signals to directly train ISD models, where each raw ECG signal consists of 4096 real values. 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. (2022). We then feed the output of those layers, along with age and sex into the ISD modes. We call this approach End-To-End Deep Learning **(E2E DL)**.

3.3.2. Supervised ECG features

We will use a pre-trained model developed in our group for feature extraction as explained by Sun et al. (2022). This model was originally developed for multilabel classification of International Classification of Diseases, 10th revision (ICD-10) codes. (Note that here we consider 1414 different ICD-10 codes, binary labels, and label each patient with 0 or more of them. (Sun et al., 2023)). We will use the predicted probabilities generated by this model for raw ECG signals as supervised ECG features. The size of these ECG features is 1414 as shown in Figure 1-b. We call this approach **RN-ICD**, which will use these supervised ECG features along with age and sex to train the ISD models.

3.3.3. Unsupervised ECG features

We will use two β-VAE models to extract embeddings/features for ECG signals: β-VAE where the building block of the encoder-decoder is based on either (1) ResNet, or (2) Temporal Convolutional Network (TCN). We will refer to the ECG features obtained from these two models as unsupervised ECG features. Supplementary Information (SI) section 1 will explain the training process of these two models. Here, we briefly describe these two models.

3.3.4. ResNet-based β-VAE

Follwoing Jang et al. (2021), we adopt the β-VAE architecture with residual connections, for our ECG dataset. This model design is created to learn single-lead ECG signals, see Figures S1 and S2. We constructed a separate model for each lead of the 12-lead ECGs. After learning,, we then use the trained model to extract ECG features, using one model for each lead. For example, lead number one of these ECGs will be fed into the model that was trained on lead number one. This approach maps each ECG lead into 64 features (Figure 1-c). We will refer to these unsupervised ECG features as **RN-β-VAE-lead#**. Then, we concatenate ECG features obtained from each lead. We call these combined ECG features as **RN-β-VAE**.

3.3.5. TCN-based β-VAE

TCN-based β-VAE approach adapt the β-VAE architecture and code provided by van de Leur et al. (2022), to our ECG dataset. Using the encoder of the trained model, we map each 12-lead ECG into a 32 embeddings/features (Figure 1-d). We call this approach TCN-β-VAE.

3.3.6. Knowledge-based ECG features

The Philips IntelliSpace ECG Machine generated 22 global measurements for each of the 12 leads (Figure 1-e)– features that are well-known to experts, such as QRS duration, RR-interval, and heart-rate to name a few (Hammad et al., 2018). We call these features **knowledge-based**.

3.4. ISD Algorithms

3.4.1. Cox-Proportional Hazard (COX-PH) Model

The Cox-PH Model (Cox, 1972) is one of the well-known statistical methods used for survival analysis. Here, we will use the Cox-PH Model to estimate patient-specific ISD using any of the ones, described earlier.

3.4.2. Neural Multi-Task Logistic Regression (N-MTLR) Model

N-MTLR (Fotso, 2018) is a modified version of MTLR (Yu et al., 2011). 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 a specific time interval. (Note: for each of them).To learn a MTLR model, we first divide the entire time horizon into m time bins. Yu et al. (2011) states that for each time bin, we estimate the survival probability as follows:

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

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 $y = y(d) = (y_1, \ldots, y_m) \in \{0,1\}^m$. Here, y_i is set to 0 if death has not occurred by the time t_i – that is, $t_i < d$. 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:

$$P_{\boldsymbol{\Theta}}(Y|\mathbf{x}) = \frac{\exp\left(\sum_{i=1}^{m} y_i(\boldsymbol{\theta}_i\mathbf{x} + b_i)\right)}{\sum_{k=0}^{m} \exp(f_{\boldsymbol{\Theta}}(\mathbf{x}, k))} \tag{2}$$

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 $[t_k, t_{k+1})$ (Yu et al., 2011). 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 along with age and sex.

3.5. Evaluation

This paper attempts to achieve two goals. Firstly, we compare the performance of ISD models (COX-PH versus N-MTLR) by utilizing ECG features obtained through various methods as discussed earlier. Secondly, 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 training sample size. The diagram in Figure 2-a shows the sample sizes utilized for each of these objectives. To accomplish the second objective, we select 7 different training sample sizes, 100, 500, 1000, 5000, 10000, 100000, and 50000. 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. To evaluate the performance of ISD models, we use three metrics: the Concordance index (C-index), Marginal L1-loss, and Integrated Brier Score (IBS). The higher value of the C-index and lower value of Marginal L1 loss and IBS show a better model performance. In Section S2, we provide the definition of these metrics.

4. Results

We evaluated the effectiveness of different ECG features in estimating ISD using two models. Tables 1 and 2 show the results of the ISD estimated by COX-PH and N-MTLR models using all training dataset (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 (MST) for both uncensored and all patients using the Kaplan-Meier (KM) method Kaplan and Meier (1958), and the Marginal L1 loss when the model predicted the median survival time for all patients. If the ECG features have the required information for the ISD task, the ISD model trained on these Features should have a smaller Marginal L1 loss than the baseline models. For the COX-PH model (Table 1), the performance of all unsupervised features (RN-β-VAE and TCN-β-VAE) and knowledge-based features are close to the performance of the baseline. RN-ICD showed a significantly better performance in all metrics than the baseline. For the N-MTLR model (Table 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 features, 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. Note that the performance metrics for each lead individually (RN-β-VAE-lead#) are lower than the 12 lead signals (TCN-β-VAE) – see Table S1. For all metrics, the ECG measurement features had slightly better performance than RN-β-VAE but had similar performance compared with unsupervised features of TCN-β-VAE.

Table 1: Survival Prediction performance using various generated embedding approaches from ECG signals to predict time until death using the COX-PH model.

Feature Approach	**Feature Size**	**Marg. L1 loss**	**C-index**	**IBS**
E2E DL	12×4096	2622.83	0.501	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
Knowledge-based	22	2672.32	0.50	0.22
MST all = 3420	-	2749.90	-	-
MST uncensord = 496	-	2615.52	-	-

Table 2: Survival Prediction performance using various generated embedding approaches from ECG signals to predict time until death using the N-MTLR model.

Feature Approach	**Feature Size**	**Marg. L1 loss**	**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	0.71	0.17
TCN-β-VAE	32	2106.28	0.72	0.17
Knowledge-based	22	2121.62	0.73	0.17
MST all = 3420	-	2749.90	-	-
MST uncensored = 496	-	2615.52	-	-

To achieve the second objective of our study, which is to analyze the impact of the training dataset sample size on the performance of the ISD model, we used the N-MTLR model as it performed better than the COX-PH model. We considered sample sizes of 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 used age and sex features as a baseline to compare the model's performance when no ECG features are used. This baseline served as a reference point to assess the contribution of ECG features to the learning process and enhancement of the ISD model's performance. Figure 3 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 to just using 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 inferior 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. Discussion

Using COX-PH and N-MTLR models and a large ECG dataset of 244,077 patients, we investigated the

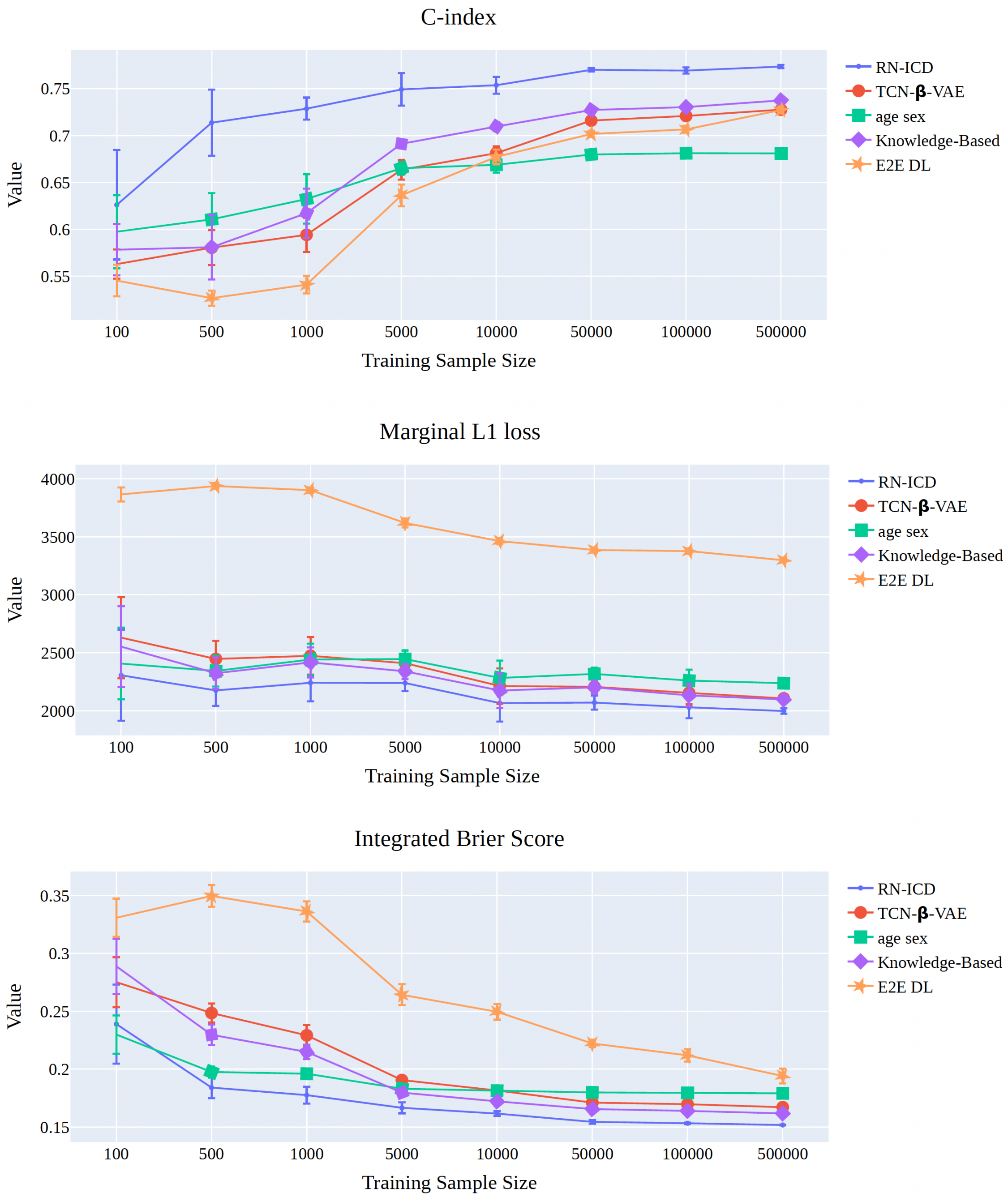

Figure 3: Performance of the N-MTLR model as a function of training sample size using various supervised, unsupervised, and knowledge-based ECG features. The points represent the mean value over the 10 experiments, and the bars represent 95% confidence intervals. (a) C-index, (b) Marginal L1 loss, (c) IBS.

performance of ISD models using raw ECG signals, and ECG features obtained from supervised and unsupervised learning as well 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 diagnostic 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. (2022), 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) (Sun et al., 2023). The results indicate that the performance of unsupervised ECG features (RN-β-VAE and TCN-β-VAE) and knowledge-based features are similar, suggesting there is no clear advantage in using knowledge-based features over unsupervised, because 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 getting better 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 comparable 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.

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 generalizability 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 bot 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 (Goldberger et al., 2000), MIT-BIH (Moody and Mark, 2001), and PTB (Bousseljot et al., 1995), do not include death-related data linked to ECGs. However, it's possible that more comprehensive clinical datasets that include ECGs, like MIMIC (Gow et al., 2023), may become available in the future. These datasets could serve as benchmark data for ECG-based ISD tasks and external validation.

6. Conclusion

In this study, we examined different methods of extracting ECG-based features, which are then used for training ISD models, using a large dataset of 244,077 patients. We used several advanced algorithms to extract these ECG features and found that supervised ECG features performed better than raw ECG signals as well as unsupervised and knowledge-based features. We found that an ISD model trained on ECG features obtained from supervised learning 1414 ICD-10 codes performed the best based on C-index and Integrated Brier score. We also found that a smaller training sample size is sufficient to have a better ISD model when using ECG features obtained from a supervised feature extractor compared with unsupervised feature extractor or knowledge-based ECG features. However, additional research is necessary to evaluate the ECG features obtained from either supervised or semi-unsupervised learning trained on a different task. In summary, this paper can inspire discussions about improving Individual Survival Distributions (ISDs) through various extraction methods and algorithms.

Acknowledgments

This work was supported by NSERC Discovery, IVADO, CIFAR through CCAI Chair funding and by the Alberta Machine Intelligence Institute (Amii)

References

Ralf Bousseljot, Dieter Kreiseler, and Allard Schnabel. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. 1995.

HR Byers, HE Landsberg, H Wexler, B Haurwitz, AF Spilhaus, HC Willett, HG Houghton, Glenn W Brier, and Roger A Allen. Verification of weather forecasts. *Compendium of Meteorology: Prepared under the Direction of the Committee on the Compendium of Meteorology*, pages 841–848, 1951.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Stephane Fotso. Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512*, 2018.

Mitchell H Gail, Louise A Brinton, David P Byar, Donald K Corle, Sylvan B Green, Catherine Schairer, and John J Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *JNCI: Journal of the National Cancer Institute*, 81 (24):1879–1886, 1989.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Seth Berkowitz, Dana Moukheiber, Parastou Eslami, et al. Mimic-iv-ecg-diagnostic electrocardiogram matched subset. 2023.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

Humza Haider, Bret Hoehn, Sarah Davis, and Russell Greiner. Effective ways to build and evaluate individual survival distributions. *The Journal of Machine Learning Research*, 21(1):3289–3351, 2020.

Mohamed Hammad, Asmaa Maher, Kuanquan Wang, Feng Jiang, and Moussa Amrani. Detection of abnormal heart conditions based on characteristics of ecg signals. *Measurement*, 125:634–644, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016.

Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *Advances in neural information processing systems*, 6, 1993.

H Iswaran, UB Kogalur, EH Blackstone, and MS Lauer. Random survival forest. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

Jong-Hwan Jang, Tae Young Kim, Hong-Seok Lim, and Dukyong Yoon. Unsupervised feature learning for electrocardiogram data using the convolutional variational autoencoder. *PloS one*, 16(12): e0260612, 2021.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2011.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

VV Kuznetsov, VA Moskalenko, and N Yu Zolotykh. Electrocardiogram generation and feature extraction using a variational autoencoder. *arXiv preprint arXiv:2002.00254*, 2020.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

George B Moody and Roger G Mark. The impact of the mit-bih arrhythmia database. *IEEE engineering in medicine and biology magazine*, 20(3):45–50, 2001.

Dan M Popescu, Julie K Shade, Changxin Lai, Konstantinos N Aronis, David Ouyang, M Vinayaga Moorthy, Nancy R Cook, Daniel C Lee, Alan Kadish, Christine M Albert, et al. Arrhythmic sudden death survival prediction using deep learning analysis of scarring in the heart. *Nature Cardiovascular Research*, 1(4):334–343, 2022.

Shi-ang Qi, Neeraj Kumar, Ruchika Verma, Jian-Yi Xu, Grace Shen-Tu, and Russell Greiner. Using bayesian neural networks to select features and compute credible intervals for personalized survival prediction. *IEEE Transactions on Biomedical Engineering*, 2023.

Sushravya Raghunath, Alvaro E Ulloa Cerna, Linyuan Jing, David P VanMaanen, Joshua Stough, Dustin N Hartzel, Joseph B Leader, H Lester Kirchner, Martin C Stumpe, Ashraf Hafez, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature medicine*, 26(6):886–891, 2020.

Weijie Sun, Sunil Vasu Kalmady, Nariman Sepehrvan, Luan Manh Chu, Zihan Wang, Amir Salimi, Abram Hindle, Russell Greiner, and Padma Kaul. Improving ecg-based covid-19 diagnosis and mortality predictions using pre-pandemic medical records at population-scale. *arXiv preprint arXiv:2211.10431*, 2022.

Weijie Sun, Sunil Vasu Kalmady, Nariman Sepehrvand, Amir Salimi, Yousef Nademi, Kevin Bainey, Justin A Ezekowitz, Russell Greiner, Abram Hindle, Finlay A McAlister, et al. Towards artificial intelligence-based learning health system for population-level mortality prediction using electrocardiograms. *NPJ Digital Medicine*, 6(1): 21, 2023.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Weijing Tang, Jiaqi Ma, Qiaozhu Mei, and Ji Zhu. Soden: A scalable continuous-time survival model through ordinary differential equation networks. *The Journal of Machine Learning Research*, 23(1): 1516–1544, 2022.

Connie W Tsao, Aaron W Aday, Zaid I Almarzooq, Cheryl AM Anderson, Pankaj Arora, Christy L Avery, Carissa M Baker-Smith, Andrea Z Beaton, Amelia K Boehme, Alfred E Buxton, et al. Heart disease and stroke statistics—2023 update: a report from the american heart association. *Circulation*, 147(8):e93–e621, 2023.

Rutger R van de Leur, Max N Bos, Karim Taha, Arjan Sammani, Ming Wai Yeung, Stefan van Duijvenboden, Pier D Lambiase, Rutger J Hassink, Pim van der Harst, Pieter A Doevendans, et al. Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders. *European Heart Journal-Digital Health*, 3(3):390–404, 2022.

Yi Yang and Hui Zou. A cocktail algorithm for solving the elastic net penalized cox's regression in high dimensions. *Statistics and its Interface*, 6(2):167–173, 2013.

Chun-Nam Yu, Russell Greiner, Hsiu-Chin Lin, and Vickie Baracos. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in neural information processing systems*, 24, 2011.

Appendix A. Training process of unsupervised ECG feature extractor models

As we discussed in the main manuscript, We use two Beta Variational AutoEncoder (β-VAE) model to encode ECG signals. The β-VAE is a variation of the VAE that includes a hyperparameter β in the loss function to disentangle the learned representation. Here, we will describe these models in more details.

A.1. ResNet-Based Model

We use β-VAE architecture with residual connections, following the proposal by Jang et al. (2021), and adapt it for our ECG dataset. This model design is created to learn single-lead ECG signals, and Figures S1 and S2 show modified diagrams of its architecture.

We constructed a separate model for each lead of the 12 lead signals. Training of each model requires configuring two adjustable parameters of the β-VAE: the embedding size and β. We chose the values 64 and 8, respectively. A schematic of the β-VAE training process is shown in Figure S1. 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 loss error. The total loss is a combination of individual losses, and the individual loss function used for training the β-VAE model is negative log-likelihood with a regularizer term, as follows:

$$l_i(\theta, \phi) = -E[\log p_\phi(x_i|z)] + \beta KL(q_\theta(z|x_i)||p(z)) \tag{3}$$

where the first term is the reconstruction loss for the i-th data point. The expectation is considered with respect to the distribution of the encoder over the features. Here, we used the mean squared error as the reconstruction loss. The second term is the Kullback-Leibler (KL) divergence between the prior distribution $p(z)$ and the encoder's distribution $q(z \mid x_i)$, with a regularization term β. In our study, we utilized the mean squared error as the reconstruction loss. After training the β-VAE model, the ECG signal is fed into the encoder of the trained model (with weights frozen), producing 64 pairs of [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 a feature size of 768) (RN-β-VAE). We use these ECG features, along with age and sex, to train the ISD models. We will also compare the performance of ECG features obtained from each individual ECG lead for estimation of ISDs. to (RN-β-VAE-lead#). (Note that we use the # sign to reflect the lead number that is used to train this model.)

A.2. TCN-Based Model

We used β-VAE architecture and code provided by van de Leur et al. (2022), but modified it to reconstruct 12-lead ECG traces of the Alberta Dataset (see Figure S3). The adjustable parameters of β-VAE are the number of features and β, which we set to 32 and 8, respectively, which we chose to be the same as the values used by van de Leur et al. (2022). The training process and loss function were the same as the RN-β-VAE model described earlier. The encoder using a convolutional deep neural network encodes the input ECG signal into 32 pairs [means, variances]. After the training, 12-lead ECG signals were fed into the encoder, and we use the encoded 32 means, along with the patient's age and sex, to estimate that patient's specific ISD curves. We call this approach TCN-β-VAE.

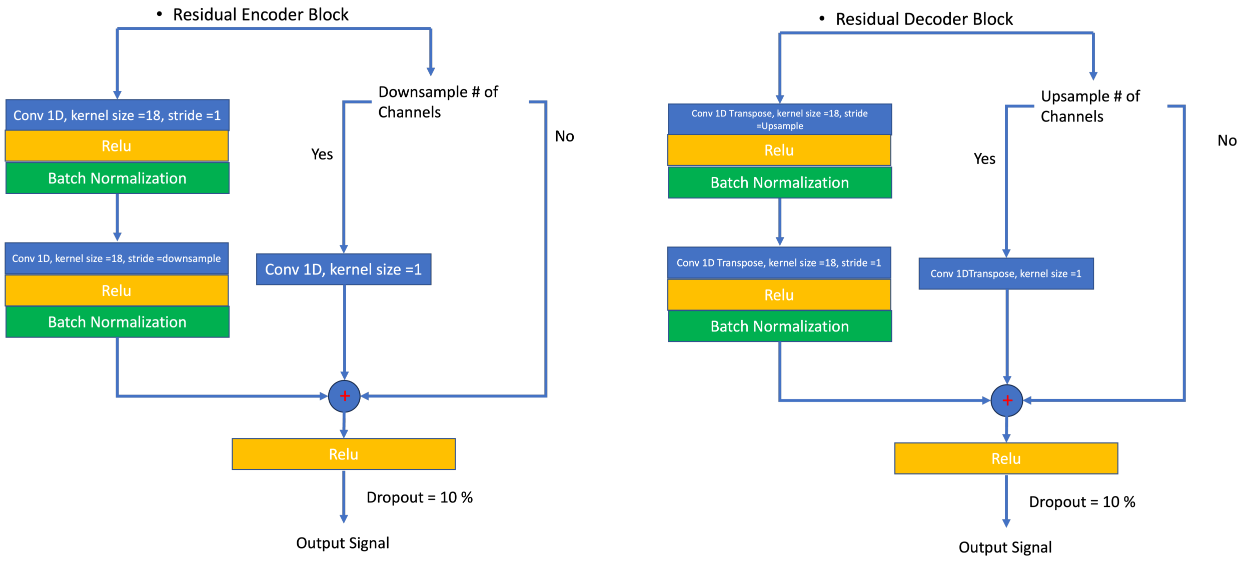

Figure S1: Building blocks of Residual Encoder and Residual Decoder used in Resnet based β-VAE.

Table S1: Survival Prediction performance using various generated embedding approaches from ECG signals to predict time until death using the N-MTLR model.

Embedding Approach	**Feature Size**	**Marg. L1 loss**	**C-index**	**IBS**
RN-β-VAE-lead#1	64	2176.0923	0.7040	0.1755
RN-β-VAE-lead#2	64	2179.3920	0.7015	0.1764
RN-β-VAE-lead#3	64	2179.4895	0.7036	0.1756
RN-β-VAE-lead# aVR	64	2173.2837	0.7053	0.1754
RN-β-VAE-lead# aVL	64	2174.6006	0.7089	0.1753
RN-β-VAE-lead# aVF	64	2173.1548	0.7061	0.1756
RN-β-VAE-lead# V1	64	2170.1717	0.7036	0.1754
RN-β-VAE-lead# V2	64	2165.4921	0.7017	0.1753
RN-β-VAE-lead# V3	64	2167.9394	0.7075	0.1754
RN-β-VAE-lead# V4	64	2194.9370	0.7001	0.1766
RN-β-VAE-lead# V5	64	2160.4920	0.7063	0.1753
RN-β-VAE-lead# V6	64	2181.7161	0.6880	0.1768

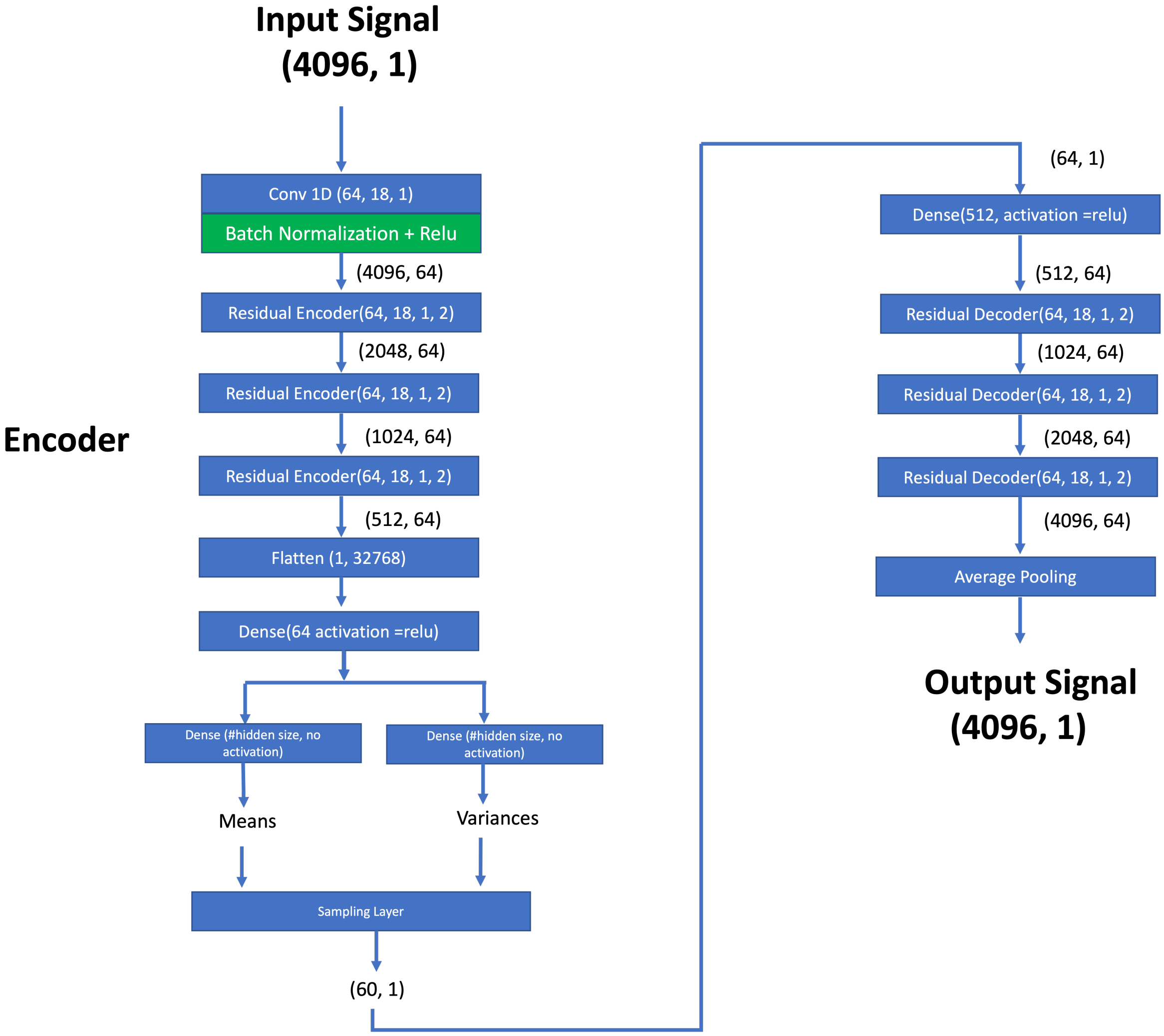
Input Signal
(4096, 1)
Conv 1D (64, 18, 1)
Batch Normalization + Relu
(4096, 64)
Residual Encoder(64, 18, 1, 2)
(2048, 64)
Residual Encoder(64, 18, 1, 2)
Encoder
(1024, 64)
Residual Encoder(64, 18, 1, 2)
(512, 64)
Flatten (1, 32768)
Dense(64 activation =relu)
Dense (#hidden size, no activation)
Dense (#hidden size, no activation)
Means
Variances
Sampling Layer
(60, 1)
(64, 1)
Dense(512, activation =relu)
(512, 64)
Residual Decoder(64, 18, 1, 2)
(1024, 64)
Residual Decoder(64, 18, 1, 2)
(2048, 64)
Residual Decoder(64, 18, 1, 2)
(4096, 64)
Average Pooling
Output Signal
(4096, 1)

Figure S2: Model's Architecture of Resnet based β-VAE.

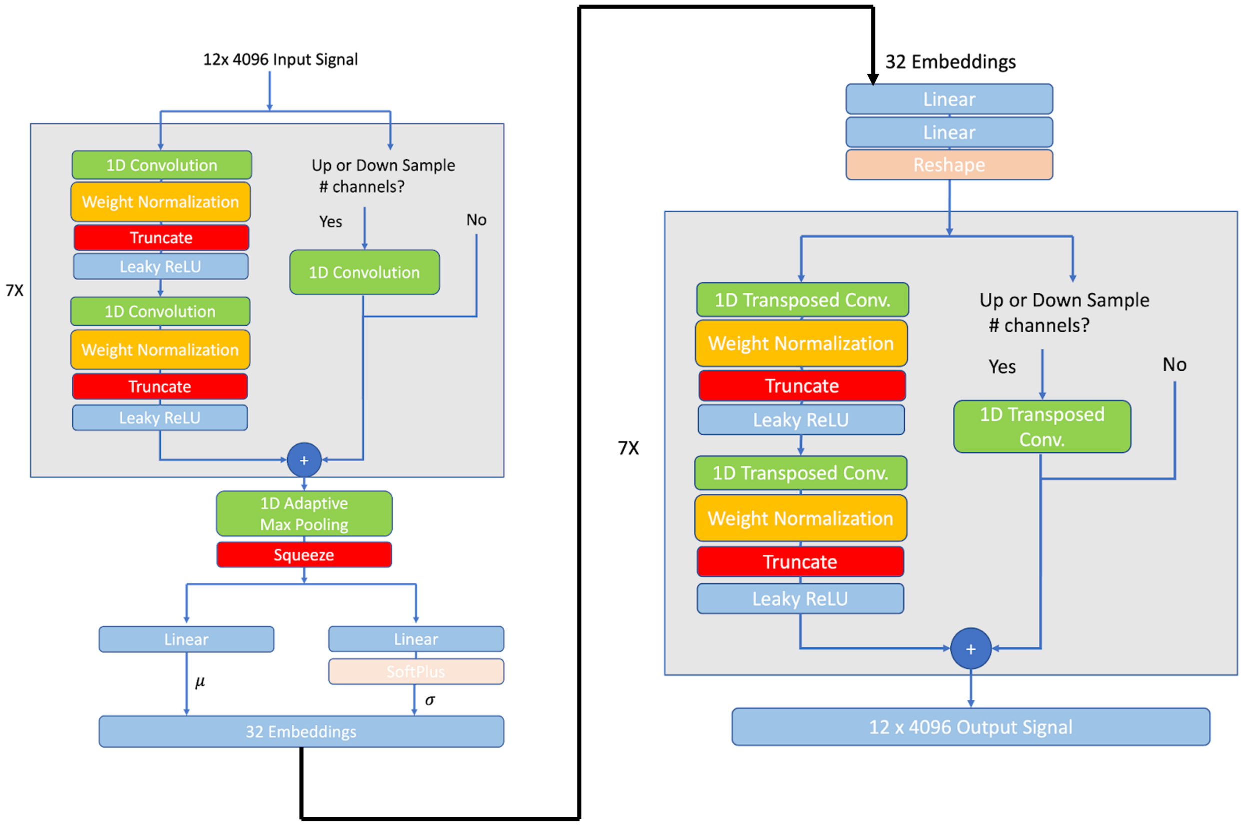

Figure S3: Model's Architecture of Resnet based β-VAE.

Appendix B. Metrics

B.1. 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 person with a shorter observed survival time also has a shorter predicted survival time according to 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:

$$C\text{-index} = \frac{\sum_{i,j}\left(1_{T_j<T_i} \cdot 1_{\eta_j<\eta_i} \cdot \delta_j\right)}{\sum_{i,j}\left(1_{T_j<T_i} \cdot \delta_j\right)} \tag{4}$$

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, which implies that randomly assigning probabilities to instances would result in a 50% probability of correct ordering. A higher C-index value shows a better model performance.

B.2. Marginal L1 Loss

To compute this metric, it is necessary to have the actual event time in order 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. 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 time c.

$$BG(c) = c + \frac{\int_c^\infty S(t)\,dt}{S(c)} \tag{5}$$

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

$$\begin{aligned} L1_{\text{margin}}(D, \hat{t}^{0.5}) = \frac{1}{\gamma}\Big[& \\ & \sum_{j \in D_{\text{uncensor}}} \left|d_j - \hat{t}_i^{0.5}\right| \\ & + \sum_{k \in D_{\text{censor}}} \alpha_k \left|BG(c_k) - \hat{t}_k^{0.5}\right|\Big] \end{aligned} \tag{6}$$

where $\gamma = |D_{\text{uncensor}}| + \sum_{k \in D_{\text{censor}}} \alpha_k$, and α_k shows the level of weight in each estimate based on the Best-Guess for each individual, and d is the true event time. We set $\alpha_k = 1 - \hat{S}_{\text{KM}}(c_k)$ to place more weight on the late censor time instances. The reason for such a weight definition as explained by Haider et al. (2020) is that individuals with early censor time give less information compared to those individuals with late censor time.

B.3. Integrated Brier Score

Brier Score (Byers et al., 1951) 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:

$$BS_t(D, \hat{S}(t|\boldsymbol{x})) = \frac{1}{D} \sum_{(\boldsymbol{x}_i, d_i) \in D} \left(I|d_i \le t| - \hat{S}(t|\boldsymbol{x}_i)\right)^2 \tag{7}$$

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.

$$IBS(\tau, D, \hat{S}(.|.)) = \frac{1}{\tau} \int_0^\tau BS_t(D, \hat{S}(t|.))\,dt \tag{8}$$

If the model accurately predicts all time points, the score will be 1, 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. (1999)

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. (1999)

If the model accurately predicts all time points, the score will be 1, and if the model always predicts 0.5, the score will be 0.25. So, a lower number indicates a better ISD model.