Explainable ECG analysis by explicit information disentanglement with VAEs

Viktor van der Valk, Douwe Atsma, Roderick Scherptong and Marius Staring

***Abstract*— Objective: The interpretation of electrocardiogram (ECG) signals is vital for diagnosis of cardiac conditions. Traditional methods rely on expert knowledge, which is time consuming, costly and potentially misses subtle features. AI has shown promise in ECG interpretation, but clinically desired model explainability is often lacking in literature.**

Methods: We introduce an explainable AI method for ECG classification by partitioning the variational autoencoder (VAE) latent space into a label-specific and a non-label-specific subset. By optimizing both subsets for signal reconstruction and one subset also for prediction while constraining the other from learning label-specific information with an adversarial network, the latent space is disentangled in a supervised manner. This latent space is leveraged to create enhanced visualizations for ECG feature interpretation by means of attribute manipulation. As a proof of concept, we predict the left ventricular function (LVF), a critical prognostic determinant in cardiac disease, from the ECG.

Results: Our study demonstrates the effective segregation of LVF-specific information within a single dimension of the VAE latent space, without compromising classification performance. We show that the proposed model improves state-of-the-art VAE methods (AUC 0.832 vs. 0.790, F1 0.688 vs. 0.640) in prediction and performs comparable to ground truth LVF (concordance 0.72 vs.0.72) in predicting survival.

Conclusion: The model facilitates the interpretation of LVF predictions by providing visual context to ECG signals, offering a general explainable and predictive AI method.

Significance: Our explainable AI model can potentially reduce time and expertise required for ECG analysis.

***Index Terms*— Deep Learning, ECG, Explainable AI, Left Ventricular Function, Myocardial Infarction, Variational Autoencoder,**

I. Introduction

THE electrocardiogram (ECG) stands as a widely employed method for evaluating cardiac morphology and function by capturing the heart's electrical signals with multiple electrodes. Clinicians rely on ECG data for diagnostic and monitoring purposes in various cardiac syndromes, often obtaining a 12-lead ECG as standard practice for disease diagnosis and progression tracking. However, the interpretation of ECG signals traditionally requires expert knowledge, where physicians identify specific patterns associated with disease.

Despite the proficiency of expert analysis, certain crucial information within a 12-lead ECG may elude human interpretation, prompting the exploration of alternative methodologies. Deep learning has demonstrated its efficacy in interpreting ECG signals for various classification tasks, among others, atrial fibrillation, tachycardia, and bradycardia detection [1], [2]. Notably, recent advances in explainable artificial intelligence (AI) algorithms have increased the ability to reveal complex features within ECG signals [3]–[6]. In a medical setting, the transparency and interpretability of AI algorithms are of paramount importance, given the necessity to understand and trust decision-making processes [2].

In this context, β-VAEs [7] have emerged as unsupervised and explainable feature generators for ECG analysis. Studies have shown that a β-VAE trained to reconstruct ECG signals can extract features that become more interpretable by visualizing the reconstructed latent space with the decoder of the β-VAE [3], a process called attribute manipulation. Analysis of these features resulted in a subset that was shown to be predictive of cardiac function, several of which were similar to known indicators of cardiac health. However, features generated for reconstruction purposes only may not be optimally suited for label-specific predictions, such as cardiac function estimation.

This paper aims to enhance the specificity and predictive performance of latent features derived from ECG signals. The task chosen for this optimization is the assessment of left ventricular function (LVF), a critical prognostic determinant in cardiac disease [8]. Traditionally, LVF assessment necessitates advanced imaging techniques and expert interpretation, however it was shown that ECG signals contain information on LVF as well [9]. A clear understanding of the relationship between LVF and ECG characteristics could facilitate patient monitoring. Thereby, derivatives of the ECG signal show promise in remote monitoring with smart devices. Leveraging these two could enable remote monitoring of LVF in patients, especially in home-based settings.

We propose a method to partition the latent space into label-specific and non-label-specific parts, aiming to enhance the

Submitted on 04-06-2024. This work was supported in part by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 860173

Viktor van der Valk is with the Department of Radiology, Leiden University Medical Center, Albinusdreef 2, 2333ZA Leiden, The Netherlands. (email: $v.o.van\_der\_valk@lumc.nl$)

Douwe Atsma is with the Department of Cardiology, Leiden University Medical Center, Albinusdreef 2, 2333ZA Leiden, The Netherlands.

Roderick Scherptong is with the Department of Cardiology, Leiden University Medical Center, Albinusdreef 2, 2333ZA Leiden, The Netherlands.

Marius Staring is with the Department of Radiology, Leiden University Medical Center, Albinusdreef 2, 2333ZA Leiden, The Netherlands.

specificity of the latent features. By optimizing both part for signal reconstruction and only one for label (LVF) prediction, while constraining the other part from learning label-specific information through an adversarial network, we aim to enhance the specificity of these features. By means of attribute manipulation, selective sampling from the label-specific latent space only, this latent space can be translated into ECG features. These features are tailored to each individual ECG and demonstrate the algorithm's classification rationale and put the heartbeat in a population context of heartbeats with varying LVF status, thereby enhancing interpretability.

In summary, this study improves the subdivision of the latent space, resulting in supervised disentanglement of label-specific features, which allow improved visualizations by means of attribute manipulation. These visualisations improve interpretability of label-specific features, by illustrating the classification rationale specific to each case. The features are additionally validated with survival analysis. This research extends previous research presented at the MICCAI 2023 conference [10].

A. Related work

Representation learning, which involves learning subspaces of data that compress or disentangle input data, is used in both generative and explainable AI (XAI). This dynamic field has produced several prominent algorithms and methods, including but not limited to StyleGANs, VAEs, and contrastive learning [11]–[14]. These techniques have also been successfully applied in the medical domain, particularly in ECG analysis [3]–[5], [15]–[18].

In addition to these, other XAI methods have proven effective in ECG analysis. For a comprehensive review of these methods, we refer to Ayano *et al.* (2022) [19]. Wagner *et al.* (2024) [20] also provide an extensive analysis of various XAI techniques for ECG feature extraction. These studies underscore the importance of deriving clinically relevant insights from AI models applied to ECGs.

This paper builds on the principles of XAI and disentanglement research to enhance the interpretability and usability of representations learned by VAEs. In recent years, VAEs have gained significant attention for their ability to generate and learn meaningful data representations. This section reviews the application of these methods with labeled data for classification tasks, with a particular focus on explainable and ECG classification.

Several studies showed the use of classic VAEs as an unsupervised ECGs feature extractor [4], [5], [9], [17], [18]. These features showed to be predictive for cardiac function, myocardial infarction, mortality and several arrhythmias. Moreover, creating artificial data by sampling from the VAE latent space was shown to be a successful ECG data augmentation strategy [6]. The more recent vector-quantized VAE (VQ-VAE) was also successfully used as an unsupervised feature extractor for arrhythmia classification and data augmentation [21].

However, all the studies mentioned above assume that optimizing features for reconstruction implicitly leads to the optimal aggregation and isolation of label-specific information in these features. This is not the case as was shown in Van der Valk *et al.* (2023) [10]. Moreover, the use of a single latent space from which relevant dimensions are selected after training, a method used in some of the studies mentioned above, has its limitations w.r.t. interpretability. It might allow visualization of ECG features important for label prediction by means of attribute manipulation, but label-specific information will most likely be spread over several dimensions, which possibly show complex interaction. This hampers interpretability of these features.

The use of supervision in VAE training could potentially address these issues. Conditional VAEs (CVAEs) [22], classification autoencoders [23], the Attribute-based Regularized VAE [24], the Task-Specific VAE [10] and especially the conditional subspace VAE [25], all use a form of supervision and share some similarities with the model proposed here. These include the use of labeled data, the supervised structuring of the latent space, and an adversarial component that prevents information from being captured in a non label-specific latent space. Several studies have explored the use of these models in ECG analysis. However, the majority of these studies focused solely on the generation of ECG data, rather than the classification or interpretation of the ECG signal [6], [26]–[29]. To our knowledge only one study explored the use of supervision in ECG classification with a VAE. Gyawali *et al.* (2018) [30] classified the origin of ventricular activation in the 12-lead ECG with two siamese CVAEs. They extended the CVAE pipeline with a separate deterministic encoder that learned to predict the label, which was input to the CVAE encoders.

Our approach differs from previous work by explicitly optimizing a VAE architecture with a labeled subspace for classification. We introduce novel elements such as using different KL-divergence loss weights for these subspaces and leveraging the interpretability of VAEs for generating signal-specific rather than population-specific explainable ECG visualizations in the context of classification.

II. Methods

A. Data

The study uses a partially labeled dataset. The unlabeled subset comprises 119,886 raw 10-second, 12-lead ECG signals recorded at a frequency of 500Hz. These signals are obtained from 7,255 patients (71% male, age 64.2 $\pm$ 12.2) between 2010 and 2022 at the Leiden University Medical Center in the Netherlands, who were diagnosed with acute coronary syndrome. The labeled subset set consists of 33,610 labeled ECGs from 2,736 patients within the same cohort. The label, the level of LVF impairment, consists of 4 ordinal categories: normal, mild, moderate, and severe. Each ECG is labeled through visual assessment of echocardiograms conducted within 3 days before or after the ECG recording. A 1-day instead of a 3-day margin is applied when the ECG is obtained within two weeks after cardiac intervention, as the LVF can still change considerably in the first weeks after cardiac intervention. Cases in which a cardiac intervention occurred between ECG and echocardiography are excluded. The relative frequencies of the LVF status were 32.8%, 56.1%, 9.0% and 2.1% for

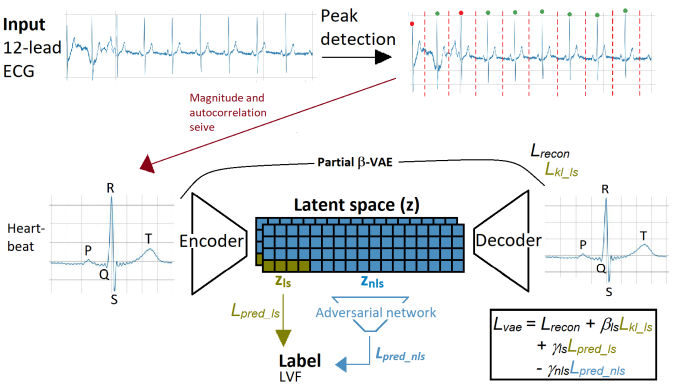


Fig. 1: **Overview of our Processing and Prediction Framework:** Initially, the 12-lead ECGs are split into individual heartbeats using RPNet [31]. Subsequently, the heartbeats are sieved in two stages: the first based on individual magnitude, identifying low magnitude as indicative of a flat signal, and the second utilizing autocorrelation to assess correlation with other heartbeats from the same signal. The remaining heartbeats (indicated with green dots) are used as input for the VAE. The encoded representation, the latent space z, is divided in a label-specific part, z_{ls} (green part), and non label-specific part, z_{nls} (blue part), i.e. z=[z_{ls}, z_{nls}]. While z_{ls} directly contributes to LVF label prediction with loss $L_{\text{pred\_ls}}$, z_{nls} serves as input for an adversarial MLP also predicting the LVF label, introducing loss $L_{\text{pred\_nls}}$. The decoder's input is sampled from z to produce the reconstructed heartbeat, subsequently compared with the original input heartbeat, resulting in the reconstruction loss L_{recon}.

respectively normal, mild, moderate, and severely impaired LVF. For both data sets, an all-cause mortality indicator is available, at least up to three years after hospital discharge, which was used in survival analysis. The study protocol, identified as nWMODIV2_2022006, received approval from the institutional review board, which waived the requirement to obtain informed consent for the use of ECG data of individual patients.

B. Data preprocessing

The initial processing of raw ECG signals involves segmentation into individual heartbeats, defined as the 400-ms intervals preceding and following the R-peak, the prominent peak in the ECG denoting ventricular depolarization, see Figure 1. All recordings are resting ECGs and the average RR-interval in the dataset is 801ms. Therefore, a 800-ms interval is chosen to strike a balance between selecting too little and thereby missing parts of the heartbeat and selecting too much and including a previous or subsequent beat. This segmentation is achieved through a peak detection method inspired by RPNet, a U-Net-structured CNN with inception blocks. RPNet was previously trained by the authors on manually labeled peak locations, as detailed in Vijayarangan *et al.*, (2020) [31]. The R-peaks are determined in lead I for all leads to prevent misalignment when the peaks are not accurately detected.

The segmented heartbeats can contain noisy or flat signals in one or more of the leads, which is the result of improperly attached electrodes or movement during acquisition. To automatically discard these heartbeats and ensure that only relevant and adequately characterized heartbeats contribute to subsequent analyses and model training, a magnitude and an autocorrelation sieve are applied. The magnitude sieve eliminates heartbeats that show a zero signal by sieving out heartbeats with an average magnitude below 1. The autocorrelation sieve discards signals exhibiting a mean autocorrelation between heartbeats below a predefined threshold. This threshold is set to 0.85, which is determined by visual inspection of the resulting heartbeats. A lower threshold does not remove enough noise signals, whereas a higher threshold removes too many true heartbeats that show naturally occurring variation. ECG signals can potentially show various forms of arrhythmia, which would result in a low autocorrelation between heartbeats. As the aim of this study is to look at the morphology of the heart, these irregular heartbeats are also sieved out. Following the application of these sieves, as a form of data augmentation, a maximum of 5 from the 2-8 retained heartbeats are randomly added to the training set, this number is chosen to prevent over-representation of similar heartbeats from the same ECG signal. Here a heartbeat consists of all 12 leads of the ECG signal and whenever 1 lead shows a loud or flat signal, the whole heartbeat is discarded, meaning all 12 leads.

C. Model overview

We introduce an innovative method for the analysis of ECG signals by employing a partial β-VAE, see Figure 1. We propose an extension of the β-VAE pipeline to incorporate labeled data, such that classification can be done with the encoder of the VAE, contrary to conditional VAEs. The primary objective of this approach is to promote the disentanglement of label-specific information within the latent representation, which can then be used to improve ECG signal classification. By sampling from a disentangled latent space, it becomes feasible to generate a spectrum of points exhibiting diverse label-specific information while maintaining consistent non label-specific information. Subsequent reconstruction of these sampled points using the VAE decoder results in a variety of ECG signals, which prove valuable for explainability purposes in ECG signal classification. In generative AI, this reconstruction process is also referred to as attribute manipulation, which will here be used to add explainability to the classification process.

To do so, the latent space, denoted z, is divided into two distinct domains: i) The label-specific latent space (z_{ls}) and ii) the non-label-specific latent space (z_{nls}), see Figure 1, with dimensions of L_{ls} and L_{nls} respectively. To extract label-specific information from the ECG signal, the VAE is extended at the bottleneck with a single fully connected layer and a softmax layer. These additional layers are trained to predict the label based on z_{ls}. Consequently, the encoder is tasked with estimating both prior distributions $p(z_{\text{ls}})$ and $p(z_{\text{nls}})$ as well as inferring the label y from the input heartbeat X, modeled

as:

$$z_{\text{ls}}, y \sim q(z_{\text{ls}}, y|X) \text{ with } y \sim q(y|z_{\text{ls}}) \text{ and } z_{\text{nls}} \sim q(z_{\text{nls}}|X).$$

The decoders objective, reconstruction of the input X, remains consistent with that of the original VAE and is defined as:

$$X \sim p(X|q(z|X)).$$

The method is designed to be portable to other VAE architectures, but will here be showcased with a relatively basic convolutional VAE. The encoder of the VAE is a CNN with 8 1D convolutional layers followed by 6 2D convolutional layers. This design choice ensures that the leads of the ECG signal undergo mixing only after traversing multiple layers, preserving the integrity of the temporal features of the signal. The decoder is an exact mirror of the encoder. The latent space at the bottleneck has size $L_{\text{ls}} + L_{\text{nls}}$ and the complete VAE consists of 18.6M parameters in total.

To further enforce the isolation of label-specific information in z_{ls}, the VAE is jointly optimized with an adversarial multilayer perceptron (MLP). This adversarial network is tasked with predicting the label exclusively from z_{nls}, see Figure 1. The negative of this adversarial loss is integrated into the overall VAE loss, discouraging any capture of label-specific information in z_{nls}. Through this dual optimization strategy, label-specific information is explicitly disentangled within the latent space z, providing a comprehensive and effective representation for the analysis of ECG signals.

D. Training of the Partial β-VAE

Training of the partial β-VAE involves the incorporation of multiple loss terms. These encompass: i) L_{recon}, the mean squared reconstruction error (MSE) quantifying the dissimilarity between the input and output ECG heartbeat, ii)L_{kl_ls} the Kullback-Leibler (KL)-divergence loss, $D_{\text{KL}}(P||Q)$ characterizing the disparity between the posterior, $q(z|X)$, and the prior, $p(z)$, which is set to be $\mathcal{N}(0, 1)$, iii) a prediction loss for label prediction derived from z_{ls}, and iv) an adversarial prediction loss for label prediction originating from z_{nls}. The first two losses are calculated for the whole dataset and the prediction loss only for the labeled subset.

Consistent with the original work by Higgins et al. [7], the KL-divergence loss is weighted by a β factor. The KL-divergence loss is only calculated for the label specific latent space and is calculated as $D_{\text{KL}}(p(z_{\text{ls}})||q(z_{\text{ls}}|X))$

The labeled subset of the data is annotated with ordinal labels (normal, mild, moderate, and severe impairment), therefore the ranked probability loss function, L_{rps} proposed by Galdran [32] is used here. This function is defined as:

$$L_{\text{rps}} = \frac{1}{nk} \sum_{i}^{N} \sum_{j}^{K} \|\mathbf{P}_{ij} - \mathbf{Y}_{ij}\|_2^2,$$

where K denotes the number of classes, N represents the number of patients, j the one-hot encoded label position, $\mathbf{Y}_i$ signifies the ordinal encoding of the labels ([0,0,0,1], [0,0,1,1], [0,1,1,1], [1,1,1,1]), and $\mathbf{P}_i$ corresponds to the cumulative sum of the output class probabilities. We propose to use a combination of the ranked probability loss and the categorical cross-entropy loss, given as:

$$L_{\text{pred}} = L_{\text{rps}} + L_{\text{CE}}.$$

This choice is informed by optimization experiments which demonstrated its effectiveness. Two prediction losses are used in training. $L_{\text{pred\_ls}}$ is employed for label prediction from z_{ls} and is weighted with a γ_{ls} factor. $L_{\text{pred\_nls}}$ is employed for the adversarial MLP predicting the label from z_{nls} and is weighted with a negative γ_{nls} factor to deter the assimilation of label-specific information in z_{nls}. The MLP is optimized with the same $L_{\text{pred\_nls}}$. This results in the MLP minimizing the prediction loss and the VAE maximizing the prediction loss of the MLP. Since the VAE controls the information that is contained in z_{nls}, maximizing the prediction loss of the MLP will result in minimizing the label specific information in z_{nls}. However, given the dynamic nature of z_{nls} during training and therefore the varying input to the MLP, the MLP is trained for 5 epochs after each epoch of the VAE training phase.

The overall loss function for the VAE is defined as follows:

$$L_{\text{VAE}} = L_{\text{recon}} + \beta_{\text{ls}} D_{\text{KL}}(p(z_{\text{ls}})||q(z_{\text{ls}}|X)) + \gamma_{\text{ls}} L_{\text{pred\_ls}} - \gamma_{\text{nls}} L_{\text{pred\_nls}}.$$

For the MLP, the loss function is expressed as:

$$L_{\text{MLP}} = L_{\text{pred\_nls}}$$

E. Feature visualization

To visually explore the pertinent features in the ECG signal crucial for LVF prediction, we employ attribute manipulation on a per-signal basis. The application of attribute manipulation in this context differs from the conventional methods employed in data generation. Here it provides a distinctive approach that offers a clear per-signal visualization of the classification rationale adopted by the encoder. The procedure involves encoding each ECG heartbeat into z_{nls} and z_{ls}, and subsequently decoding the encoded signal to obtain the reconstructed signal, denoted as S_{recon}. We assume $L_{\text{ls}} = 1$ here, but other sampling strategies could be adopted when $L_{\text{ls}} > 1$. The above mentioned procedure is applied iteratively to six distinct sets of z_{ls}, covering the mean values for the four LVF groups (S_1, S_2, S_3, S_4), as well as two additional "extreme" z_{ls} values. These extreme values are obtained by sampling the mean of the severely impaired LVF group minus its standard deviation (S_0) and the mean of the normal LVF group plus its standard deviation (S_5). Consequently, a spectrum of reconstructed ECG signals is generated, representing various degrees of LVF impairment. This spectrum can be interpreted as the algorithm's utilization of distinct features within the signal for LVF classification.

F. Survival analysis

In addition to the primary evaluation metrics, we perform a supplementary analysis to validate the performance of the latent space representation (z_{ls}) by means of survival analysis. The rationale behind this approach lies in understanding that LVF status serves as a significant indicator of cardiac health

and exhibits predictive value for mortality. The anticipation is that the performance of z_{ls} in predicting mortality will align closely with the predictive capacity demonstrated by the LVF labels. Survival analysis assesses the time until an event of interest occurs, such as mortality, and compares survival distributions between different groups. The primary survival analysis method employed in this study is the Kaplan-Meier estimator, a non-parametric approach, which provides an estimate of the survival function. The log-rank test is used to compare survival curves between different groups. The survival function, denoted as $S(t)$, represents the probability of survival beyond time t. The evaluation of the survival function is done with the concordance index, which evaluates how well a model ranks the survival times of patients. The Cox proportional hazards model is also applied to assess the impact of z_{ls} on survival outcomes while considering covariates. The model is represented as:

$$h(t) = h_0(t) \exp(\beta_{\mathrm{cox}} z_{\mathrm{ls}}),$$

where $h(t)$ is the hazard function at time t, $h_0(t)$ is the baseline hazard function, β_{cox} represents the hazard ratio for z_{ls} [33].

III. Experiments and Results

A. Experiments

1) Experimental settings: In all experiments the combined latent space, z, with a dimensionality of L was set to 600, large enough to ensure it was not a limiting factor in reconstruction. Both data sets were combined and divided into a training set (85%) and a test set (15%). The 5-fold cross-validation was performed with the training set with again a ratio of 85:15 between the training and the validation set. The validation set was split into two equal subsets, one used for early stopping and one used for hyper-parameter optimization. All data splits were grouped by patient, which means that no patient in the training set was used in the validation or test set, and stratified by label in the case of labeled data splits. Training was done until convergence, i.e. until the loss on the validation set stopped improving for 30 epochs. To avoid overfitting, balanced sampling, early stopping, regularization with the Adam optimizer with weight decay, and L2 regularization of the fully connected layer were used. To prevent gradient explosion He initialization was used [34], to prevent overflow, the standard deviations of the posterior $q(z|X)$ were restricted to a [-10, 3] interval before the sampling step.

The networks were build and trained in the PyTorch 2.1.0 framework and trained on a Quadro RTX 6000 GPU with CUDA 12.1 [35], [36]. The implementation of our models will be made publicly available via GitHub at `https://github.com/ViktorvdValk/Interpretable_VAE_for_ECG`.

B. Feature evaluation

The LVF predictions of the models are evaluated with 3 classification metrics. The Area Under the Receiver Operator Characteristic Curve (AUROC) and the F1 score are used for evaluation of the binary split (severe/moderate vs. mild/normal), given the clinical significance of this split and to facilitate comparison with prior studies. Significant difference between AUROC scores was calculated as proposed by DeLong *et al.* (1988) [37]. Another prediction metric, the relative correctness rate, RCR, that takes into account the ordinality of the classes, based on the prediction MSE between the predicted and the ground truth class [38], is used to assess the ordinal classification of the label. The RCR is class balanced and is designed so that the perfect classification results in $\mathrm{RCR} = 1$ and mean classification, ($\hat{y} = \overline{y}$), results in $\mathrm{RCR} = 0$. RCR is defined as follows:

$$\mathrm{RCR} = 1 - \frac{1}{4} \sum_{j=0}^{3} \frac{1}{N_j} \sum_{i \in \mathcal{S}_j} \left(\frac{y_i - \hat{y}_i}{y_i - \overline{y}} \right)^2 .$$

Here, $\mathcal{S}_j$ is the set of all patients with LVF label j with j = severe, moderate, mild, normal and $N_j = |\mathcal{S}_j|$. The reconstruction of the heartbeats by the models is evaluated with the mean squared reconstruction error (MSE). For inter-model comparison of the MSE, the paired t-test was used. For 95% confidence interval calculation bootstrapping was used with $n = 1000$. In line with the output class used in L_{rps}, the cumulative sum of the models' output class probabilities was used to predict the final class for RCR calculation.

1) Impact of the β factor: In this experiment, we investigate the influence of β_{ls}. Notably, given that the pipeline does not incorporate sampling from the posterior $q(z_{\mathrm{nls}}|X)$, it can be argued that structuring this portion of the latent space is unnecessary, particularly if it compromises reconstruction quality, as observed in prior work by Van der Valk *et al.* (2023) [10]. Consequently, in the proposed algorithm, β_{nls} was set to 0. Furthermore, we examined the impact of β_{ls} on prediction and reconstruction to elucidate the importance of structuring and disentangling the latent space (z_{ls}), which is already subjected to a prediction loss.

2) Impact of the label-specific split: We investigated the advantages of implementing a label-specific split within the latent space z. Although such a split constrains both reconstruction and prediction capacity, it reduces the susceptibility to overfitting and enhances the interpretability of the model, particularly when L_{ls} is small. L_{ls} plays a crucial role in promoting structure within z_{ls}, thereby facilitating sampling and enhancing the interpretability of the reconstructed signals. The interpretability of the visualization of the features made with attribute manipulation is directly related to L_{ls}, because of the influence of the dimensionality of z_{ls} on the sampling process. Interaction effects between dimensions of z_{ls} when $L_{\mathrm{ls}} > 1$, make attribute manipulation less interpretable.

3) Ablation experiments: In the ablation experiments we investigate the influence of the prediction loss and the adversarial network. Both extensions to the pipeline promote either direct of indirect disentanglement of the prediction relevant information in the input. Since L is large and thus not a limiting factor, it can be assumed that the addition of the adversarial network forces the model to use the label-specific latent space for signal reconstruction also. This is assessed using a linear regression classifier that aims to predict the label from z_{nls}.

TABLE I: Results for different values of β_{ls}, the KL-divergence weight for z_{ls}, are shown here for both reconstruction (MSE) and prediction (AUROC, F1 and RCR). All metrics show the average (and 95% confidence interval) of 5-fold cross-validation on the validation set. Significant difference, $p < 0.025$, in the AUROC and MSE metrics from β_{ls}=0 for all folds is indicated with an asterisk (*)

β_{ls}	MSE ↓	AUROC ↑	F1 ↑	RCR ↑
0	**2.10**	**0.868**	**0.737**	**0.637**
	[2.09-2.11]	[0.864-0.871]	[0.734-0.741]	[0.630-644]
0.1	2.96*	0.811*	0.596	0.498
	[2.95-2.97]	[0.808-0.815]	[0.592-0.600]	[0.490-0.505]
0.5	2.37	0.501*	0.469	0.275
	[2.36-2.38]	[0.495-0.506]	[0.468-0.469]	[0.267-0.283]
1	2.30	0.556*	0.469	0.275
	[2.29-2.31]	[0.551-0.556]	[0.469-0.469]	[0.266-0.283]

4) Baseline methods: To give context to the reconstruction, prediction and interpretability values, our method is compared against both principal component analysis (PCA) and the VAE XAI as proposed in Van der Leur *et al.* (2022) [9]. PCA can be considered a general, unsupervised deterministic feature extractor that does not use any form of machine learning. The VAE XAI is a β-VAE that is specifically designed for unsupervised feature extraction from the ECG signal. The features generated with the VAE XAI were shown to contain information of the LVF in a general population of patients who had an ECG taken. In their paper the authors reported an AUC of 0.90 when predicting the LVF from the 12-lead ECG with a latent space of 32. The authors also reported a similar AUC (0.91) for a supervised CNN trained with the LVF labels. Since both PCA and the VAE XAI do not have a classification layer at the bottleneck, a linear regression classifier is used to predict the label from z. For PCA the space spanned by the principal components is used as the latent space z with L the number of principal components used for reconstruction and prediction.

C. Results

1) Hyperparameter tuning: The prediction and reconstruction tasks have different complexity and data scales. Several adjustments are made to prevent overfitting on the prediction task, which is less complex and is trained on a subset of the data used for reconstruction. γ_{ls} is set to 0 for the first 100 epochs, then gradually increases to 2 in the 400 epochs following. This schedule helps to prevent overfitting on the prediction task while ensuring enough training epochs for the reconstruction task. The value of γ_{ls} is tuned to achieve optimal network prediction performance. γ_{nls} is set to 15 from the beginning, as the adversarial loss does not promote overfitting of the prediction task. Setting γ_{nls} larger than 15 did not improve the capture of any label-specific information in L_{nls}.

Table I presents the impact of the KL-divergence weight β_{ls} on reconstruction and prediction. Both reconstruction and prediction seem better with small values of β_{ls}. The results indicate that reconstruction and especially prediction are limited by the KL-divergence loss. $\beta_{\text{ls}} = 0$ gave both best reconstruction and best prediction results. Therefore, $\beta_{\text{ls}} = 0$ will be used in the rest of the paper.

TABLE II: Results for different values of L_{ls} (and thus L_{nls}, given that $L_{\text{ls}} + L_{\text{nls}} = 600$) are shown here for both reconstruction (MSE) and prediction (AUROC, F1, RCR). Setting $L_{\text{ls}} = 600$ essentially means not spitting L, since the whole latent space is then optimized for both reconstruction and prediction. All metrics show the average (and 95% confidence interval) of 5-fold cross-validation on the validation set. Significant difference, $p < 0.025$, in the AUROC and MSE metrics from L_{ls}=1 for all folds is indicated with an asterisk (*)

L_{ls}	MSE ↓	AUROC ↑	F1 ↑	RCR ↑
1	2.10	0.868	**0.737**	0.637
	[2.09-2.11]	[0.864-0.871]	[0.734-0.741]	[0.631-0.643]
2	2.02	0.871	0.726	0.637
	[2.01-2.03]	[0.868-0.874]	[0.722-0.730]	[0.631-0.643]
5	1.99	0.872	0.723	**0.658**
	[1.98-2.00]	[0.870-0.875]	[0.720-0.727]	[0.652-0.664]
600	**1.87**	**0.873**	0.719	0.649
	[1.86-1.88]	[0.870-0.876]	[0.715-0.722]	[0.643-0.654]

Table II summarizes the effect of L_{ls} on reconstruction and prediction. Setting $L_{\text{ls}} = 600$ effectively eliminates the split, as the full latent space is optimized for both reconstruction and prediction. Reducing L_{ls} imposes a constraint on the VAE's reconstruction capacity; however, this effect is minimal, as indicated by the slight reduction in reconstruction quality for $L_{\text{ls}} = 1$ compared with $L_{\text{ls}} > 1$. Importantly, lowering L_{ls} does not compromise prediction performance, suggesting that all label-specific information can be captured in a single latent dimension. Minimizing L_{ls} also improves interpretability (as explained in the Experiments section) and, for these reasons, $L_{\text{ls}} = 1$ was used in all subsequent experiments.

2) Ablation results: Table III shows the results of different ablations on the pipeline. Separately, the prediction loss and, to a lesser extent, the addition of the adversarial network contribute to enhancing the prediction quality of the model. When both are used together, the addition of the adversarial network provides negligible benefit to the prediction quality over a network trained solely with the prediction loss. However, less overfitting was observed on the prediction task, when both were used in the pipeline. The benefit of each addition on prediction performance comes at the cost of the reconstruction quality, which is slightly reduced by each term. This degradation occurs because these components impose additional constraints on the latent space z. The prediction quality from z_{nls}, as indicated by AUROC (Adv. Netw.) in Table III, is similar in models with and without the adversarial loss. Preventing the model from capturing label-specific information in z_{nls} appears to be a challenging task in which the adversarial network does not fully succeed.

3) Benchmark comparison: Table IV shows the comparison of the proposed model with PCA and with the VAE XAI [9] on the test set. The proposed model performs better than the state-of-the-art in LVF prediction, using only one latent dimension to store label specific information, instead of 32 (VAE XAI) or 600 (PCA). This confirms that the proposed model can successfully separate (the majority of) the label

TABLE III: Results of ablation experiments. The influence of the adversarial network (adv. netw.) and γ are shown here for both reconstruction (MSE) and prediction (AUROC, F1, and RCR). The AUROC for the label prediction with z_{nls} is also shown. All metrics show the average (and 95% confidence interval) of 5-fold cross-validation on the validation set. Significant difference, $p < 0.025$,in the AUROC and MSE metrics from the proposed method for all folds is indicated with an asterisk (*)

Adv. Netw.	γ_{ls}	MSE ↓	AUROC ↑	AUROC (z_{nls}) ↓	F1 ↑	RCR ↑
Yes	2	2.10 [2.09-2.11]	0.868 [0.864-0.871]	**0.857** [0.854-0.860]	**0.737** [0.734-0.741]	**0.637** [0.631-0.642]
Yes	0	0.96* [0.94-0.98]	0.753* [0.743-0.764]	0.837 [0.829-0.844]	0.668 [0.659-0.677]	0.442 [0.435-0.450]
No	2	1.92 [1.91-1.92]	**0.870** [0.867-0.873]	0.853 [0.849-0.856]	0.720 [0.716-0.724]	0.630 [0.622-0.637]
No	0	**0.82*** [0.82-0.83]	0.535* [0.530-0.540]	0.853 [0.850-0.857]	0.469 [0.469-0.469]	0.274 [0.265-0.282]

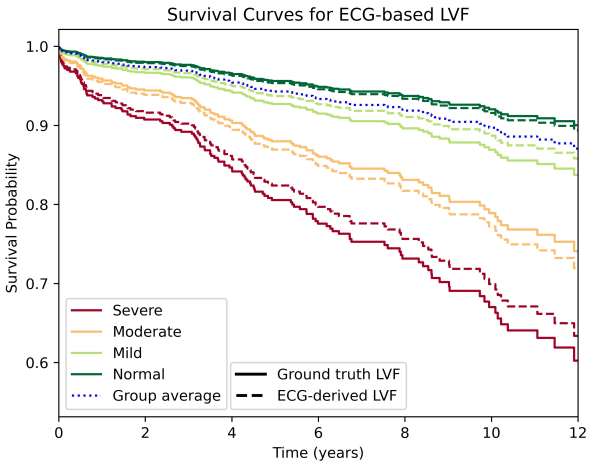


Fig. 2: Survival curves depicting the stratification of patients into distinct groups based on ECG-derived and ground truth LVF predictions. Clear delineation is observed among the four distinct groups. This underscores the predictive capacity of both ECG-derived and ground truth LVF in delineating survival outcomes, revealing significant differences in mortality risk among the identified groups.

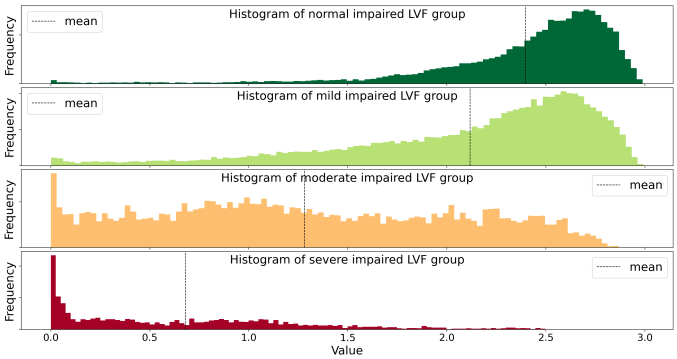


Fig. 3: The histograms show the distributions of the continuous ECG-derived LVF values for the 4 LVF groups. The distributions show some overlap, which indicates that the prediction is reasonably good but not perfect. The means of the distributions are indicated with a dotted line.

specific information and reduce it to 1 dimension. While PCA gives reasonable classification and superior reconstruction results, the visualization of the 600 first principal components is cumbersome and therefore not suitable for meaningful attribute manipulation. On the other hand, the proposed model outperforms the VAE XAI, which would be more suitable for meaningful attribute manipulation, by a large amount in both reconstruction and prediction.

4) Cox regression: Table V presents a comparative analysis between two Cox regression models, one using ground truth LVF and the other the ECG-derived LVF prediction. The concordance index and the significance of hazard ratios demonstrate nearly identical performance between the two models. This observation suggests that the ECG-derived LVF captures mortality-related information in patients to a similar extent as the LVF obtained from echocardiograms. Additionally, Figure 2 illustrates the predictive value of ECG-derived LVF for mortality outcomes in patients following myocardial infarction.

5) Evaluation of interpretability: Figure 3 shows the histograms of predicted LVF values for the 4 LVF groups. The histograms show separation to a reasonable extent, especially between the mild and the moderate groups. The ground truth LVF is a continuous value, which is discretized in 4 groups in clinical practice. The predicted LVF reintroduces this continuous character, which might be of additional value.

Figure 4 presents a visualization example of attribute manipulation of the LVF label using the proposed VAE. Notably, a reconstruction closely tracking the original heartbeat can be seen as a confirmation that the heartbeat is not an out-of-distribution heartbeat, thereby reinforcing the prediction's reliability. The attribute manipulation showcases the distribution of heartbeats across a spectrum of ECG-derived LVF values, which essentially puts the heartbeat in a population context. Heartbeats shaded in yellow to green represent heartbeats classified as having minimal LVF impairment while retaining non-LVF-related characteristics. Conversely, heartbeats shaded in yellow to dark red indicate greater impairment. The magnitude of the discrepancy between the red and green heartbeats underscores the importance of leads and peaks within the heartbeat for LVF classification. This observation not only elucidates the model's main reason behind the classification decisions, but could also be viewed as potential avenues for desired future improvement. Visual inspection of the ECG spectra by an interventional cardiologist confirmed the relevance of the indicated ECG features. Several patterns known to indicate LVF status where recognized in various attribute manipulated heartbeat distributions, see Figure 4. Some of these patterns are routinely used by cardiologists to estimate cardiac function, others were known but not used, since the patterns are deemed

TABLE IV: Comparison with benchmarks. The proposed method is compared with PCA and the VAE XAI network from [9] for both reconstruction (MSE) and prediction (AUROC, F1 and RCR). The AUROC and F1 scores are shown for the prediction of the normal/mild vs. the moderate/severe groups. All metrics show the average (and 95% confidence interval) of 5-fold cross-validation on the test set. Significant difference, $p < 0.025$, in the AUROC and MSE metrics from the proposed method for all folds indicated with an asterisk (*)

Model	L_{ls} (L)	MSE ↓	AUROC ↑	F1 ↑	RCR ↑
Proposed	1 (600)	2.24 [2.18-2.301]	**0.832** [0.829-0.834]	**0.688** [0.681-0.695]	**0.591** [0.581-0.601]
VAE XAI [9]	32 (32)	32.8* [32.0-33.5]	0.790* [0.786-0.794]	0.640 [0.632-0.647]	0.524 [0.514-0.533]
PCA	600 (600)	**0.356*** [0.354-0.356]	0.815* [0.812-0.818]	0.659 [0.652-0.665]	0.480 [0.470-0.489]

TABLE V: Hazard ratios for Cox regression models incorporating both ground truth LVF and ECG-derived LVF, with age included as an additional predictor. All predictors exhibit high significance levels ($p < 0.005$), and the concordance index for survival prediction using ground truth LVF versus ECG-derived LVF predictions demonstrates remarkable similarity. These findings suggest that ECG-derived LVF holds promise as a reliable alternative to echocardiogram-derived ground truth LVF in predicting mortality outcomes

var	**hazard ratio**	**p**	**concordance ↑**
age	1.98 [1.63-2.40]	< 0.005	0.72
ECG-derived LVF	0.59 [0.48-0.71]	< 0.005	
age	2.04 [1.69-2.47]	< 0.005	0.72
ground truth LVF	0.59 [0.48-0.73]	< 0.005	

too subtle to estimate from a single ECG without context. The model's features are complementary to already known ECG patterns that are associated with reduced LVF, such as LBBB and the presence of Q-waves. In such cases the models' features can help with the quantitative assessment of pathological patterns. For example, the T-wave is an important prognostic factor in myocardial infarction, but quantitative assessment of T-wave changes due to myocardial infarction is currently not part of the clinical routine in the non-acute setting. In this case the models output, see Figure 4 was considered helpful in providing context.

IV. Discussion

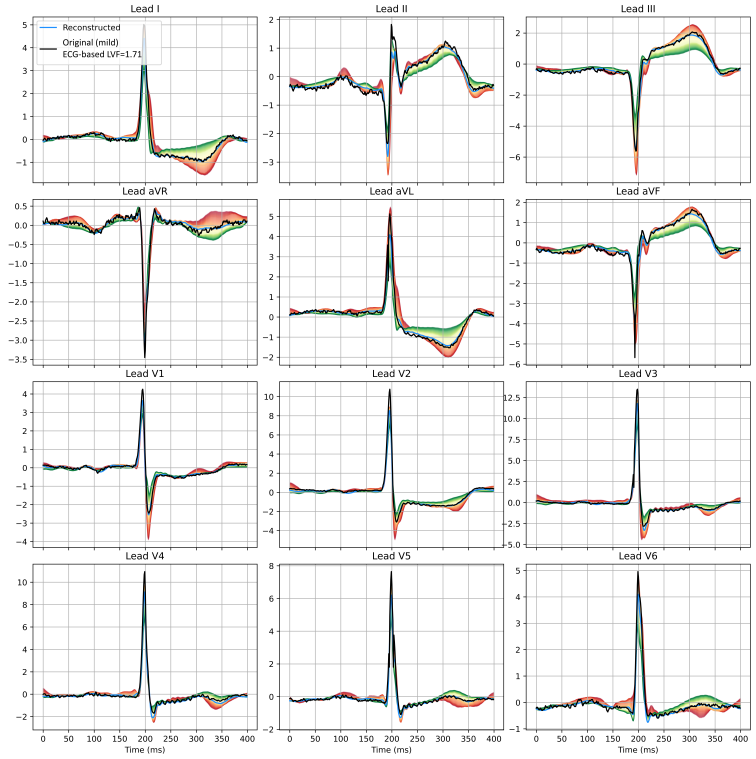
Explicit disentanglement of the latent space through supervised learning enhances interpretability in classification tasks when combined with attribute manipulation. In this study, we illustrate this approach using a VAE framework, where we i) partition the latent space into a label-specific and non-label-specific part, ii) introduce a classification loss, and iii) leverage an adversarial network into the VAE optimization process. Ablation experiments confirm that these extensions collectively contribute to achieving classification (AUROC$= 0.832$) and reconstruction performance (MSE$= 2.24$) comparable or better than the state-of-the-art, while adding interpretation possibilities. i) Previous studies have shown that incorporating a classification loss encourages the explicit encoding of information relevant for classification in the latent space, thereby enhancing classification accuracy compared to unsupervised methods [10], [23]. ii) Subdividing the latent space does not necessarily improve classification performance. However, it isolates and compresses label-specific information into a dedicated part of the latent space, which can be as small as one dimension (see Table II). This partitioning helps visualize the model's decision-making process through attribute manipulation. It places the signal in a population context, making signal-specific features important for classification easier to interpret (see Figure 4). The signal-specific projections used here differ from the population-specific ECG features described by Van der Valk et al. (2023) and Van der Leur et al. (2022) [9], [10]. iii) The inclusion of an adversarial network serves as a regularization mechanism to prevent overfitting on the classification task, as was shown in Table III.

Survival analysis confirmed the effective isolation of LVF-specific information through the ECG-derived LVF, demonstrating performance comparable to the ground truth LVF, as presented in Table V. Although the ECG-based LVF does not serve as a perfect LVF predictor (AUROC $= 0.832$), it encapsulates all survival-related information present in the ground truth.

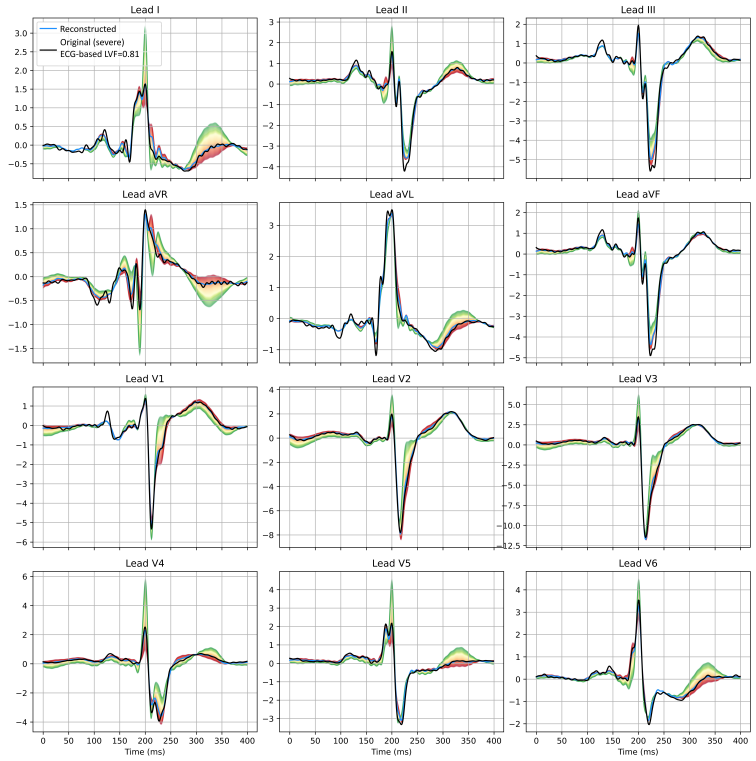
The KL-divergence loss in our framework does not provide the same benefits as in conventional VAEs, as evidenced by the negligible impact of the KL-divergence loss weight (β) in our experiments. This can be attributed to two factors: Firstly, for attribute manipulation we do not sample from the non-label-specific part of the latent space, therefore disentanglement of this part seems unnecessary. Secondly, the prediction loss, already promotes a structure on z_{ls}, as can be seen in Figure 3. Namely, the models prediction is a linear combination of z_{ls} (a single fully connected layer without activation). This likely diminishes the benefit of the KL-divergence on the ordering of the label-specific latent space as well.

In summary, our study shows a novel application of VAEs in explainable (ECG) classification. While previous research has shown the potential of VAEs in feature extraction and data augmentation [6], [26]–[29], our approach goes beyond by explicitly optimizing the VAE architecture with a labeled subspace for classification. Our method is able to do explainable classification on a per-signal basis, while performing on par with or better than the state-of-the-art in ECG classification. We thereby contribute to the advancement of explainable AI algorithms in the medical context, where transparency and interpretability are essential.

The successful outcomes of our study have important implications for clinical practice. By enhancing the specificity of latent features derived from ECG signals, our method allows a more accurate and reliable assessment of LVF, a critical prognostic determinant in cardiac disease [8]. The ultimate aim of the ECG analysis pipeline as proposed here is integration in clinical practice. The lack of context or a



a



b

Fig. 4: Attribute manipulation in 2 heartbeats with mild (a) and severe (b) LVF impairment. The black and blue line respectively show the original and the reconstructed signal. The green to red distributions are the result of attribute manipulation and could be interpreted as the model's rationale to classify the signal with the LVF value shown in the legend. The yellow to green range shows heartbeats that would be classified as having a mildly impaired to normal LVF and the yellow to red range shows heartbeats that would be classified as having a moderate to severely impaired LVF. (a) The ECG shows clear Q-waves, a negative deflection preceding the R-peak, in lead II, III, aVF. The Q-wave pattern is known in clinical cardiology to indicate prior myocardial infarction. The model takes the depth of this peak into account, but also uses the T-wave, the positive deflection after the R-peak. (b) The ECG shows a pattern known as left bundle branch block (LBBB), indicated by the dominant S wave in lead V1, V2 and V3 and the notched R waves in lead I, AVL V5 and V6. The model shows that not only QRS-characteristics, but also T-wave patterns, are important for prediction of the LVF.

'normal or healthy' signal to compare against in plain ECG analysis, is one of the reasons why ECG analysis is hard and necessitates extensive expertise. Providing a 'normal or healthy' signal for comparison presents complexities due to the diverse patient characteristics influencing ECG morphology. Therefore, even in patients not affected by disease, the range of patterns considered normal is wide [39]–[41]. Our pipeline aims to give such a 'normal or healthy' version of the signal, while preserving patient-specific features. When implemented in ECG-software visuals, as shown in Figure 4, can be helpful in providing context to otherwise harder to distinguish ECG characteristics, such as T or Q-wave improvement or decline. Besides, providing context, the visuals also help to assess the quality of the classification and the clinical implication that is attributed to it. Assessment of the difference between the original signal and the reconstruction serves as a quality control measure. A large difference can be used as an indicator of classification inaccuracies, which in some cases are related to signal noise. Clinical integration of our pipeline holds promise for reducing both the time and expertise required for ECG analysis, facilitating patient monitoring, and enabling remote monitoring in home-based settings where remote monitoring with smart devices is becoming increasingly common. If LVF can be determined accurately from ECG measurements, this could reduce the need for echocardiogram acquisition, which is time-consuming, costly, and can only be performed in a hospital. Moreover, explainable automation of ECG interpretation may allow less specialized healthcare personnel to reliably assess LVF. Accurate LVF prediction from 12-lead ECGs also opens the path towards extending these capabilities to 1-lead ECGs, making home-based monitoring a realistic opportunity for scalable and continuous assessment of cardiac function.

Although our study demonstrates promising results, it is important to acknowledge its limitations and identify directions for future research.

One key limitation lies in the adversarial network's inability to effectively prevent label-specific information from being encoded in z_{nls}. The dimensionality of the latent space (L_{nls}) may facilitate easier adversarial label prediction, undermining the network's ability to suppress label-specific information. In addition, the interaction between L_{nls} and the hyperparameters γ_{ls} and γ_{nls} remains unexplored in this study, but could play a role in mitigating the capture of label-specific information in z_{nls}. Considering the challenges associated with training adversarial networks, future work should investigate the use of extended training schedules to enhance their effectiveness [42].

The joint optimization of the prediction and reconstruction tasks introduced additional challenges, necessitating careful tuning of the hyperparameter γ_{ls} to mitigate overfitting. These challenges stem from the differing complexities and data scales associated with the two tasks. To address this, the reconstruction task was independently trained for several epochs before initiating joint optimization. While this approach was effective, alternative strategies to achieve comparable results were not explored in this study.

Another notable limitation of this study is the imperfect prediction of LVF and the observed overlap in predicted LVF values across true LVF classes, as depicted in Figure 3. This overlap can be partially attributed to the labeling process, where ground truth LVF values are rounded to the nearest perceived class, introducing variability near class boundaries. Substituting LVF with left ventricular ejection fraction (LVEF) as the target variable could potentially enhance predictive performance. Unfortunately, concurrent LVEF measurements were unavailable for most of the ECGs analyzed in this study. Previous research has demonstrated that the 12-lead ECG may not encompass all the information required for perfect LVF prediction, which can also explain the overlap [4], [10]. Nonetheless, accurate LVF prediction is critical to ensure the validity of attribute manipulation as illustrated in Figure 4. Incorporating additional predictors that provide more detailed information about LVF status could improve overall predictive accuracy and enhance the robustness of the explainability framework.

In clinical practice, ECG signals are typically preprocessed through methods such as baseline wander correction and denoising to enhance signal quality and standardization [43]. These steps were not applied in our study, and their inclusion may further improve model performance and clinical applicability. In contrast, some preprocessing procedures used here, such as heartbeat segmentation and exclusion of irregular beats, may restrict applicability in contexts like atrial fibrillation, ectopic beats or certain tachycardias, where the information is in the irregular beat itself [44]. Moreover, pathologies where R-peak detection is complicated such as bundle branch blocks or cardiomyopathies may hinder appropriate heartbeat segmentation, which also restricts the applicability of the current pipeline.

The generalizability of our findings to broader patient populations, different clinical settings, and for predicting cardiac parameters beyond LVF remains to be established. Importantly, while this work primarily serves as a proof of concept for interpretable ECG classification, the underlying methodology is not limited to ECG. The approach is applicable to any encoder–decoder architecture, future research could extend it to other cardiac populations or even to entirely different medical domains.

V. Conclusion

In conclusion, our study represents a significant step forward in the field of ECG analysis by enhancing the specificity and interpretability of latent features derived from ECG signals. By improving the classification rationale specific to each case, our method has the potential to enhance clinical decision-making and patient care in cardiac disease monitoring.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 860173.

Conflict of Interest Statement

The authors have no relevant conflicts of interest to disclose.

ACKNOWLEDGMENT

REFERENCES

[1] G. D. Clifford *et al.*, "AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017," in *Computing in Cardiology (CinC)*, 2017, pp. 1–4.

[2] E. A. P. Alday *et al.*, "Classification of 12-lead ECGs: the PhysioNet/computing in cardiology challenge 2020," *Physiological measurement*, vol. 41, no. 12, p. 124003, 2020.

[3] R. R. van de Leur *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.

[4] T. Van Steenkiste, D. Deschrijver, and T. Dhaene, "Interpretable ECG beat embedding using disentangled variational auto-encoders," in *IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, 2019, pp. 373–378.

[5] J. H. Jang *et al.*, "Unsupervised feature learning for electrocardiogram data using the convolutional variational autoencoder," *PLoS ONE*, vol. 16, no. 12, pp. 1–16, December 2021. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0260612

[6] V. V. Kuznetsov, V. A. Moskalenko, and N. Y. Zolotykh, "Electrocardiogram generation and feature extraction using a variational autoencoder," *arXiv*, pp. 1–6, 2020. [Online]. Available: http://arxiv.org/abs/2002.00254

[7] I. Higgins *et al.*, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=Sy2fzU9gl

[8] J. Parker and R. Case, "Normal left ventricular function." *Circulation*, vol. 60, no. 1, pp. 4–12, 1979.

[9] R. R. van de Leur *et al.*, "Inherently explainable deep neural network-based interpretation of electrocardiograms using variational auto-encoders," *medRxiv*, 2022.

[10] V. van der Valk *et al.*, "Joint optimization of a beta-vae for ecg task-specific feature extraction," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, H. Greenspan *et al.*, Eds. Cham: Springer Nature Switzerland, 2023, pp. 554–563.

[11] T. Karras, S. Laine, and T. Aila, " A Style-Based Generator Architecture for Generative Adversarial Networks ," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 43, no. 12, pp. 4217–4228, Dec. 2021. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/TPAMI.2020.2970919

[12] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *CoRR*, vol. abs/1312.6114, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:216078090

[13] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *Ieee Access*, vol. 8, pp. 193 907–193 934, 2020.

[14] P. Chormai *et al.*, "Disentangled explanations of neural network predictions by finding relevant subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[15] Y. Skandarani *et al.*, "Generative adversarial networks in cardiology," *Canadian Journal of Cardiology*, vol. 38, no. 2, pp. 196–203, 2022.

[16] C. T. Wei *et al.*, "Contrastive heartbeats: Contrastive learning for self-supervised ecg representation and phenotyping," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1126–1130.

[17] O. Atamny *et al.*, "Outlier detection in ecg," in *2023 Computing in Cardiology (CinC)*, vol. 50. IEEE, 2023, pp. 1–4.

[18] Y. Cho *et al.*, "Artificial intelligence algorithm for detecting myocardial infarction using six-lead electrocardiography," *Scientific reports*, vol. 10, no. 1, p. 20495, 2020.

[19] Y. M. Ayano *et al.*, "Interpretable machine learning techniques in ecg-based heart disease classification: a systematic review," *Diagnostics*, vol. 13, no. 1, p. 111, 2022.

[20] P. Wagner *et al.*, "Explaining deep learning for ecg analysis: Building blocks for auditing and knowledge discovery," *Computers in Biology and Medicine*, vol. 176, p. 108525, 2024.

[21] H. Liu *et al.*, "Using the vq-vae to improve the recognition of abnormalities in short-duration 12-lead electrocardiogram records," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105639, 2020.

[22] A. Pagnoni, K. Liu, and S. Li, "Conditional variational autoencoder for neural machine translation," *arXiv preprint arXiv:1812.04405*, 2018.

[23] Q. Zhu and R. Zhang, "A classification supervised auto-encoder based on predefined evenly-distributed class centroids," *arXiv preprint arXiv:1902.00220*, 2019.

[24] A. Pati and A. Lerch, "Attribute-based regularization of latent spaces for variational auto-encoders," *Neural Computing and Applications*, vol. 33, pp. 4429–4444, 2021.

[25] J. Klys, J. Snell, and R. Zemel, "Learning latent subspaces in variational autoencoders," *Advances in neural information processing systems*, vol. 31, 2018.

[26] Y. Xia, W. Wang, and K. Wang, "Ecg signal generation based on conditional generative models," *Biomedical Signal Processing and Control*, vol. 82, p. 104587, 2023.

[27] E. Adib, F. Afghah, and J. J. Prevost, "Synthetic ecg signal generation using generative neural networks," *arXiv preprint arXiv:2112.03268*, 2021.

[28] Y. Sang, M. Beetz, and V. Grau, "Generation of 12-lead electrocardiogram with subject-specific, image-derived characteristics using a conditional variational autoencoder," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[29] M. Beetz, A. Banerjee, and V. Grau, "Multi-domain variational autoencoders for combined modeling of mri-based biventricular anatomy and ecg-based cardiac electrophysiology," *Frontiers in physiology*, vol. 13, p. 886723, 2022.

[30] P. K. Gyawali *et al.*, "Learning disentangled representation from 12-lead electrograms: application in localizing the origin of ventricular tachycardia," *arXiv preprint arXiv:1808.01524*, 2018.

[31] S. Vijayarangan *et al.*, "RPnet: A deep learning approach for robust R peak detection in noisy ECG," in *International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 345–348.

[32] A. Galdran, "Performance metrics for probabilistic ordinal classifiers," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 357–366.

[33] 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.

[34] K. He *et al.*, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[35] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[36] NVIDIA, P. Vingelmann, and F. H. Fitzek, "Cuda, release: 10.2.89," 2020. [Online]. Available: https://developer.nvidia.com/cuda-toolkit

[37] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, pp. 837–845, 1988.

[38] L. Gaudette and N. Japkowicz, "Evaluation methods for ordinal classification," in *Advances in Artificial Intelligence: 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009 Kelowna, Canada, May 25-27, 2009 Proceedings 22*. Springer, 2009, pp. 207–210.

[39] C. C. Tan, T. M. Hiew, and B. Chia, "Right chest electrocardiographic patterns in normal subjects," *Chest*, vol. 97, no. 3, pp. 572–575, 1990.

[40] E. Baiocco *et al.*, "Early repolarization: When is it a normal pattern?" *New Concepts in ECG Interpretation*, pp. 111–117, 2019.

[41] P. Ahmadi *et al.*, "Age and gender differences of basic electrocardiographic values and abnormalities in the general adult population; tehran cohort study," *BMC Cardiovascular Disorders*, vol. 23, no. 1, p. 303, 2023.

[42] M. M. Saad, R. O'Reilly, and M. H. Rehmani, "A survey on training challenges in generative adversarial networks for biomedical image analysis," *Artificial Intelligence Review*, vol. 57, no. 2, p. 19, 2024.

[43] M. Blanco-Velasco, B. Weng, and K. E. Barner, "Ecg signal denoising and baseline wander correction based on the empirical mode decomposition," *Computers in biology and medicine*, vol. 38, no. 1, pp. 1–13, 2008.

[44] K. I. Shine, J. A. Kastor, and P. M. Yurchak, "Multifocal atrial tachycardia: clinical and electrocardiographic features in 32 patients," *New England Journal of Medicine*, vol. 279, no. 7, pp. 344–349, 1968.