Joint Optimization of a β-VAE for ECG Task-Specific Feature Extraction

Viktor van der Valk[1(✉)], Douwe Atsma[3], Roderick Scherptong[3], and Marius Staring[2]

[1] TECObiosciences GmbH, Landshut, Germany
viktorvandervalk@gmail.com
[2] Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands
[3] Department of Cardiology, Leiden University Medical Center, Leiden, The Netherlands

Abstract. Electrocardiography is the most common method to investigate the condition of the heart through the observation of cardiac rhythm and electrical activity, for both diagnosis and monitoring purposes. Analysis of electrocardiograms (ECGs) is commonly performed through the investigation of specific patterns, which are visually recognizable by trained physicians and are known to reflect cardiac (dis)function. In this work we study the use of β-variational autoencoders (VAEs) as an explainable feature extractor, and improve on its predictive capacities by jointly optimizing signal reconstruction and cardiac function prediction. The extracted features are then used for cardiac function prediction using logistic regression. The method is trained and tested on data from 7255 patients, who were treated for acute coronary syndrome at the Leiden University Medical Center between 2010 and 2021. The results show that our method significantly improved prediction and explainability compared to a vanilla β-VAE, while still yielding similar reconstruction performance.

Keywords: Explainable AI · ECG · β-VAE · feature extraction · LVF prediction

1 Introduction

The electrocardiogram (ECG), is one of the most widely used methods to analyze cardiac morphology and function, by measuring the electrical signal from the heart with multiple electrodes. ECG data is used by clinicians for both diagnostic and monitoring purposes in various cardiac syndromes. A 12-lead ECG is routinely obtained in patients to diagnose and monitor disease development. However, for the interpretation of the ECG signal, the knowledge of an expert is required. Physicians usually analyze the ECG through the recognition of specific patterns, known to be associated with disease. This however requires substantial expertise, and potentially additional relevant information exists in a 12-lead ECG

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 554–563, 2023.
https://doi.org/10.1007/978-3-031-43895-0_52

missed by human interpretation. Deep learning has already proven its usefulness in the interpretation of the ECG signal in multiple classification challenges [1,3] and more recently also in feature discovery by means of explainable AI algorithms [2,7,9,10,15]. The explainablity of AI algorithms is especially valued in medical settings, where trusting a black box AI algorithm is undesirable [1].

VAEs and in particular β-VAEs have been used as unsupervised explainable ECG feature generators in the explainable AI algorithms mentioned above [6]. It was shown that a β-VAE, trained on reconstruction of the ECG signal, is able to extract features from the ECG signal that can be made more interpretable by visualization of reconstructed latent space samples with the decoder of the β-VAE [10]. This is a first step towards an explainable deep learning pipeline for ECG analysis. However, the features generated by a β-VAE when only trained to minimized reconstruction loss, are likely not optimal for task specific predictions.

The aim of this paper is to explore further improvement of the latent features by improving their explainability and prediction performance. This is clinically relevant but unexplored for the post myocardial infarction setting. We propose to improve explainability by reducing the dimension of the latent space to a level more manageable for human assessment, while encouraging outcome specific information to be captured in a small part of the latent space, and while maintaining ECG reconstruction performance for visual assessment. To achieve this, we propose a novel method to jointly optimize the β-VAE with a combination of a task specific prediction loss for a subset of the latent space, and KL-divergence and reconstruction loss for the entire latent space. The task chosen to optimize here is left ventricular function (LVF), one of the most important determinants of prognosis in patients with cardiac disease. Current assessment of LVF requires advanced imaging methods and interpretation by a trained professional. The ECG, on the other hand, can be obtained by a patient at home. In combination with automated analysis this would facilitate remote monitoring of LVF in patients.

2 Methods

2.1 Data

To train the models for both reconstruction and LVF prediction, two datasets were used: i) A non-labeled dataset consisting of 119,886 raw 10 s 12-lead ECG signals taken at 500 Hz from 7255 patients diagnosed with acute coronary syndrome between 2010 and 2021 at the Leiden University Medical Center, the Netherlands; ii) A labeled dataset of 33,610 ECGs from 2736 patients of the same cohort. This dataset was labeled by visual assessment of an echocardiogram performed within 3 days before or after the ECG. The label categories, normal, mild, moderate and severe impairment were binarized for model training. When the ECG was taken within two weeks after cardiac intervention a 1-day margin was used. If a cardiac intervention was performed between ECG and echocardiography, the case was excluded. 11.5% of the ECGs were labeled with a moderate to severe impaired LVF. The institutional review board approved

the study protocol (nWMODIV2_2022006) and waived the obligation to obtain informed consent.

2.2 Data Preprocessing

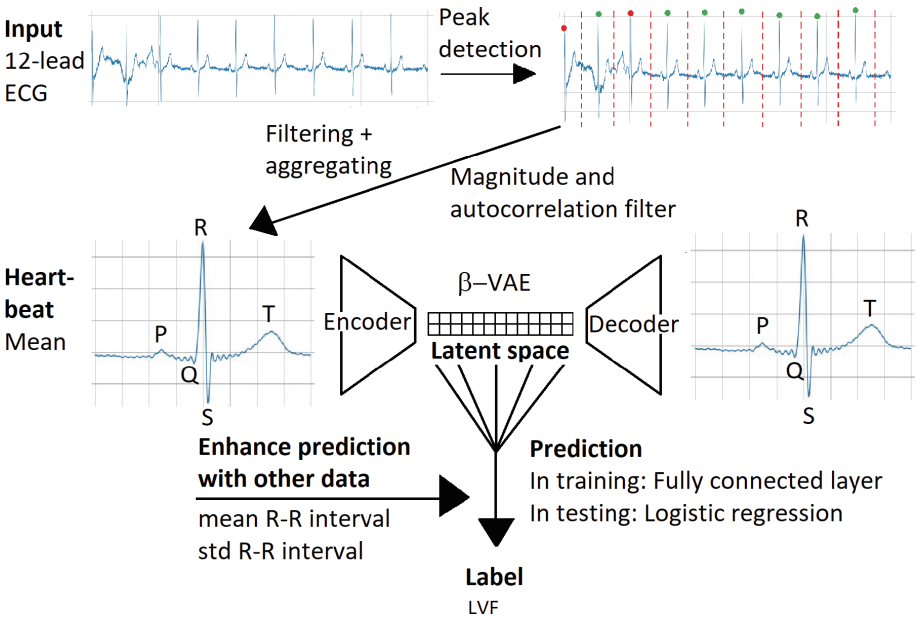


Fig. 1. Preprocessing, feature extraction and prediction pipeline.

The raw ECG signals were first split into separate heartbeats (400ms before and after the R-peak, the largest peak in the ECG, that represents depolarization of the ventricles) with a peak detection method inspired by RPNet, a U-Net structured CNN with inception blocks, that was trained on manually labeled peak locations [16]. The heartbeats were then filtered with a magnitude and an autocorrelation filter. The magnitude filter removed heartbeats with an average magnitude below a set threshold. The autocorrelation filter removed signals where both the mean and maximum autocorrelation between the heartbeats were below a set threshold. These two criteria were used, since ECG signals showing multiple rhythms can result in low mean autocorrelation, but, if not noisy, will not result in low maximum autocorrelation. The remaining heartbeats were then averaged per ECG lead. The μ and σ of the intervals of the subsequent R peaks were used as an additional feature for LVF prediction.

2.3 Model Overview

To investigate a general improvement to the VAE feature extraction pipeline [7,9,10,15], the proposed method was tested with two architectures: i) A small

VAE with 300k parameters consisting of an encoder and a mirrored decoder. Both parts contained 7 2D convolutional layers, of which 3 were residual layers, with respective channel sizes of [8,16,32,64,64,64,64] and a kernel size of 5; ii) A second larger VAE from the FactorECG pipeline as proposed by Van de Leur *et al.* (2022) [10] with 50M parameters. The VAEs were both extended at the bottleneck (the latent space, of size L), with a single fully connected layer for output prediction, in this case the LVF label, see Fig. 1. The μ and σ of the RR intervals (time between two subsequent R peaks), were added to the input of the prediction layer, since the information represented by these features is lost in averaging the heartbeats. To maintain explainability of the extracted features, only one fully connected layer is used, as otherwise the features will become weighted combinations of the latent space values, which makes visualization with the decoder and subsequent interpretation complex. However, for pure prediction performance, additional fully connected layers may have been helpful. The extracted features, again with the μ and σ of the RR interval, were subsequently analyzed with logistic regression using regularized l1 and l2 penalties on the LVF prediction task, ignoring the output of the prediction layer in the VAE. The VAEs were build and trained in the PyTorch 1.12 framework and trained on a Quadro RTX 6000 GPU with CUDA 11.4 [12,13], while for logistic regression we used the Scikit-learn toolbox [14]. The implementation of our models will be made publicly available via GitHub at https://github.com/ViktorvdValk/Task-Specific-VAE.

2.4 Model Training

The β-VAE was first pretrained in a self-supervised manner with the mean heartbeats of all filtered ECG signals, minimizing i) the mean squared reconstruction error (MSE) between the input and output ECG, and ii) the KL-divergence between the output of the encoder and the standard normal distribution. The KL-divergence loss was weighted with a β factor, like in the original paper [6]. This pretrained VAE was then fine-tuned in two-steps, first the encoder and decoder were fixed, and only the prediction layer was trained, then all layers were trained end-to-end. This training scheme was used to ensure more stable training. For these fine-tuning steps, the loss function was complemented with a binary cross-entropy loss, which was weighted with a γ factor. The *task naive* VAE resulting from pretraining was compared to the *task specific* VAE resulting from both fine-tuning steps. For pretraining, both datasets were combined and split in a training (85%) and a test set (15%). 5-fold cross validation was done with the training set with again an 85%:15% ratio between training and validation set. For fine-tuning, the same procedure was used on just the labeled dataset, making sure labeled ECGs were in the same set in both cases. All data splits were grouped by patient and stratified by label in case of the labeled data splits. Both pretraining and fine-tuning were done until convergence, i.e. until the loss on the validation set stopped improving for 25 epochs. This was done to prevent the advantage of additional training of the *task specific* network. To prevent overfitting, balanced sampling and regularization by means of drop out

layers and the Adam optimizer with weight decay were used, this was especially necessary in the fine-tuning phase. To prevent gradient explosion, gradient clipping and He initialization were used [5].

2.5 Feature Evaluation

The differences between the features from the *task naive* and *task specific* VAEs, were compared w.r.t. reconstruction and prediction. For reconstruction, both MSE and correlation between input and output ECG, and for prediction the Area Under the Receiver Operator Characteristic Curve (AUROC) and the macro-averaged F1 score were used. Significant difference between AUROC scores was calculated as proposed in Hanley & McNeil (1983) [4]. For visualization of the representation of a latent space feature f in a so called factor traversal, all features except f were sampled at their mean, while f was sampled in a range between $\mu - 3\sigma$ and $\mu + 3\sigma$. Using these samples as input for the decoder, creates a representation of that feature, which can give insight in ECG features that are important for LVF prediction.

2.6 Baseline Methods

As a baseline method, a principal component analysis (PCA) was performed on the preprocessed ECGs, to extract features. PCA can be considered an ordered task naive linear feature extractor that focuses on the axis of the largest variance, in contrast to the VAEs which are non-ordered non-linear feature extractors, that are optimized for reconstruction. A logistic regression predictor with just sex and age as input was used as an additional baseline.

3 Experiments and Results

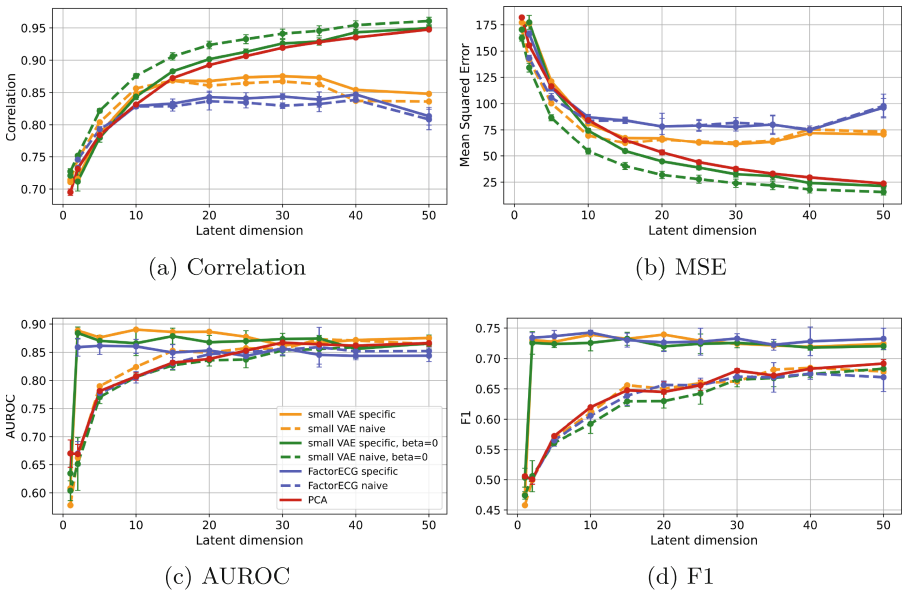
3.1 Experiments

The proposed pipeline contains several hyper-parameters, of which the latent space size L was optimized in this study. The influence of the β parameter was also briefly addressed. L was optimized for its importance in the explainability and the reconstruction and prediction quality of the model. A higher L increases the complexity of the model, and consequently decreases its explainability. An L that is too low, on the other hand, restricts the capacity of the model for reconstruction and prediction. The PCA baseline method was considered to give an upper bound of L, since the number of principal components, the PCA analog for L, indicates how many values would be needed to capture sufficient information.

3.2 Hyperparameter Optimization

The influences of γ on prediction and reconstruction performance was small and was therefore fixed to 500. The influence of L on prediction quality can

be seen in Fig. 2. The PCA baseline performs more or less equal to the *task naive* networks for all L. For the *task specific* networks, the F1 scores are higher than their *task naive* counterparts and the PCA baseline, especially for lower L. The *task specific* VAEs already reached their best prediction performance starting at $L = 2$, as compared to the *task naive* VAEs and the PCA baseline that reach their best prediction performance from $L = 30$. The influence of L on reconstruction can be seen in Fig. 2a and b. All networks perform equal to the PCA baseline for low latent dimensions. The reconstruction for the small VAE and the FactorECG VAE does not seems to improve any further for respectively $L >$20 and $L > 15$, where the PCA baseline reconstruction keeps improving with L. However, setting β to 0 and thereby ablating the variational nature of the VAEs prevents this stagnation of reconstruction performance. The *task specific* networks perform equally well as their *task naive* counterparts, which suggests that the additional joint optimization does not have a major negative impact on reconstruction. The optimization shows that the relevant information for LVF prediction in the ECG signal can be captured in just two features by both VAEs. Reconstruction, on the other hand, requires at least 10/15 features for the VAEs to reach maximum performance. Therefore, in another experiment, a *split task* VAE was trained, in which 8 of the latent space features where only optimized for reconstruction and only 2 also for prediction.



(a) Correlation (b) MSE

(c) AUROC (d) F1

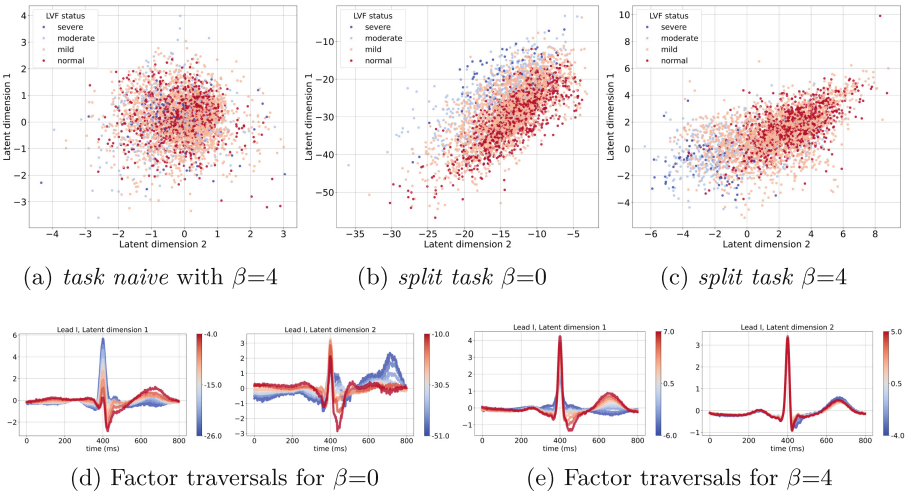
Fig. 2. Influence of the latent dimension L on reconstruction quality: (a) correlation and (b) MSE, and on prediction quality: (c) AUROC and (d) F1-score, for various models. Plotted are the mean and standard deviation of 5-fold cross-validation on the validation set.

Table 1. Reconstruction and LVF prediction comparison on the test set for the *task naive* and *task specific* architectures. The results show the μ of 5-fold cross-validation. AUROC is shown with its 95% confidence interval.* p-value < 0.01 between AUROC of task naive and specific method for all folds.

Architecture	L	MSE	Correlation	AUROC	F1
Sex and age	2	–	–	0.556 (0.520–592)	0.474
PCA	2	147	0.724	0.656 (0,624–0.688)	0.496
Small VAE task naive	2	133	0.739	0.686 (0.655–0.716)	0.503
Small VAE task specific	2	164	0.711	0.842* (0.822–0.861)	0.682
Small VAE split task	2 (10)	76.5	0.838	0.839 (0.819–0.859)	0.695
Small VAE split task $\beta = 0$	2 (10)	73.6	0.838	0.846 (0.823–0.862)	0.674
FactorECG [10] task naive	2	139	0.735	0.685 (0.654–0.715)	0.507
FactorECG [10] task specific	2	161	0.724	0.770* (0.745–0.796)	0.695
PCA	10	77.2	0.826	0.761 (0.735–0.787)	0.580
Small VAE task naive	10	63.1	0.854	0.803 (0.781–0.826)	0.586
Small VAE task specific	10	70.6	0.847	0.853* (0.834–0.871)	0.679
FactorECG [10] task naive	10	84.6	0.820	0.770 (0.745–0.796)	0.579
FactorECG [10] task specific	10	87.2	0.822	0.833* (0.813–0.854)	0.707

3.3 Results on the Test Set

Table 1 shows the results on the test set for $L = 2$ and $L = 10$. The *(split) task specific* networks significantly outperform their *task naive* counterparts, the PCA baseline, and the sex and age benchmark w.r.t. LVF prediction. Figure 3a, 3b and 3c show that the *split task*, in contrast to the *task naive* small VAE with $\beta = 0$ and $\beta = 4$ can be used to encode the ECG signals to a landscape that visually separates the signals based on LVF status reasonably well. The factor traversals in Fig. 3e and d show an example of the interpretation of the latent features. Setting β to 0, creates features that appear visually less informative.



(a) *task naive* with β=4 (b) *split task* β=0 (c) *split task* β=4

(d) Factor traversals for β=0 (e) Factor traversals for β=4

Fig. 3. Comparison of the latent space for different values of β, for the small VAE. For *task specific* methods, the scatter plots show the two dimensions of the latent space that are optimized for prediction: (a) *task naive* ($\beta = 4$); (b) *split task* ($\beta = 0$); (c) *split task* ($\beta = 4$). The latent space factor traversals (d) and (e) show the visual representation of the features for Lead I of the 12-lead ECG signal: (d) $\beta = 0$; (e) $\beta = 4$.

4 Discussion

Joint optimization of a β-VAE successfully generated features that contain more information about LVF, without hampering reconstruction of the ECG signal. We hypothesize that the β-VAEs have multiple optima for ECG reconstruction of which only some generate features that are relevant for LVF prediction. This study shows that joint optimization will favor this desired subset of optima, and that this is true for different architectures. In addition, we showed that jointly optimizing only a subset of the latent space features for prediction, results in aggregation of the predictive information, thereby improving explainability.

The AUROC score of the FactorECG VAE prediction is similar when compared to van der Leur *et al.* (2022) [10] (AUROC≈0.9 for $L = 36$). However, the proposed small VAE achieved equal if not better reconstruction and prediction performance with less than 1% of the parameters as shown in Fig. 2.

The F1 score is considered more robust than the AUROC score with data imbalance, which is the case here [8]. From Fig. 2d we can therefore conclude that the *task specific* networks outperform the *task naive* networks for any L. The differences between the *task specific* networks and their *task naive* versions in prediction, at similar reconstruction, indicate that the ECG signal can be summarized with a set of latent features of which only a subset is important for LVF prediction. The joint optimization promotes the extraction of this subset especially when L is small. Figure 2a and b show that the PCA baseline outperforms both VAEs in reconstruction for $L > 20$ when $\beta = 4$, but not for β

= 0. This indicates that the VAEs are restricted in reconstruction by the KL-divergence loss. This loss was shown to promote feature disentanglement and a gradient in the latent space [11]. Figure 3d and e show that without this loss ($\beta = 0$) the latent features are more complex to interpret. This could be explained as a reduction of the disentanglement of the features resulting from the absence of the KL-divergence loss. However, Fig. 3b and c both show a gradient in the latent space, which suggests that the prediction loss on its own also promotes a gradient in the latent space. Moreover, Fig. 3c shows dependence, and thus a lack of disentanglement, between the latent features even when $\beta = 4$. This complex interplay between the three losses used in the joint optimization, is very relevant for the explainability aspect of this method, but beyond the scope of the current study. We aim to examine the complex interplay in future work. In conclusion, the proposed joint optimization improves both explainability and prediction performance of VAEs by extraction of a smaller set of LVF specific features from the ECG signal. This could reduce the need of more advanced imaging methods, currently needed to measure the LVF. This opens the way for remote monitoring of left ventricular function in patients.

Acknowledgment. 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.

References

1. Alday, E.A.P., et al.: Classification of 12-lead ECGs: the PhysioNet/computing in cardiology challenge 2020. Physiol. Meas. **41**(12), 124003 (2020)
2. Basu, S., Wagstyl, K., Zandifar, A., Collins, L., Romero, A., Precup, D.: Early prediction of Alzheimer's disease progression using variational autoencoders. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 205–213. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_23
3. Clifford, G.D., et al.: AF classification from a short single lead ECG recording: the PhysioNet/computing in cardiology challenge 2017. In: Computing in Cardiology (CinC), pp. 1–4 (2017)
4. Hanley, J.A., McNeil, B.J.: A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology **148**(3), 839–843 (1983)
5. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
6. Higgins, I., et al.: beta-VAE: learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations (2017). http://openreview.net/forum?id=Sy2fzU9gl
7. Jang, J.H., Kim, T.Y., Lim, H.S., Yoon, D.: Unsupervised feature learning for electrocardiogram data using the convolutional variational autoencoder. PLoS ONE **16**(12), 1–16 (2021)
8. Jeni, L.A., Cohn, J.F., De La Torre, F.: Facing imbalanced data-recommendations for the use of performance metrics. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 245–251. IEEE (2013)

9. Kuznetsov, V.V., Moskalenko, V.A., Zolotykh, N.Y.: Electrocardiogram generation and feature extraction using a variational autoencoder. arXiv, pp. 1–6 (2020). http://arxiv.org/abs/2002.00254
10. van de Leur, R.R., et al.: Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders. Eur. Heart J.-Digit. Health **3**(3), 390–404 (2022)
11. Mathieu, E., Rainforth, T., Siddharth, N., Teh, Y.W.: Disentangling disentanglement in variational autoencoders. In: International Conference on Machine Learning, pp. 4402–4412. PMLR (2019)
12. NVIDIA: Vingelmann, P., Fitzek, F.H.: CUDA, release: 10.2.89 (2020). http://developer.nvidia.com/cuda-toolkit
13. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems, pp. 8024–8035 (2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
14. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
15. Van Steenkiste, T., Deschrijver, D., Dhaene, T.: Interpretable ECG beat embedding using disentangled variational auto-encoders. In: IEEE International Symposium on Computer-Based Medical Systems (CBMS), pp. 373–378 (2019)
16. Vijayarangan, S., Vignesh, R., Murugesan, B., Preejith, S., Joseph, J., Sivaprakasam, M.: 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), pp. 345–348 (2020)