2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)

Interpretable ECG Beat Embedding using Disentangled Variational Auto-Encoders

Tom Van Steenkiste, Dirk Deschrijver and Tom Dhaene
Department of Information Technology, IDLab
Ghent University - imec
Ghent, Belgium
tomd.vansteenkiste@ugent.be

***Abstract*—Electrocardiogram signals are often used in medicine. An important aspect of analyzing this data is identifying and classifying the type of beat. This classification is often done through an automated algorithm. Recent advancements in neural networks and deep learning have led to high classification accuracy. However, adoption of neural network models into clinical practice is limited due to the black-box nature of the classification method. In this work, the use of variational auto encoders to learn human-interpretable encodings for the beat types is analyzed. It is demonstrated that using this method, an interpretable and explainable representation of normal and paced beats can be achieved with neural networks.**

***Keywords*-interpretable model; ECG beat classification; deep learning; disentangled variational auto encoder;**

I. Introduction

Electrocardiogram (ECG) measurements are used throughout all branches of medicine. An important step in interpreting the ECG data is detecting the beats and subsequently classifying each beat per type. As the number of beats to classify can be high, and human-based classification is a time-consuming task, automated approaches for beat classification have been investigated. Many powerful machine learning approaches have been proposed, including recent advancements in neural networks and deep learning techniques [1], [2], [3]. However, the adoption of such neural network models into clinical practice is restricted by the lack of model interpretability. This model interpretability by human experts is crucial in medicine to ensure trustworthiness of the results [4].

A straightforward approach for creating an interpretable machine learning model is building a standard rule-based classifier. Yet, with this approach, the improved modeling capabilities of neural network methods cannot be exploited.

To reduce the complexity of neural network models, and hence improve their interpretation, dimensionality reduction techniques such as auto encoder (AE) models have been proposed. These models reduce the complexity by forcing the model to use a lower-dimensional embedding of the input data. Nevertheless, there can still be complex interactions across individual dimensions of the embedding, also known as channels, leading to further difficulties in interpretation of the model decision.

To disentangle such interactions and further reduce the complexity of the embedding, disentangled variational auto encoders (β-VAE) were introduced [5]. These models are capable of learning disentangled generative embeddings by forcing the model to represent the information in as few dimensions as possible, while using a probabilistic interpretation of the embedding. During training, a generative model is created that allows analysts to measure and see the impact of the position within a specific dimension of the embedding. In doing so, the reason for a specific model decision can be traced back to an embedding that has independent and explainable parameters.

In this work, the use of such a β-VAE is investigated for creating an interpretable and explainable ECG beat embedding that can subsequently be used in a classification system. The method is used to distinguish normal from paced beats in the MIT BIH arrhythmia dataset [6]. It is demonstrated that this interpretable embedding can justify why a beat was classified as normal or paced as the model learns a characteristic set of base beats.

In section II, the β-VAE is explained. Then, in section III, the experimental setup is provided and in Section IV the results are presented and discussed. Finally, conclusions are made in Section V.

II. Variational Auto-Encoders

An AE is an unsupervised deep learning model used for creating a lower dimensional embedding, also known as latent representation, of the input data. This embedding is subsequently used in, among others, classification, detection or compression algorithms. In recent studies, AE models were used for classification [7], [8] and compression of ECG data [9].

A typical AE model consists of two parts: the encoder and decoder, as shown in Fig. 1. The model is then trained using standard deep learning algorithms and a loss function representing the reconstruction loss L_R. Auto encoders are powerful methods to compute an embedding. However, determining the size of the embedding is not straightforward and complex interactions across different dimensions can be created during training.

To get an interpretable embedding, the variational AE (VAE) model can be used [10]. It transforms regular AE models into probabilistic methods. The embedding layer of Fig. 1 is exchanged for two vectors of equal size Z_μ and Z_σ

2372-9198/19/$31.00 ©2019 IEEE
DOI 10.1109/CBMS.2019.00081

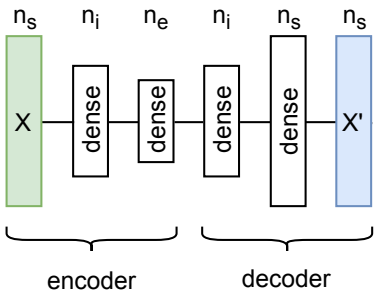


Figure 1: Auto encoder with input dimensionality n_s, intermediate layer of size n_i and embedding of size n_e. The input is represented by X and the reconstructed output is represented by X'.

followed by a sampler drawing a random sample from the distribution $\mathcal{N}(Z_\mu, Z_\sigma)$, as shown in Fig. 2. This random sample is then used by the decoder part. The decoder of a VAE is also known as a generator as this part of the model can generate new samples from the input distribution $\mathcal{X}$ given a sample from the embedding.

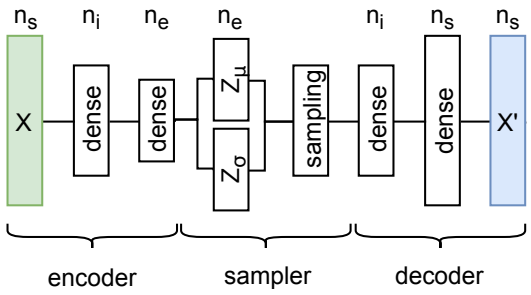


Figure 2: Variational Autoencoder with similar parameters as Fig. 1. The sampler draws a sample from $\mathcal{N}(Z_\mu, Z_\sigma)$.

During training, independence and interpretability of the embedding dimensions is encouraged by the addition of the KL-divergence D_{KL} to the loss function of the model. It is computed between $\mathcal{N}(Z_{\mu_i}, Z_{\sigma_i})$ and the standard normal $\mathcal{N}(0, 1)$ for each dimension i of the embedding. This encourages the embedding to consist of independent and standard normally distributed dimensions. The effect of this is enhanced in β-VAE by the addition of a hyperparameter β resulting in loss $= L_R + \beta D_{KL}$. This hyperparameter balances the latent embedding capacity, also known as channel capacity, with the independence and standard normal distribution constraints [5].

The resulting model is capable of automatic discovery of independent, interpretable embeddings. More details are presented in [5], [10].

III. Experimental Setup

A. Dataset

To analyze the proposed β-VAE method for creating an interpretable embedding, and compare it against normal AE, the MIT-BIH Arrhythmia dataset [6] is used. All patients with paced beats are included and an equal amount of patients with normal beats are added. To accurately test the capabilities of the models, the data is split in a separate training and test set. The patient identifiers for each set are given in Table I.

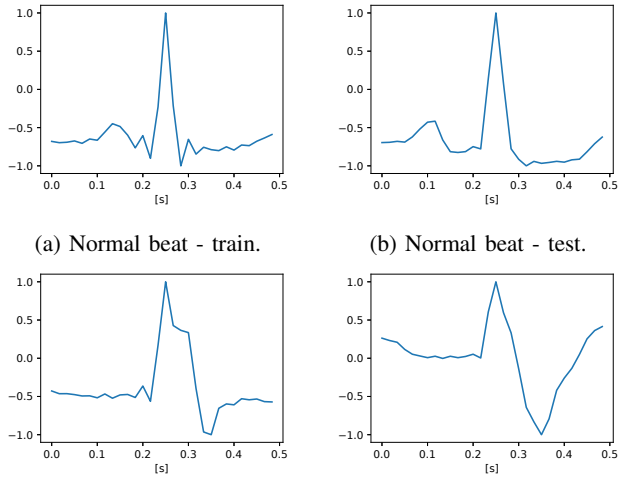
Table I: Distribution of patients across train and test set.

	normal	**paced**
train	101, 106	102, 104
test	103, 105	107, 217

The annotations included in the database are used to detect and categorize the beats. Only normal and paced beats are included in the experiment.

The ECG signal is passed through a fifth-order Butterworth bandpass filter with lower cutoff frequency of 1Hz and upper cutoff frequency of 60Hz for mild noise removal. The epochs of data have a duration of 1 second, sampled at 60Hz, with the beat centered in the epoch.

Then, the signal is normalized between $[-1, 1]$ and the center 0.5 seconds of data is extracted. This results in 30 samples per epoch (=n_s). Examples of the resulting epochs are shown in Fig. 3.



(a) Normal beat - train. (b) Normal beat - test.

(c) Paced beat - train. (d) Paced beat - test.

Figure 3: Random samples form train and test dataset used in this analysis. There is a visual and interpretable difference between the beats.

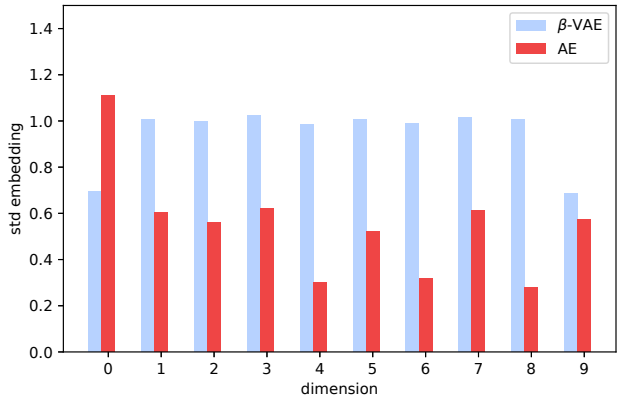
B. Modeling

Both the β-VAE as the AE model are constructed with an embedding size n_e of 10 nodes. The intermediate layer

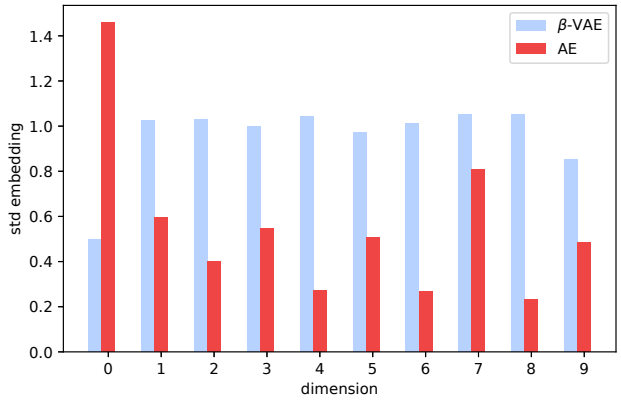
has size n_i of 20 nodes. All dense layers have linear activation functions. Each model is trained for 50 epochs on batches containing 128 samples and is optimized using the AdaDelta [11] optimizer. The root-mean-squared-error is used to represent the reconstruction loss L_R.

IV. Results and Discussion

The β-VAE model aims to summarize each dimension of the embedding with an independent standard normal distribution. This is shown in Fig. 4 where the standard deviation across the embedding for the train and test set is shown. Only two dimensions of the embedding are not close to the standard normal. In β-VAE models, these dimensions encode the information while the others have an insignificant contribution to the output of the decoder. Fig. 4 also shows the standard deviation of the embedding for the AE model. There is no pattern and the contributions of the separate input dimensions cannot be extracted or interpreted from the standard deviation for traditional AE models.



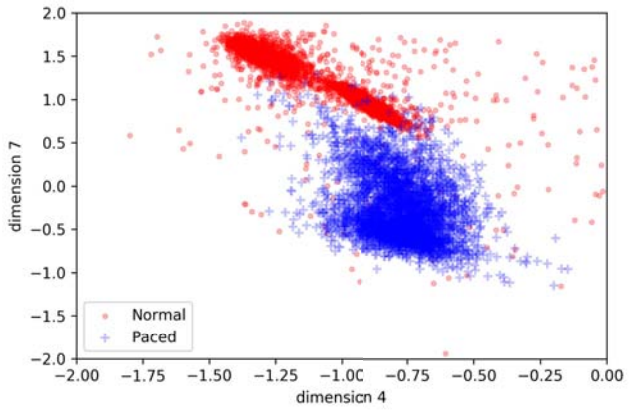
(a) Training set.



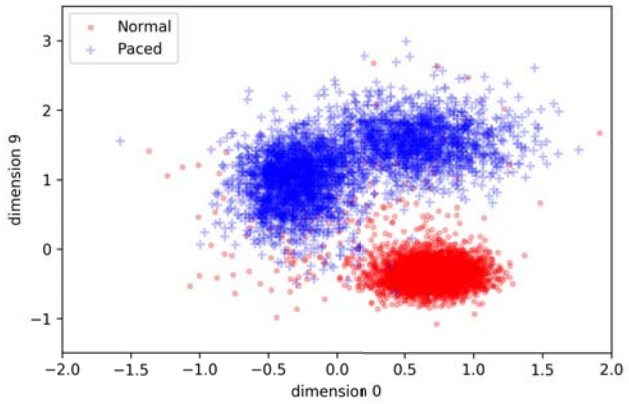
(b) Test set.

Figure 4: Standard deviation of the embedding, evaluated on the train and test dataset. The β-VAE model aims for an independent standard normal distribution in each dimension.

The position of beat types within the embedding can be visualized as shown in Fig. 5 where two random dimensions are chosen for the AE model (Fig. 5a) and the significant dimensions are chosen for the β-VAE model (Fig. 5b).



(a) AE model.

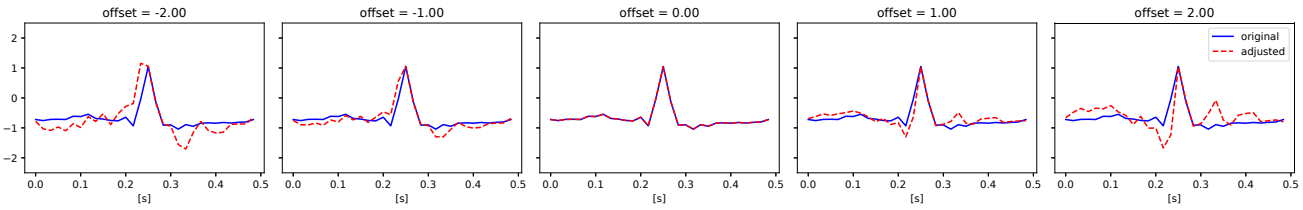


(b) β-VAE model.

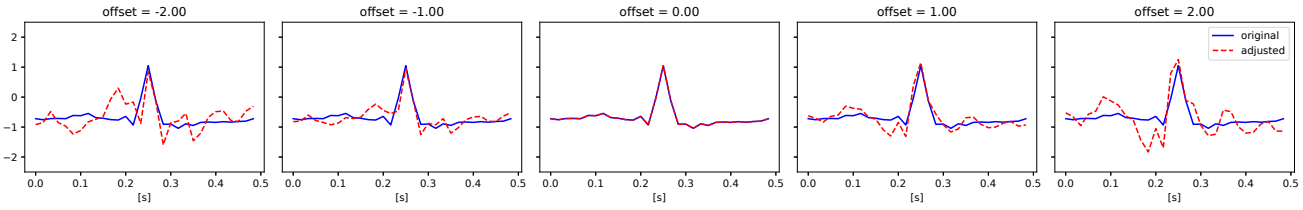
Figure 5: Learned embedding of the models using the test dataset. Both models are able to form separate clouds for normal and paced beats.

Both models are able to create separate clouds for the normal and paced beats. The difference between the models is in the interpretation of the position within the embedding. This can be analyzed by perturbing each dimension individually as shown in Fig. 6 for two random dimensions of the AE model. The resulting decoded epochs consist of many peaks and valleys and no longer contain a recognizable beat pattern. No interpretation can be linked with any dimension as the embedding is a complex combination of all 10 dimensions and only makes sense at very specific locations. Because of this, the AE model cannot be considered a generative model. Changes in the embedding do not always lead to valid samples from the original input distribution $\mathcal{X}$.

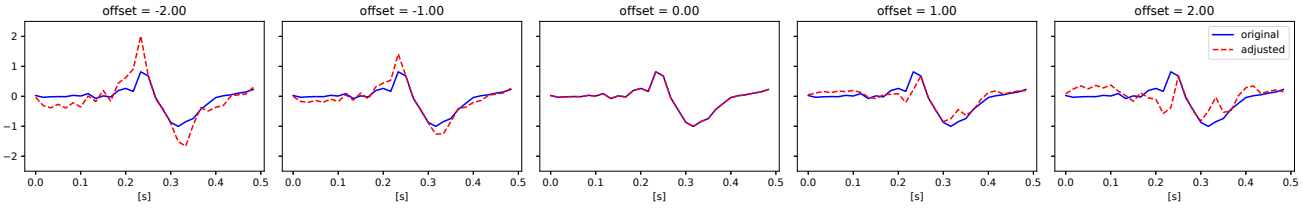
When this experiment is repeated for the two significant



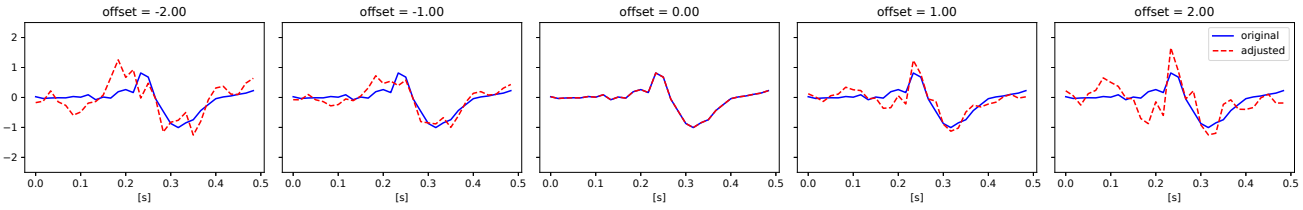
(a) Normal beat perturbed in dimension 4.



(b) Normal beat perturbed in dimension 7.



(c) Paced beat perturbed in dimension 4.



(d) Paced beat perturbed in dimension 4.

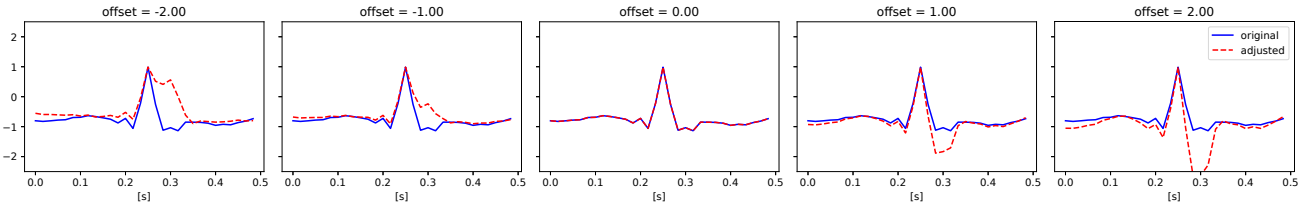
Figure 6: Perturbing the embedding of an AE model. Perturbing only one dimension generates unrecognizable patterns. Hence, the position within the embedding is not interpretable by human experts.

dimensions of the β-VAE model, the results are different, as shown in Fig 7. Now, perturbing one of the dimensions leads to a smooth transition with identifiable beat patterns. When comparing the resulting decoded version with the random samples of the database in Fig. 3, it is clear that distinct beat shapes are being learned as base for the embedding. The two dimensions now encode a physical shape of the beat and can be changed independently. Any beat can be represented as a combination of these base beats in a learned beat space.

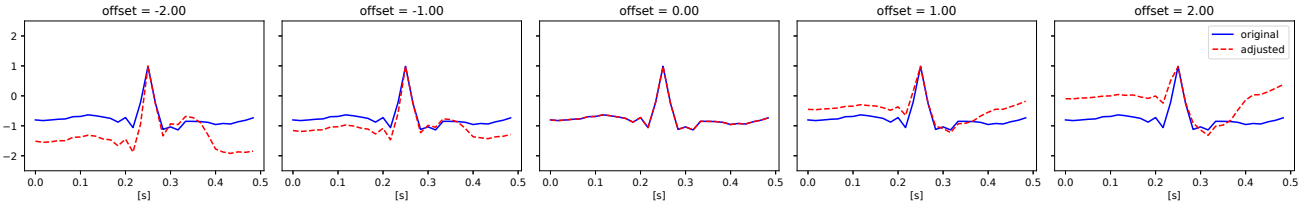
Another useful feature of the β-VAE is the ability to automatically learn the amount of required dimensions for the embedding. When the embedding is perturbed in one of the dimensions close to the independent standard normal, the decoded output does not change. This is shown in Fig. 8 where a random insignificant dimension of the embedding is perturbed. From this, it is clear that the decoder model does not take into account these dimensions and that the embedding does not need the full ten embedding nodes n_e to model the problem as two are sufficient.

With these two dimensions, the entire beat space can be visualized as shown in Fig. 9 where the decoder is evaluated with embeddings at the four corners of the embedding space. Each beat can be represented as a smooth transition within this beat space and the position within this beat space indicates the prominence of specific beat features.

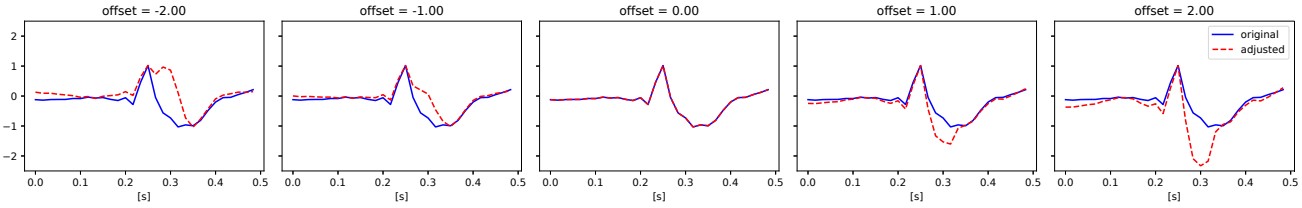
Even with this simple neural network model, the proposed method is able to learn and extract an independent, interpretable and explainable embedding. The embedding uses the minimal amount of dimensions required to model the beats and is made up of several characteristic beat types. In future work, the proposed methodology can be combined with a classification layer as demonstrated in literature [7],



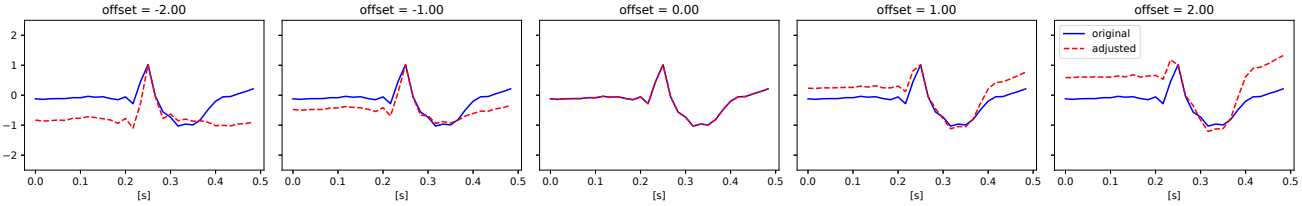
(a) Normal beat perturbed in dimension 0.



(b) Normal beat perturbed in dimension 9.



(c) Paced beat perturbed in dimension 0.



(d) Paced beat perturbed in dimension 9.

Figure 7: Perturbing the embedding of normal and paced beats using a β-VAE. An interpretable evolution of the beat can be seen. Human experts can interpret and analyze the physical meaning of the position withing the embedding.

[8] to create an interpretable and explainable beat classifier. In addition, the model can be extended with state-of-the-art deep learning architectures and trained with multiple beat types.

V. Conclusion

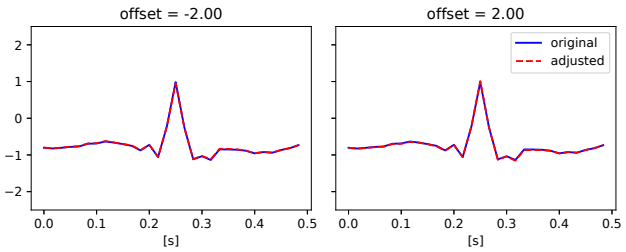
ECG beat classification is an important aspect of ECG analysis and is used in various branches of medicine. State-of-the-art neural network and deep learning models are capable of achieving a high classification accuracy. However, there is no human interpretable explanation for the classification decision of the model. By extending deep learning models to include a β-VAE embedding as illustrated in this work, representative beat patterns can be identified leading to interpretable, explainable and independent embedding dimensions. The resulting neural network model is no longer black-box and beats can be represented as combinations of learned independent base beats.

References

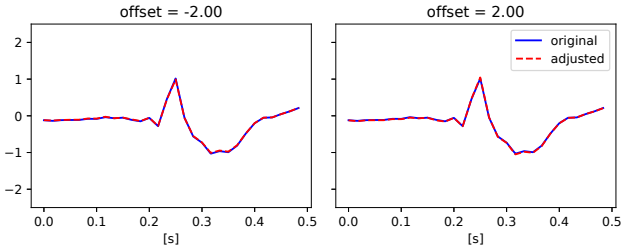
[1] M. M. Al Rahhal, Y. Bazi, H. AlHichri, N. Alajlan, F. Melgani, and R. R. Yager, "Deep learning approach for active classification of electrocardiogram signals," *Information Sciences*, vol. 345, pp. 340–354, 2016.

[2] U. R. Acharya, H. Fujita, O. S. Lih, Y. Hagiwara, J. H. Tan, and M. Adam, "Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network," *Information sciences*, vol. 405, pp. 81–90, 2017.

[3] O. Yildirim, P. Pawel, R. San Tan, and U. R. Acharya, "Arrhythmia detection using deep convolutional neural network with long duration ECG signals," *Computers in biology and medicine*, vol. 102, pp. 411–420, 2018.

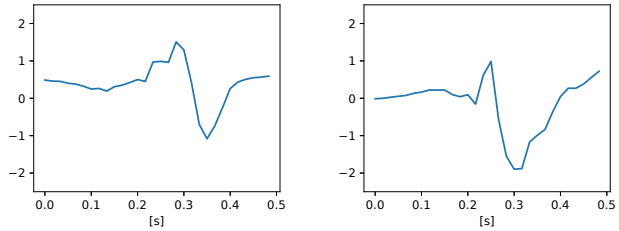


(a) Normal beat perturbed in dimension 4.



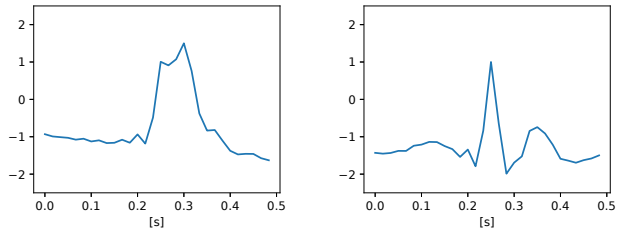
(b) Paced beat perturbed in dimension 4

Figure 8: Perturbing the embedding of a normal and a paced beat using a β-VAE. The decoded beat does not change when the embedding is perturbed in a dimension with standard deviation close to one.



(a) Embedding = [-2,2].

(b) Embedding = [2,2]



(c) Embedding = [-2,-2]

(d) Embedding = [2,-2]

Figure 9: Decoding the edges of the embedding for the β-VAE model in the significant dimensions 0 and 9. The values of the other dimensions are fixed at zero.

[4] A. Vellido, J. D. Martín-Guerrero, and P. J. Lisboa, "Making machine learning models interpretable." in *ESANN*, vol. 12. Citeseer, 2012, pp. 163–172.

[5] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[6] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001.

[7] Y. Xia, H. Zhang, L. Xu, Z. Gao, H. Zhang, H. Liu, and S. Li, "An automatic cardiac arrhythmia classification system with wearable electrocardiogram," *IEEE Access*, vol. 6, pp. 16 529–16 538, 2018.

[8] K. Ochiai, S. Takahashi, and Y. Fukazawa, "Arrhythmia detection from 2-lead ecg using convolutional denoising autoencoders," *KDD18 Deep Learning Day*, 2018.

[9] O. Yildirim, R. San Tan, and U. R. Acharya, "An efficient compression of ecg signals using deep convolutional autoencoders," *Cognitive Systems Research*, vol. 52, pp. 198–211, 2018.

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

[11] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.