Proceedings of Machine Learning Research 136:194–212, 2020 Machine Learning for Health (ML4H) 2020

DeepHeartBeat: Latent trajectory learning of cardiac cycles using cardiac ultrasounds

Fabian Laumer* FABIAN.LAUMER@INF.ETHZ.CH
Gabriel Fringeli* FGABRIEL@STUDENT.ETHZ.CH
Alina Dubatovka ALINA.DUBATOVKA@INF.ETHZ.CH
Laura Manduchi LAURA.MANDUCHI@INF.ETHZ.CH
Joachim M. Buhmann JBUHMANN@INF.ETHZ.CH
ETH Zürich, Department of Computer Science

Editors: Emily Alsentzer⊗, Matthew B. A. McDermott⊗, Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy‡, Stephanie L. Hyland‡

Abstract

Echocardiography monitors the heart movement for noninvasive diagnosis of heart diseases. It proves to be of profound practical importance as it combines low-cost portable instrumentation and rapid image acquisition without the risks of ionizing radiation. However, echocardiograms produce high-dimensional, noisy data which frequently proved difficult to interpret. As a solution, we propose a novel autoencoder-based framework, DeepHeartBeat, to learn human interpretable representations of cardiac cycles from cardiac ultrasound data. Our model encodes high dimensional observations by a cyclic trajectory in a lower dimensional space. We show that the learned parameters describing the latent trajectory are well interpretable and we demonstrate the versatility of our model by successfully applying it to various cardiologically relevant tasks, such as ejection fraction prediction and arrhythmia detection. As a result, DeepHeartBeat promises to serve as a valuable assistant tool for automating therapy decisions and guiding clinical care.

Keywords: unsupervised representation learning, sequence modeling, interpretability

1. Introduction

The assessment of cardiac function proved to be crucial for the diagnosis and prognosis of patients suffering from ventricular dysfunction (Bellenger et al., 2000; Roger et al., 2011). Impairment of cardiac function, also known as heart failure, is a rapidly growing global health issue. Although the underlying causes vary according to sex, age, ethnicity, comorbidities and environment, the majority of cases remain preventable (Ziaeian and Fonarow, 2016) by minimally invasive diagnostics like echocardiography. However, to correctly quantify cardiac function and diagnose dysfunction, expensive and time consuming medical imaging methods are often required. In developing countries, the lack of available diagnostic modalities may delay care at a more advanced stage of illness, potentially resulting in increased hospitalizations for heart failure or treatment for more extensive cardiovascular diseases (Jeemon et al., 2014). The versatility of ultrasound, on the other hand, may provide increased access to necessary cardiovascular imaging. Echocardiography is indeed the most commonly used noninvasive modality as it combines portable instrumentation, rapid image acquisition and high temporal resolution, without the risks

© 2020 F. Laumer*, G. Fringeli*, A. Dubatovka, L. Manduchi & J. M. Buhmann.

of ionizing radiation (Lang et al., 2015). Its clinical benefits are substantial in countries with limited access to diagnostic equipment, as it offers sensitive case detection at a fraction of common expenses (Beaton et al., 2015; Richter et al., 1990; Sippel et al., 2011). For the above reasons, automatic disease prediction based on affordable medical instruments, such as echocardiography, provides a solution to enhance health care for people with cardiovascular diseases and for those who are at high cardiovascular risk, regardless of their economic realities. This diagnostic tool could alleviate a shortage of medical expertise in countries that have been adversely affected by the migration of their health professionals (Naicker et al., 2009).

In recent years, machine learning has been successfully applied to the detection of different cardiovascular diseases (Zhang et al., 2018; Ouyang et al., 2020)like ... . However, the applicability of machine learning models is still limited due to their black-box nature. This shortcoming is particularly problematic in medicine, where the reasons behind a certain prediction are as important as the prediction itself. As an additional challenge, cardiac ultrasound videos are high-dimensional and contain noisy information. Thus, it is highly desirable to reduce the dimensionality of the data by analysing its periodic patterns and by retaining its discriminative features, to facilitate downstream classification tasks.

For the aforementioned reasons, we propose DeepHeartBeat, a novel autoencoder-based model, to learn an interpretable low dimensional representation of echocardiogram videos (ECHOs) and electrocardiograms (ECGs) in an unsupervised way. Our approach explicitly models the periodic features of the cardiac cycle, providing meaningful insights for practitioners. By enforcing periodic trajectories in the latent space, DeepHeartBeat extracts periodic features of the data as well as their frequencies. The learnt latent embeddings enable then to solve further downstream tasks, such as diagnosis, anomaly detection, denoising, heart rate prediction and heart cycle alignment. We apply our framework to ECHO data, showing human level performance in ejection fraction prediction, and to electrocardiograms, showing good results in arrhythmia detection. Due to its successful application to multiple use cases, our method serves as a promising step towards a general purpose model for the interpretation of periodic sequences such as ECHOs or ECGs.

2. Related work

The growing amount of medical data and the time costs associated with its labeling makes unsupervised learning particularly important in the medical applications. Deep generative models have recently achieved great success in unsupervised representation learning. They generate a compressed latent representation of the data that captures the explanatory factors of the observed input. Some of the most commonly used and efficient approaches are Autoencoders (LeCun et al. (2015)), Variational Autoencoders (Kingma and Welling, 2013) and Generative Adversarial Networks (Goodfellow et al., 2014; Kulkarni et al., 2015). However, such compressed representations are often not humanly intelligible and this opaqueness has often prevented their widespread usage in the medical domain. Disentangled representations, on the other hand, represent an important step towards the direction of learning interpretable encodings of the input, as they recover the independent explanatory factors of variation of the data. Many architectures have been developed in the literature to learn disentangle representations by modifying the deep generative models presented above (Vondrick et al., 2016; Chen et al., 2016; Higgins et al., 2016). Most of these works assume the input data

to be i.i.d., however, in some domains, data naturally come in sequences, where the observations are temporally correlated. While many unsupervised sequence-to-sequence approaches for representation learning already exist (Sutskever et al., 2014; Srivastava et al., 2015), learning temporal representations with disentangled factors of variation is still an open problem. A promising approach was presented by Louis et al. (2019) where the authors modelled the temporal progression of Alzheimer's disease using an encoder-decoder neural network architecture, which maps a patient's brain MRI images taken at different points in time onto a straight trajectory in a latent space, therewith decoupling state and progression velocity of the disease. Another approach presented by Denton et al. (2017) leverages the temporal coherence of video and uses a novel adversarial loss to learn a representation that factorizes each frame into a stationary part and a temporally varying component.

In the medical domain, several works focused on applying machine learning to cardiac ultrasound video data. Madani et al. (2018) trained a convolutional neural network to classify standard views, Zhang et al. (2018) developed a pipeline that automates key aspects of ECHO interpretation, including identifying views, delineating individual cardiac chambers and detecting specific diseases. Ouyang et al. (2020) present a video-based deep learning algorithm that surpasses the performance of human experts in the tasks of segmenting the left ventricle, estimating ejection fraction and assessing cardiomyopathy. Using convolutional neural networks, Ghorbani et al. (2019) showed that deep learning applied to echocardiography can identify local cardiac structures, estimate cardiac function, and predict systemic phenotypes.

While these models are specifically tailored to different tasks, our feature extractor model can be considered task agnostic as it is able to extract informative and well interpretable features which subsequently can be used for a wide range of tasks, achieving comparable or even better performance than previous approaches. Furthermore, the proposed model is applicable to both video data as well as high frequency wave data.

3. Method

We present an autoencoder-based framework for learning cyclic latent trajectories of periodic sequences[1]. An overview of the architecture is sketched in Figure 1.

3.1. Cyclic latent trajectory

We define a sequence of observations as $(\mathbf{y}_j, t_j)_{j=1}^n$ where $\mathbf{y}_j \in \mathbb{R}^D$ represents a sample at each point in time t_j, i.e. for video data $\mathbf{y}_j$ corresponds to the pixels of a particular frame and n is the number of frames. We model the observations of a sequence as following a trajectory over time in a lower dimensional space $\mathcal{Z} = \mathbb{R}^d$, with $d \in \mathbb{N}$, as from now on referred to as latent space. We integrate prior knowledge about the periodicity of the sequence of a particular subject i into our model by using the following cyclic trajectory:

$$\begin{aligned} \boldsymbol{\ell}_i(t) = &\cos(2\pi f_i(t-\tau_i))\boldsymbol{e}_1 \\ &+ \sin(2\pi f_i(t-\tau_i))\boldsymbol{e}_2 \\ &+ \sum_{j=3}^{d} b_i^{(j)} \boldsymbol{e}_j, \end{aligned} \tag{1}$$

where $(\mathbf{e}_1, \ldots, \mathbf{e}_d)$ is the canonical basis of $\mathcal{Z}$ in which the sequences are embedded. The frequency parameter, $f_i > 0$, corresponds to the number of cycles per time unit and the offset parameter τ_i allows the sequences to start at different points in time within the cycle. The b-parameters characterise the

1. The code with experiments is available at: https://github.com/laumerf/DeepHeartBeat

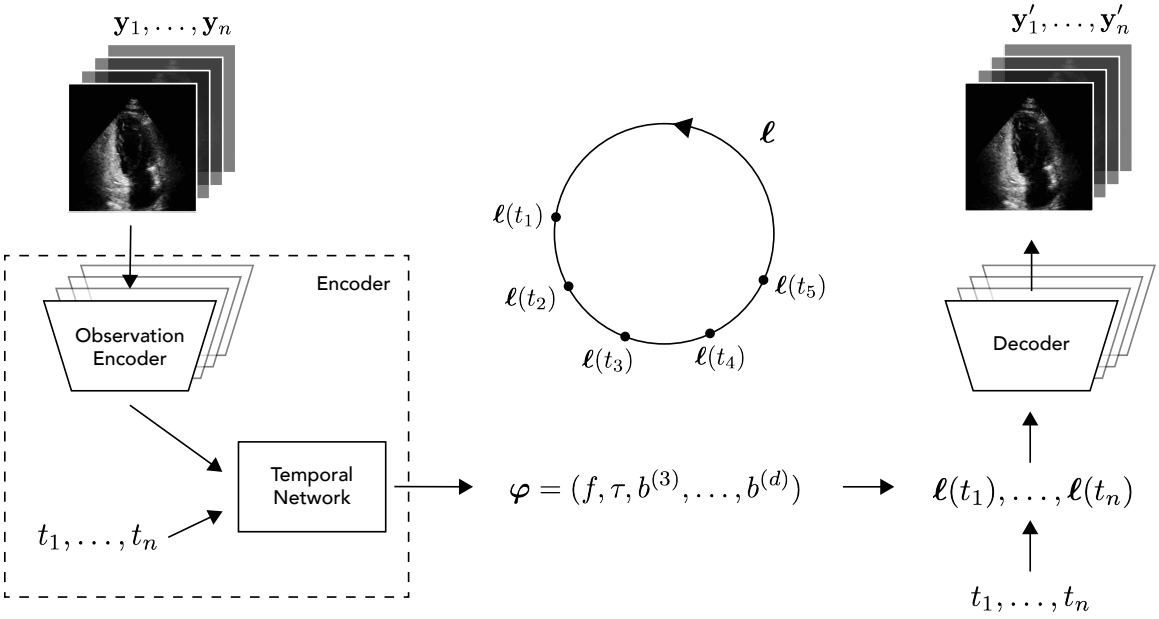

Figure 1: Overview of the proposed network architecture. First, the encoder extracts features from each observation of an input sequence, i.e. echocardiogram. Then, it aggregates the information over time with a temporal network and outputs the parameters ($\boldsymbol{\varphi}$) describing the latent trajectory of the input sequence. Conversely, the decoder takes as input the embeddings on the latent trajectory $\boldsymbol{\ell}$ and maps them back to the observation space.

shape of the signal[2]. The first two dimensions of the latent space are used to describe a circle such that the projection of the trajectory onto the plane spanned by $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ is a unit circle centered at the origin. We define $\boldsymbol{\varphi}_i = (f_i, \tau_i, b_i^{(3)}, \ldots, b_i^{(d)})$ to be the trajectory parameters of subject i.

3.2. Autoencoder Model

The model consists of an encoder and a decoder part. The encoder maps a sequence of observations $(\mathbf{y}_j, t_j)_{j=1}^n$ to trajectory parameters $\boldsymbol{\varphi}$ describing a circular trajectory $\boldsymbol{\ell}(t)$ given by Equation (1). The observation times $t_1, \ldots, t_n$ are subsequently used to calculate the embeddings $\boldsymbol{\ell}(t_1), \ldots, \boldsymbol{\ell}(t_n)$ in the latent space from which the decoder reconstructs the input sequence $\mathbf{y}'_1, \ldots, \mathbf{y}'_n$.

3.2.1. Encoder

The encoder is composed of two neural networks, an observation encoder network (OEN) and a temporal neural network (TNN). The OEN serves as a feature extractor for the observations $\mathbf{y}_j$. These features combined with the corresponding time information t_j are then fed into the TNN, i.e. a LSTM (Graves and Schmidhuber, 2005). The TNN aggregates the inputs over time and outputs the latent parameters $\boldsymbol{\varphi}_i$ for each sequence i.

2. Examples of how different b parameters influence the shape of ECGs can be found in Appendix B.

3.2.2. Decoder

The sample at time t in the temporal sequence i corresponds to the embedding $\boldsymbol{\ell}_i(t) \in \mathcal{Z}$ in the latent space. The decoder maps the embeddings on the latent trajectory $\boldsymbol{\ell}_i$ back to the observation space. Note that the decoder does not directly process the encoder's output $\boldsymbol{\varphi}$, but rather it processes the embeddings lying on the latent trajectory parameterised by the encoder's output.

3.3. Training

Let $\mathcal{D} = \{(\mathbf{y}_{1j}, t_{1j})_{j=1}^{n_1}, \ldots, (\mathbf{y}_{Nj}, t_{Nj})_{j=1}^{n_N}\}$ be a set of N observation sequences and let $\{(\mathbf{y}'_{1j})_{j=1}^{n_1}, \ldots, (\mathbf{y}'_{Nj})_{j=1}^{n_N}\}$ be the corresponding reconstructions, i.e. the output of the decoder. We train the neural network weights by minimising a cost function c, defined as

$$c(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\sigma^2}\frac{1}{n_i}\sum_{j=1}^{n_i} L(\mathbf{y}'_{ij}, \mathbf{y}_{ij}) + r(\boldsymbol{\varphi}_i)\right), \tag{2}$$

which consists of a reconstruction error term and a regulariser r balanced by a trade-off parameter σ^2. L defines a loss-function, i.e. L_1-loss for high frequency wave data and L_2-loss for videos. To give equal weight to all subjects independent of the number of observation in a sequence, we average the frame reconstruction errors for each video. The regulariser forces the shape parameters $b_i^3, \ldots, b_i^d$ to be near the origin, and it is defined as

$$r(\boldsymbol{\varphi}_i) = \sum_{j=3}^{d} (b_i^{(j)})^2. \tag{3}$$

The trade-off parameter σ^2 is estimated based on the loss function. In order to increase stability during training, especially when using small batch sizes N_B, we update σ^2 after each training step by calculating an exponential moving average:

$$\sigma^2_{t+1} \leftarrow (1-\eta)\frac{1}{N_B}\sum_{i\in B}\frac{1}{n_i}\sum_{j=1}^{n_i} L(\mathbf{y}'_{ij}, \mathbf{y}_{ij}) + \eta\sigma_t^2, \tag{4}$$

where parameter $\eta \in [0, 1)$ controls the speed with which the estimate of σ^2 is updated. We train the neural network weights using the Adam optimizer (Kingma and Ba, 2015). A fraction of the sequences is used to compute the following reconstruction error required for early stopping:

$$c_{\text{val}}(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i} L(\mathbf{y}'_{ij}, \mathbf{y}_{ij}). \tag{5}$$

The model is implemented in Python using TensorFlow (version 2.2.0). For more detailed information on the model architectures and training procedure for the different experiments the reader is referred to Appendix A.

4. Datasets

4.1. Echocardiogram video data

For the general assessment of our cardiac cycle mode we use the publicly available EchoNet-Dynamic dataset (Ouyang et al., 2020) consisting of 10,030 apical four-chamber view echocardiograms collected at Stanford Medicine. The published videos have a resolution of 112x112 pixels. The dataset is accompanied by ejection fraction values of the left ventricle and frame numbers of end-systole and end-diastole frames determined by medical practitioners.

4.2. ECG data

To show that our framework is general enough to handle different data modalities, we apply it to electrocardiograms (ECGs) as an example of waveform data. We use the PhysioNet/CinC Challenge 2017 dataset (Clifford et al. (2017)) consisting of 8,528 single-lead

ECG recordings between 9 and 61 seconds in length labeled as one of four classes: recording shows normal sinus rhythm, atrial fibrillation (AF), an alternative rhythm, or is too noisy to be classified.

5. Experiments

In the following, we provide a thorough empirical assessment of our proposed method on the two challenging cardiac datasets described in Section 4. In particular, we use the echocardiogram video data for heart rate detection (5.1.1), echocardiogram video alignment (5.1.2), ejection fraction prediction (5.1.3), and denoising of semantic segmentations of the cardiac chambers (5.1.4). Additionally, we show that DeepHeartBeat can successfully extract the heart rate from ECGs (5.2.1), detect anomalies using ECG noise labels (5.2.2), and classify ECG sequences into atrial fibrillation and other rhythms (5.2.3).

5.1. Echocardiogram video data

The proposed method is able to recognise the heart beat as the fundamental frequency of the cyclic model, as demonstrated in this section, and, thereby, it enables semantic alignment of echocardiogram recordings. Furthermore, based on the latent parameter vector ($\boldsymbol{\varphi}$) learnt by the model, it is possible to reliably predict the ejection fraction using straightforward regression methods.

5.1.1. Heart rate detection

We conjecture that a single cycle along the trajectory $\boldsymbol{\ell}_i$ of a given subject parameterised by $\boldsymbol{\varphi}_i$ corresponds to one complete cardiac cycle. Given the parameterisation of $\boldsymbol{\ell}_i$ this holds if and only if the frequency parameter f_i corresponds to the heart rate in beats per second.

Due to missing ground truth heart rates for the videos in the EchoNet-Dynamic dataset, we compare the heart rates extracted by our model to the heart rates determined by rank-2 Robust Non-negative Matrix Factorisation (RNMF), which was shown to reliably identify the periodic pattern present in echocardiogram videos (Dukler et al., 2018). By using RNMF, we successfully determined the heart rate as described in Appendix C for 8,798 subjects, which are used for comparison with our model.

For a given subject i, let $f_{i,\text{model}}$ and $f_{i,\text{RNMF}}$ denote the model rate and RNMF rate in beats per minute respectively. We define the deviation ϵ_i for subject i as the relative difference with respect to the smaller of the two rates:

$$\epsilon_i = \frac{|f_{i,\text{model}} - f_{i,\text{RNMF}}|}{\min\{f_{i,\text{model}}, f_{i,\text{RNMF}}\}}. \tag{6}$$

The choice of this symmetric metric is motivated by the fact that we cannot assume that $f_{i,\text{RNMF}}$ is the true heart rate. In fact, either rate could be more accurate and, to account for this, ϵ_i is the maximum relative error when assuming either rate to correspond to the ground truth.

When fitting our model five times to the EchoNet-Dynamic dataset (each time using a different train/validation split), we obtain mean deviations between 4.4% and 6.7%, and median deviations ranging from 1.7% to 3.1%.

5.1.2. Semantic Alignment

We expect latent embeddings that are located on the same position $s \in [0, 1)$ on the unit circle to be semantically equivalent, i.e. they should encode the same stage of the cardiac cycle.

To test this hypothesis, we take advantage of the EchoNet-Dynamic dataset. Each video is accompanied by the frame numbers of the end-systole frame and the end-diastole frame, from which we derive the end-systole time $t_{i,\text{systole}}$ and the end-diastole time $t_{i,\text{diastole}}$.

For each subject i, we determine $s_{i,\text{systole}}$ by solving the following equations:

$$\cos(2\pi s_{i,\text{systole}}) = \cos(2\pi f_i(t_{i,\text{systole}} - \tau_i))$$
$$\sin(2\pi s_{i,\text{systole}}) = \sin(2\pi f_i(t_{i,\text{systole}} - \tau_i)).$$

We also compute $s_{i,\text{diastole}}$ similarly. Figure 2 shows the distribution of $s_{i,\text{systole}}$ and $s_{i,\text{diastole}}$. We can observe a concentration of the end-systole values at 0.28 and a concentration of end-diastole values at 0.92.

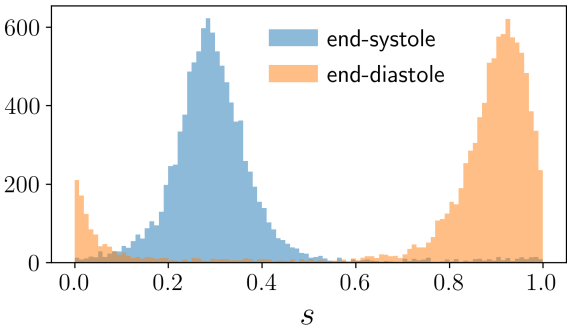

Figure 2: Distribution of end-systole values $s_{i,\text{systole}}$ (blue) and end-diastole values $s_{i,\text{diastole}}$ (orange).

To obtain a quantitative measurement for the semantic alignment, we discretise the values of $s_{i,\text{systole}}$ by rounding to the nearest 100^{th}. Let s_{systole} denote the most frequent value. The absolute deviation of a given value of $s_{i,\text{systole}}$ from s_{systole} is defined as

$$|\Delta s_{i,\text{systole}}| = \min\{|s_{\text{systole}} - s_{i,\text{systole}}|, 1 - |s_{\text{systole}} - s_{i,\text{systole}}|\}. \quad (7)$$

Figure 3 illustrates this concept for two subjects using the projection of the latent trajectory onto the $(\mathbf{e}_1, \mathbf{e}_2)$-plane. We follow the same procedure to calculate the end-diastole deviations $|\Delta s_{i,\text{diastole}}|$. When fitting our model five times to the EchoNet-Dynamic dataset, we achieve mean (median) end-systole deviations between 0.068 (0.047) and 0.085 (0.058). The mean (median) end-diastole deviations range from 0.069 (0.046) to 0.084 (0.056). This time interval corresponds to approx. 3-4 frames deviation from the ground truth assuming a heart rate of 60 bpm and a frame rate of 50 Hz.

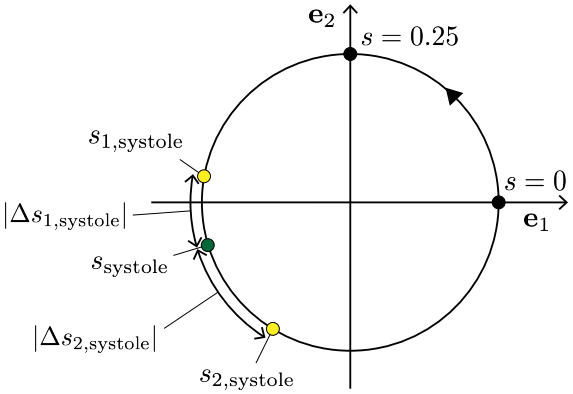

Figure 3: End-systole deviation on the projection of the latent trajectory onto the $(\mathbf{e}_1, \mathbf{e}_2)$-plane

Further experiments and figures for heart rate prediction and semantic alignment are summarized in Appendix D.

5.1.3. Ejection fraction prediction

The ejection fraction of the left ventricle serves as a widely used biomarker of cardiac health. It measures the ratio of the blood volume of the left ventricle ejected during a heart cycle and the left ventricle's end-diastolic volume. The videos of the EchoNet-Dynamic dataset are accompanied by ejection fraction measurements as determined by human experts.

For predicting the ejection fraction, we first determine the latent trajectory parameters $\boldsymbol{\varphi}_i$ of each subject as given by our cardiac cycle model. These parameters are subsequently used as inputs to a neural network which is trained to predict the ejection fraction. We

Table 1: MAE, RMSE, and R^2 scores for different ejection fraction prediction methods.

Methods	MAE	RMSE	R^2
DeepHeartBeat	$\mathbf{6.34 \pm 0.209}$	$\mathbf{8.59 \pm 0.264}$	$\mathbf{0.506 \pm 0.0303}$
R2+1D (Ouyang et al., 2020)	7.35	9.53	0.40
R3D (Ouyang et al., 2020)	7.63	9.75	0.37
MC3 (Ouyang et al., 2020)	6.59	9.39	0.42
EchoNet-Dynamic, Ouyang et al. (2020)	4.05	5.32	0.81

choose a simple neural network with two hidden layers, each one with 1024 nodes and ReLU activation functions. The weights are inferred by an Adam optimiser with learning rate 10^{-4} and mini-batches of size 4.

Repeating this procedure five times results in a MAE of 6.34 ± 0.209, a RMSE of 8.59 ± 0.264, and an R^2 score of 0.506 ± 0.0303 when evaluating the performance on a completely hold out test split. The standard deviation of the corresponding metric is calculated based on 5 different training runs of DeepHeartBeat. Our method is conceptually comparable to the three models presented by Ouyang et al. (2020) which predict the ejection fraction based on the full-length videos without using any addition averaging methods. We are able to surpass the performance of all these three models, of which the best-performing one achieves a MAE of 6.59, a RMSE of 9.39, and an R^2 score of 0.42. We note that the best model presented by Ouyang et al. (2020), named EchoNet-Dynamic, surpasses our performance. However, it is not directly comparable to our work as it applies averaging over subsequences and uses additional human annotations for training. The results are summarized in Table 1.

To put those numbers into perspective, the mean standard deviation of the ejection fraction measurements of different medical practitioners based on the same echocardiogram video amounts to 8.3 (Cole et al., 2015), which is comparable to DeepHeartBeat.

5.1.4. Denoising

To illustrate the denoising effect of our model we apply it to the semantic segmentation of ECHOs as inferred by the neural network model of Zhang et al. (2018). The first column of Figure 4 shows the time evolution of an echocardiogram video of a single patient and the middle column depicts the corresponding frame wise segmentations. One can see that the segmentations are often unsatisfactory for certain frames. For example, the first frame in Figure 4 does not correctly depict the left atrium and the segmentation borders are subject to irregular variations. We fit our model to the segmented videos of the EchoNet-Dynamic dataset by modifying the decoder to output a probability distribution over the six possible segment classes and by choosing the negative log-likelihood as the loss function L in Equation (2). The right column of Figure 4 shows the maximum likelihood reconstructions of the frames in the middle column as calculated by our model. The artefacts described previously have been removed resulting in a smooth deformation over time. For more examples please consult Appendix E.

5.2. Electrocardiograms

5.2.1. Heart rate detection

To show the applicability of DeepHeartBeat to waveform data, we repeat the heart beat experiment for ECG data. To track the progres-

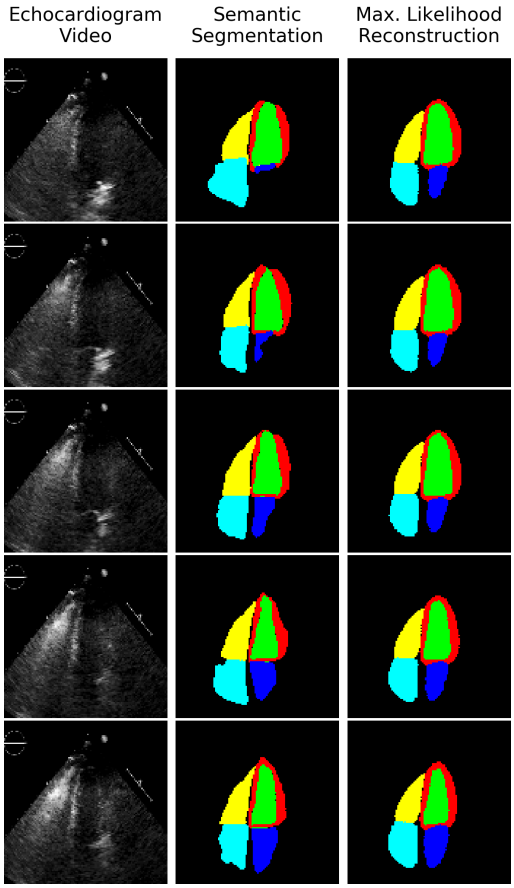

Figure 4: Left: Consecutive raw cardiac ultrasound frames. Center: Corresponding segmentation on a frame by frame basis. Right: The maximum likelihood reconstructions as given by DeepHeartBeat.

sion of the heart rate during the recording, we extract sequences of 2.75 seconds with a stride of 0.01 seconds and encode each sequence separately. Figure 5 shows that the f_i parameter of the trajectory indeed corresponds to changes of the heart rate. Furthermore, one can notice that the phase parameter τ_i expresses periodic behaviour corresponding to the heart cycle. In contrast to the echocardiogram data, the ground truth for ECGs can be obtained by extracting R-peaks and calculating time between them. In Figure 10 in Appendix D one can see how the phase linearly changes as we continuously calculate and plot τ and that its periodicity is aligned with the heart beats.

5.2.2. Anomaly detection

The presence of noise renders the reconstruction and classification of ECG signal a challenging task as such recordings do not show regular periodic behaviour anymore. Since our model is designed to capture periodicity of the input signal, we postulate the hypothesis that the quality of reconstruction can be informative for detecting such noisy signals. As noisy recordings contain only 3.3% of the dataset, we formulate the noise detection as an anomaly detection problem. We encode and reconstruct the full ECG recordings to show that reconstruction error is indeed a good predictor for the noise class yielding an AUC score of 0.81.

5.2.3. Arrhythmia classification

Medical sequence classification defines another clinically relevant downstream tasks. Following the PhysioNet-Challenge, we focus on atrial fibrillation (AF) detection. We combine the trajectory parameters $\boldsymbol{\varphi}$ with the reconstruction error and use them as features for training a SVM classifier to distinguish between AF and other rhythms present in the data. The b parameters encode information about the shape of a normal heart beat vs. an AF heart beat as visualized in Figure 6. Our representation allows to reach a balanced accuracy of 0.70 on the hold out test set provided by PhysioNet Challenge consisting of 300 ECG recordings. This result is on par with human expert performance in AF classification (Hannun et al., 2019). Note that we use only one trajectory embedding per ECG which does not allow to track short changes

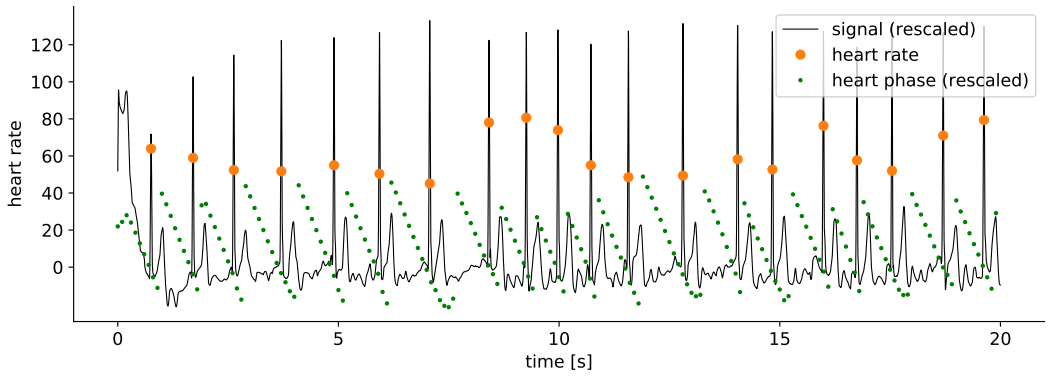

Figure 5: We depict the change in heart rate over a sliding subsequence of an ECG recording starting at the moment when R-peaks occur. Note that ECG signal and heart phase are re-scaled for visualization purposes.

within the signal. Aggregating trajectory parameters from different subsequences of the original recording should further improve the results, but is left for future work since it does not directly relate to the representation learning approach presented in this paper.

6. Conclusion

We presented a novel autoencoder-based model to learn geometrically interpretable low dimensional representations of ECHOs and ECGs in an unsupervised way. The model maps high dimensional observations to a cyclic trajectory in a lower dimensional space. We evaluated our model on large publicly available data sets in a variety of experiments and, thereby, touching upon several aspects of practical interest including heart rate detection, as well as end-systole and end-diastole identification. Furthermore, we demonstrated that the extracted shape parameters can be used for reliably detecting arrhythmia in ECGs and we achieve human comparable performance in predicting the ejection fraction based on cardiac ultrasound data. Due to its successful application to multiple use cases, our method is a promising step towards a general purpose model for extracting interpretable and informative features from periodic sequences.

6.1. Limitation and future work

Our approach assumes a constant cycle frequency, a clear limitation for its applicability to patients suffering e.g. from arrhythmia. For future work, we suggest to overcome this deficiency by adapting the parameterization to capture phase and frequency changes which would further increase the applicability of our model.

In the past, machine learning has been used to detect and classify a wide range of cardiac conditions based on ECG and ECHO data. Motivated by our promising experiments, it would be interesting to explore to what degree the features extracted by DeepHeartBeat with its separation of dynamic and static features will improve on those works. Especially, in situation where task-specific labels are scarce, using another larger task-agnostic dataset for training DeepHeartBeat, could improve the quality of the extracted task-specific features, and, consequently, also

the performance of the down-stream task of interest.

Acknowledgments

FL, AD and LM have been supported by PHRT - SHFN / SWISSHEART Failure Network (JMB, PI); we thank Julia Vogt for valuable discussions.

References

A. Beaton, J. Lu, T. Aliku, Peter N. Dean, L. Gaur, J. Weinberg, Justin Godown, P. Lwabi, Grace Mirembe, E. Okello, A. Reese, Ashley Shrestha-Astudillo, Tyler Bradley-Hewitt, Janet N Scheel, C. Webb, R. McCarter, Greg J Ensing, and C. Sable. The utility of handheld echocardiography for early rheumatic heart disease diagnosis: a field study. *European heart journal cardiovascular Imaging*, 16 5:475–82, 2015.

N.G Bellenger, M.I Burgess, S.G Ray, A Lahiri, A.J.S Coats, J.G.F Cleland, and D.J Pennell. Comparison of left ventricular ejection fraction and volumes in heart failure by echocardiography, radionuclide ventriculography and cardiovascular magnetic resonance. Are they interchangeable? *European Heart Journal*, 21(16): 1387–1396, 08 2000. ISSN 0195-668X. doi: 10.1053/euhj.2000.2011. URL https://doi.org/10.1053/euhj.2000.2011.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

Gari Clifford, Chengyu Liu, Benjamin Moody, Li-wei Lehman, Ikaro Silva, Qiao Li, Alistair Johnson, and Roger Mark. Af classification from a short single lead ecg recording: the physionet computing in cardiology challenge 2017. 09 2017. doi: 10.22489/CinC.2017.065-469.

Graham D. Cole, Niti M. Dhutia, Matthew J. Shun-Shin, Keith Willson, James Harrison, Claire E. Raphael, Massoud Zolgharni, Jamil Mayet, and Darrel P. Francis. Defining the real-world reproducibility of visual grading of left ventricular function and visual estimation of left ventricular ejection fraction: impact of image quality, experience and accreditation. *International Journal of Cardiovascular Imaging*, 31(7): 1303–1314, 2015. ISSN 15730743. doi: 10.1007/s10554-015-0659-1.

Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*, pages 4414–4423, 2017.

Yonatan Dukler, Yurun Ge, Yizhou Qian, Shintaro Yamamoto, Baichuan Yuan, Long Zhao, Andrea L. Bertozzi, Blake Hunter, Rafael Llerena, and Jesse T. Yen. Automatic valve segmentation in cardiac ultrasound time series data. In *Medical Imaging 2018: Image Processing*, page 69, 2018. ISBN 9781510616370. doi: 10.1117/12.2293255.

A. Ghorbani, D. Ouyang, Abubakar Abid, Bryan D. He, J. Chen, R. Harrington, D. Liang, E. Ashley, and J. Zou. Deep learning interpretation of echocardiograms. *bioRxiv*, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6): 602–610, 2005. ISSN 08936080. doi: 10.1016/j.neunet.2005.06.042.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory

electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65, 2019.

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

P. Jeemon, D. Prabhakaran, M. Huffman, S. Harikrishnan, and Stephen R. Leeder. A race against time ii: the challenge of cardiovascular diseases in developing economies. 2014.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–15, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in neural information processing systems*, pages 2539–2547, 2015.

R. Lang, L. Badano, V. Mor-Avi, J. Afilalo, A. Armstrong, L. Ernande, F. Flachskampf, E. Foster, S. Goldstein, T. Kuznetsova, P. Lancellotti, D. Muraru, Michael H. Picard, E. R. Rietzschel, L. Rudski, K. Spencer, W. Tsang, and J. Voigt. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging. *European heart journal cardiovascular Imaging*, 16 3: 233–70, 2015.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

Maxime Louis, Raphaël Couronné, Igor Koval, Benjamin Charlier, and Stanley Durrleman. Riemannian Geometry Learning for Disease Progression Modelling. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11492 LNCS:542–553, 2019. ISSN 16113349. doi: 10.1007/978-3-030-20351-1_42.

Ali Madani, Ramy Arnaout, Mohammad Mofrad, and Rima Arnaout. Fast and accurate view classification of echocardiograms using deep learning. *NPJ digital medicine*, 1(1):1–8, 2018.

S. Naicker, J. Plange-Rhule, R. Tutt, and J. B. Eastwood. Shortage of healthcare workers in developing countries–africa. *Ethnicity & disease*, 19 1 Suppl 1:S1–60–4, 2009.

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580 (7802):252–256, 2020.

J. Richter, A. Dengler, E. Mohammed, G. M. Ali, I. Abdel-Rahim, C. Kaiser, and E. Doehring-Schwerdtfeger. Results of echocardiographic examinations in a regional hospital of central sudan. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 84 5:749–52, 1990.

V. Roger, A. Go, D. Lloyd-Jones, R. Adams, J. Berry, T. Brown, M. Carnethon, S. Dai, G. de Simone, E. S. Ford, C. Fox, H. Fullerton, C. Gillespie, K. Greenlund, S. Hailpern, J. Heit, P. M. Ho, V. Howard, B. Kissela, S. J. Kittner, D. Lackland, J. H.

Lichtman, L. Lisabeth, D. M. Makuc, G. M. Marcus, A. Marelli, D. B. Matchar, M. M. McDermott, J. Meigs, C. S. Moy, D. Mozaffarian, M. Mussolino, G. Nichol, N. P. Paynter, W. Rosamond, P. Sorlie, R. Stafford, T. N. Turan, M. Turner, N. D. Wong, and J. Wylie-Rosett. Heart disease and stroke statistics–2011 update: a report from the american heart association. *Circulation*, 123 4:e18–e209, 2011.

Stephanie Sippel, K. Muruganandan, A. Levine, and Sachita Shah. Review article: Use of ultrasound in the developing world. *International Journal of Emergency Medicine*, 4:72 – 72, 2011.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *Advances in neural information processing systems*, pages 613–621, 2016.

Jeffrey Zhang, Sravani Gajjala, Pulkit Agrawal, Geoffrey H Tison, Laura A Hallock, Lauren Beussink-Nelson, Mats H Lassen, Eugene Fan, Mandar A Aras, ChaRandle Jordan, et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation*, 138(16):1623–1635, 2018.

B. Ziaeian and G. Fonarow. Epidemiology and aetiology of heart failure. *Nature Reviews Cardiology*, 13:368–378, 2016.

Appendix A. Implementation details

A.1. Echocardiogram video data

The layer-by-layer description of the neural network architecture of the encoder can be found in Table 2 (observation encoder network) and Table 3 (temporal neural network). Table 4 describes the decoder architecture.

Layer type	**Output shape**
Input (frame, pixel values scaled to [0, 1]).	$112 \times 112 \times 1$
4×4 conv. 8 filters. stride 2. ReLU	$55 \times 55 \times 8$
4×4 conv. 16 filters. stride 2. ReLU	$26 \times 26 \times 16$
4×4 conv. 16 filters. stride 2. ReLU	$12 \times 12 \times 16$
4×4 conv. 16 filters. stride 2. ReLU	$5 \times 5 \times 16$
Flatting.	400

Table 2: Neural network architecture of the observation encoder network (OEN) part of the encoder.

Layer type	**Output shape**
Input (sequence of frame embeddings).	$m \times 401$
Bidirectional LSTM. 128 units per direction. Tanh.	256
Fully-connected. d units. Linear.	$d = 128$

Table 3: Neural network architecture of the temporal neural network (TNN) part of the encoder. The size of the output depends on the dimensionality of the latent space. m denotes the sequence length.

Layer type	**Output shape**
Input (point in latent space).	$d = 128$
Fully-connected. 400 units. Linear.	400
Reshaping.	$5 \times 5 \times 16$
4×4 transp. conv. 16 filters. stride 2. ReLU	$12 \times 12 \times 16$
4×4 transp. conv. 16 filters. stride 2. ReLU	$26 \times 26 \times 16$
4×4 transp. conv. 8 filters. stride 2. ReLU	$55 \times 55 \times 8$
4×4 transp. conv. 1 filter. stride 2. Sigmoid	$112 \times 112 \times 1$

Table 4: Neural network architecture of the decoder. The size of the input depends on the dimensionality of the latent space.

For the Adam optimiser, we choose a learning rate of $\alpha = 5 \cdot 10^{-4}$ and exponential decay rates of $\beta_1 = 0.9$ for the first and $\beta_2 = 0.999$ for the second moment estimates. We run 200 iterations per epoch and use mini-batches consisting of $N_B = 32$ subsequences of duration

at least two seconds randomly sampled from the set of training videos. For the exponential average update of the reconstruction/regularity trade-off parameter σ^2 in equation (4) we select $\eta = 0.99$. The L_2 norm was used as the loss function L.

A.2. Electrocardiogram

The layer-by-layer description of the neural network architecture of the encoder can be found in Table 5 (observation encoder network) and Table 6 (temporal network). Table 7 describes the decoder architecture.

Layer type	**Output shape**
Input (single value in $\mathbb{R}$).	1
Fully-connected. 16 units. ReLU.	16

Table 5: Neural network architecture of the observation encoder network (OEN) part of the encoder.

Layer type	**Output shape**
Input (sequence of frame embeddings).	$m \times 17$
Bidirectional LSTM. 128 units per direction. Tanh.	$m \times 256$
Bidirectional LSTM. 128 units per direction. Tanh.	256
Fully-connected. d units. Linear.	$d = 8$

Table 6: Neural network architecture of the temporal neural network (TNN) part of the encoder. The size of the output depends on the dimensionality of the latent space. m denotes the sequence length.

Layer type	**Output shape**
Input (point in latent space).	$d = 8$
Fully-connected. 128 units. ReLU.	128
Fully-connected. 128 units. ReLU.	128
Fully-connected. 1 unit. Linear.	1

Table 7: Neural network architecture of the decoder. The size of the input depends on the dimensionality of the latent space.

For the Adam optimiser, we choose a learning rate of $\alpha = 5 \cdot 10^{-4}$. The exponential decay rates are set to $\beta_1 = 0.9$ for the first and $\beta_2 = 0.999$ for the second moment estimates. We run 1000 iterations per epoch and use mini-batches consisting of $N_B = 64$ subsequences of duration between 1.5 and 4.0 seconds randomly sampled from the set of training subjects. For the exponential average update of the reconstruction/regularity trade-off parameter σ^2 in equation (4) we select $\eta = 0.99$. The L_1 norm was used as the loss function L.

To prevent the model from encoding multiple heart cycles for a single cycle along the latent trajectory, we penalise frequencies f that correspond to heart rates below 40 bpm by extending the regularisation term in Equation (3) to

$$r(\boldsymbol{\varphi}_i) = \sum_{j=3}^{d} (b_i^{(j)})^2 + \max\left\{0, \frac{40}{60} - f_i\right\}.$$

Appendix B. Visualization b-parameters for ECG

We visualize the reconstruction of a normal heart beat and an AF heart beat. In order to visualize how the different b-parameters influence the shape of an ECG, we linearly interpolate each b-parameter from a normal signal to the corresponding AF parameter and reconstruct the generated signal over one heart beat. In Figure 6, each b is changed separately while keeping the other bs fix. In Figure 7, we linearly interpolate all parameters simultaneously from a normal signal representation to an AF representation.

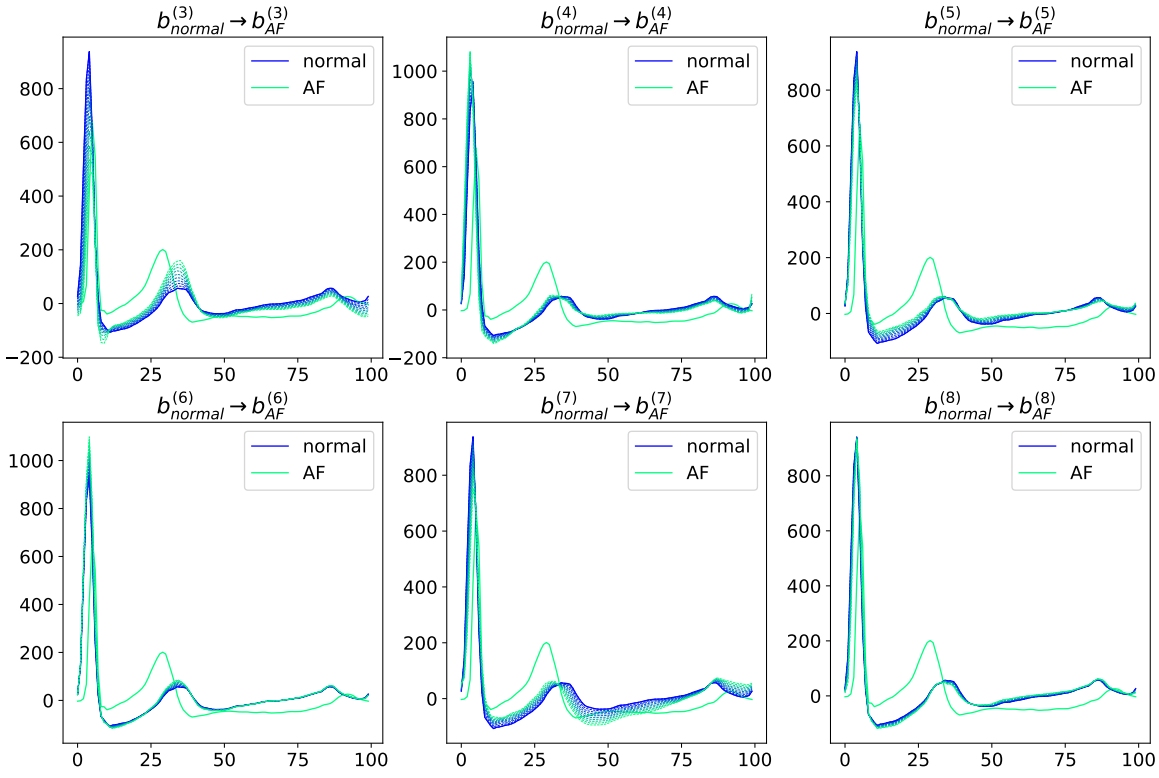

Figure 6: We visualize the reconstructed signals over a heartbeat when linearly interpolating each b parameter of a normal signal to the b parameter of an AF signal.

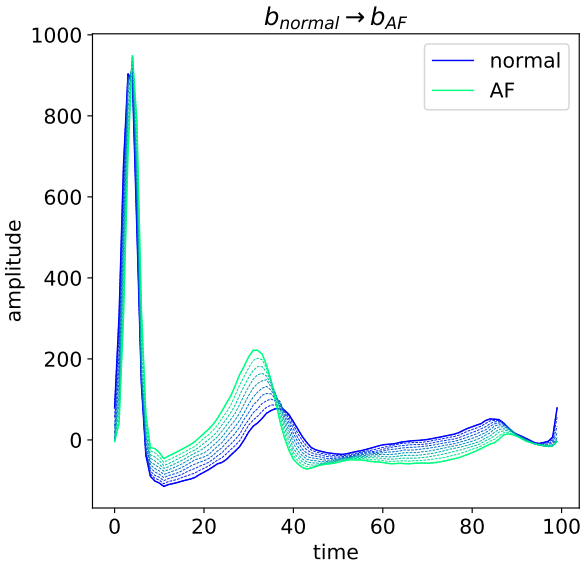

Figure 7: We visualize the reconstructed signals over a heartbeat when linearly interpolating the b vector of a normal signal to the b vector corresponding to an AF signal.

Appendix C. Rank-2 Robust Non-negative Matrix Factorisation for heart rate detection

We describe the procedure used to determine the heart rate of a echocardiographic video based on rank-2 Robust Non-negative Matrix Factorisation (RNMF) (Dukler et al., 2018).

Assume we are given an echocardiographic video consisting of m frames with n pixels each, which we represent as a non-negative matrix $\mathbf{X} \in \mathbb{R}_+^{n \times m}$. RNMF aims to find matrices $\mathbf{W} \in \mathbb{R}_+^{n \times k}$, $\mathbf{H} \in \mathbb{R}_+^{k \times m}$, and $\mathbf{S} \in \mathbb{R}^{n \times m}$ such that

$$\mathbf{X} \approx \mathbf{WH} + \mathbf{S}. \tag{8}$$

This is achieved by minimising the energy function

$$f(\mathbf{W}, \mathbf{H}, \mathbf{S}) = \|\mathbf{X} - \mathbf{WH} - \mathbf{S}\|_F^2 + \lambda \|\mathbf{S}\|_1 \tag{9}$$

where $\lambda > 0$ controls the sparsity in $\mathbf{S}$.

We choose $k = 2$, $\lambda = 0.1$, and determine $\mathbf{W} \in \mathbb{R}_+^{n \times 2}$, $\mathbf{H} \in \mathbb{R}_+^{2 \times m}$, and $\mathbf{S} \in \mathbb{R}^{n \times m}$ using the iterative thresholding approach presented by (Dukler et al., 2018). If successful, the coefficients in both rows of $\mathbf{H}$, denoted by $\mathbf{h}_1$ and $\mathbf{h}_2$, should exhibit a periodic pattern. To determine the frequency of this pattern we fit the simple sine model

$$h_i \approx a \sin(2\pi t_i f + d) + bt + c, \quad i = 1, \ldots, m \tag{10}$$

to both $\mathbf{h}_1$ and $\mathbf{h}_2$ independently. t_i denotes the time of the i^{th} frame and h_i denotes the i^{th} element of either $\mathbf{h}_1$ or $\mathbf{h}_2$. We optimise a, b, c, d, f by minimising the mean squared error. Figure 8 shows an example for an optimal sine fit.

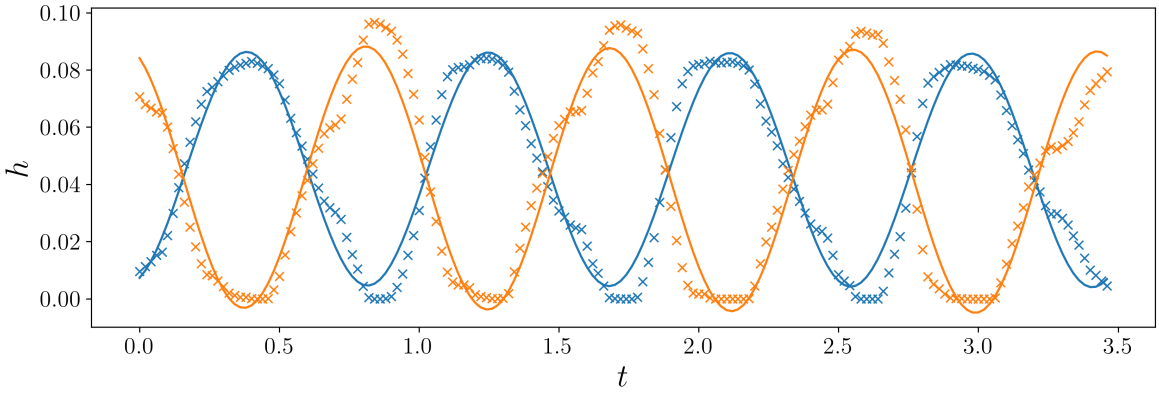

Figure 8: Sine model fits for $\mathbf{h}_1$ (blue) and $\mathbf{h}_2$ (orange) of a given echocardiographic video. The crosses show the values of the elements of $\mathbf{h}_1$ and $\mathbf{h}_2$. The solid lines are the corresponding optimal sine model fits.

We end up with two frequency values f_1 and f_2, one for each row in $\mathbf{H}$, corresponding to the heart rate in beats per second. The final heart rate in beats per minute is $f^* = 60\frac{f_1+f_2}{2}$. As we do not always get reasonable heart rates, we define the heart rate detection to be successful if the difference $|f_1 - f_2|$ amounts to less than 10% of $\min\{f_1, f_2\}$ and if the final value f^* ends up in the range $[45, 180]$.

Appendix D. Heart rate and semantic alignment

D.1. Echocardiogram

We use the trained model to continuously extract the frequency f as well as the phase τ of a sliding window consisting of 90 frames (Figure 9). Our model is able to track the phase of the cardiac cycle as well as changing heart rate over an echocardiographic video. This could for example be used for detecting arrhythmia solely based on echocardiographic videos as shown in the lower part of Figure 9.

D.2. ECG

We use the trained model to continuously extract the frequency f as well as the phase τ of a sliding window consisting of .75 seconds. Our model is able to track the phase/frequency (Figure 10) of the cardiac cycle as well as changing heart rate (Figure 5) over an ECG recording.

Appendix E. Denoising semantic segmentations

Some additional qualitative examples of improved and denoised semantic segmentation. The segmented cardiac regions are: left ventricle blood pool (green), left atrium blood pool (blue), left ventricle myocardium (red), right ventricle blood pool (yellow), and right atrium blood pool (cyan).

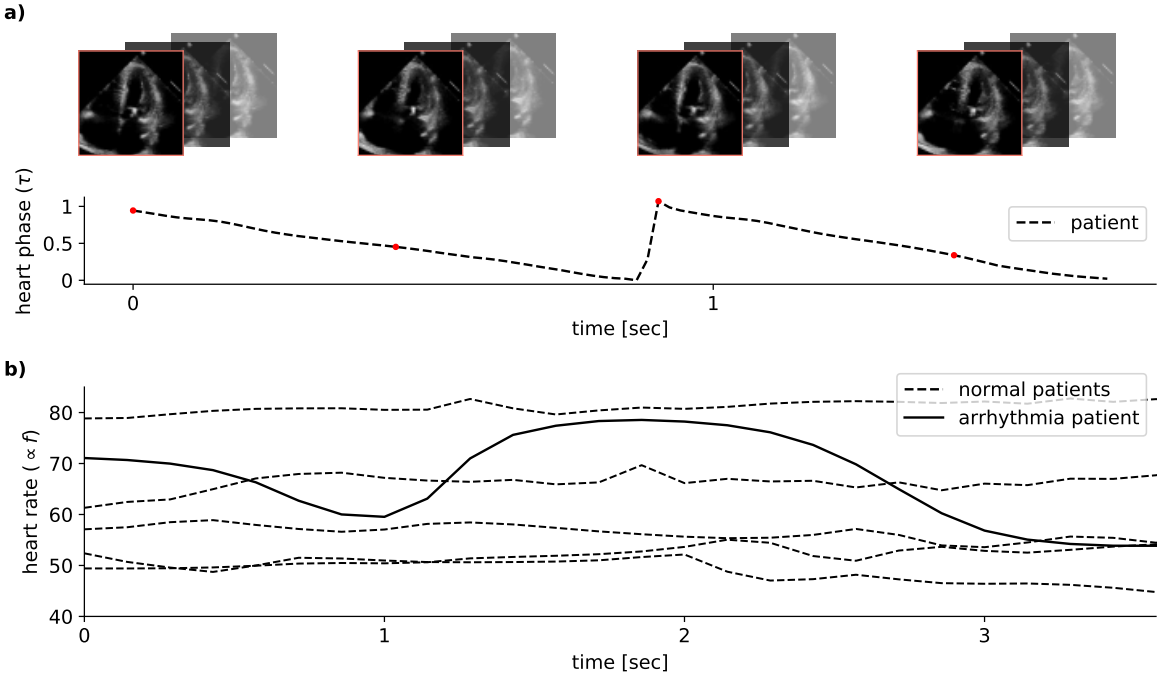

Figure 9: We depict the change in heart phase and heart rate over a sliding subsequence of an ECHO. In the upper part, one can see how the phase linearly changes as we continuously calculate and plot τ. In the bottom part, we track the heart rate over a couple of heart beats. The difference between normal patients and an arrhythmia patient is clearly visible.

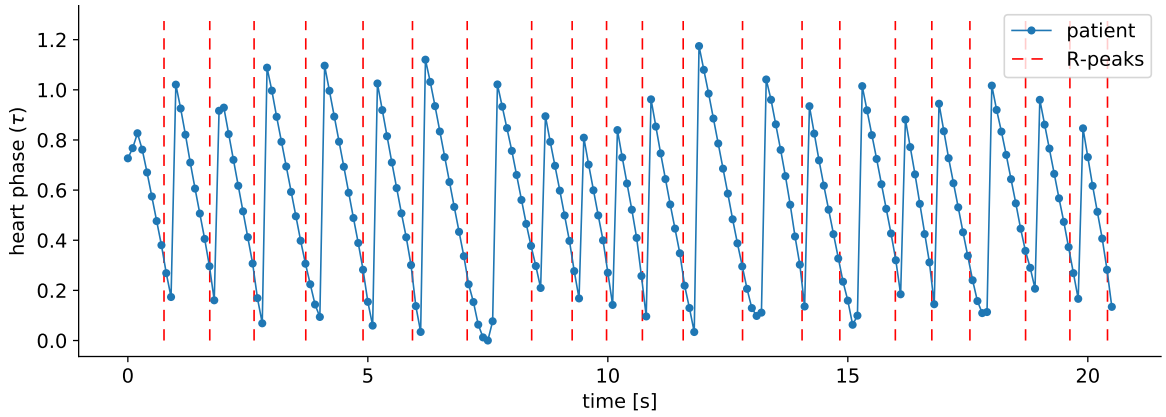

Figure 10: We depict the change in heart phase over a sliding subsequence of an ECG recording. One can see how the phase linearly changes as we continuously calculate and plot τ and that its periodicity is aligned with the heart beats.

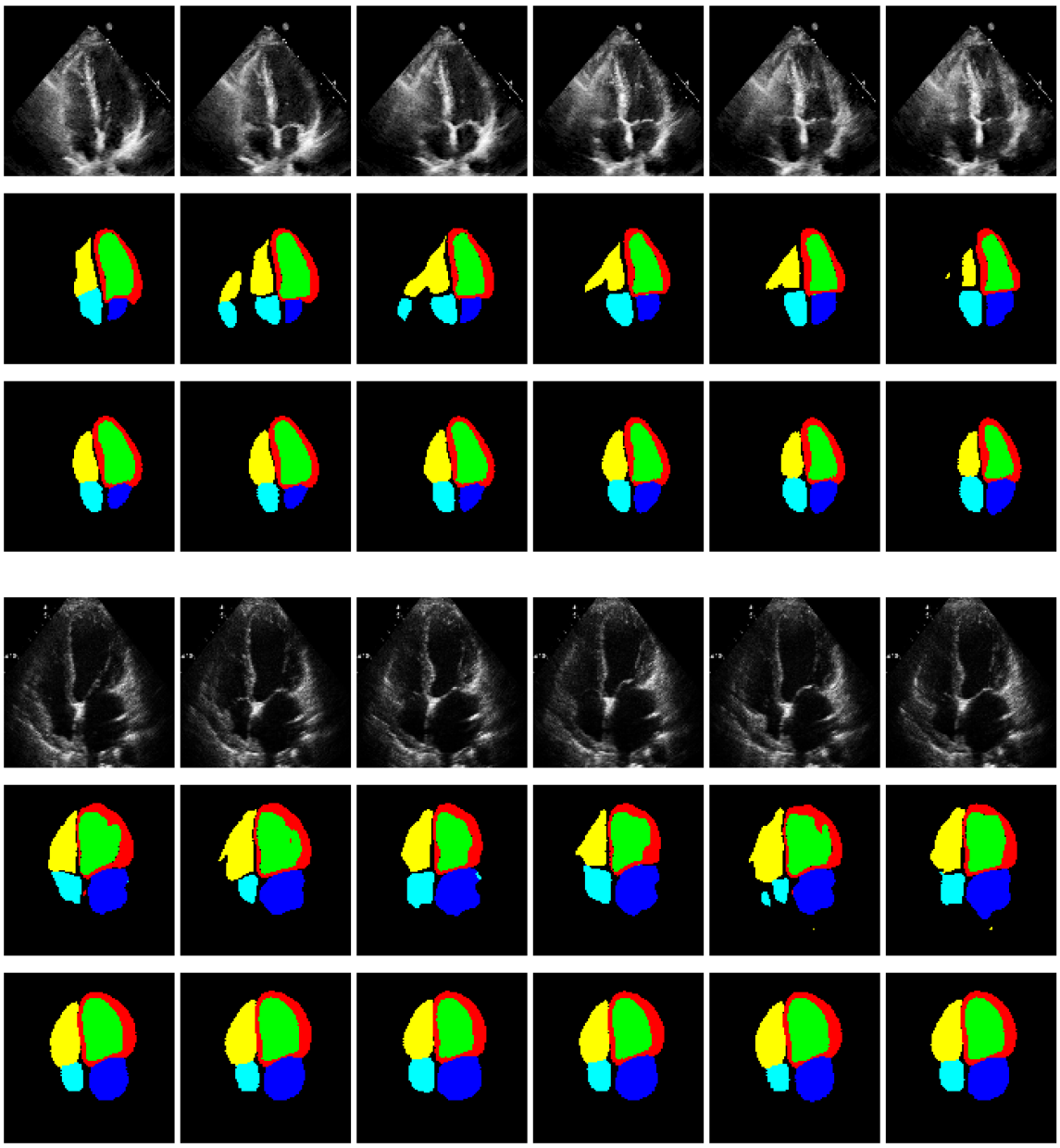

Figure 11: Upper row: Series of consecutive raw cardiac ultrasound frames. Middle row: Semantic segmentation (Zhang et al., 2018). Bottom row: Maximum likelihood reconstruction as given by DeepHeartBeat.

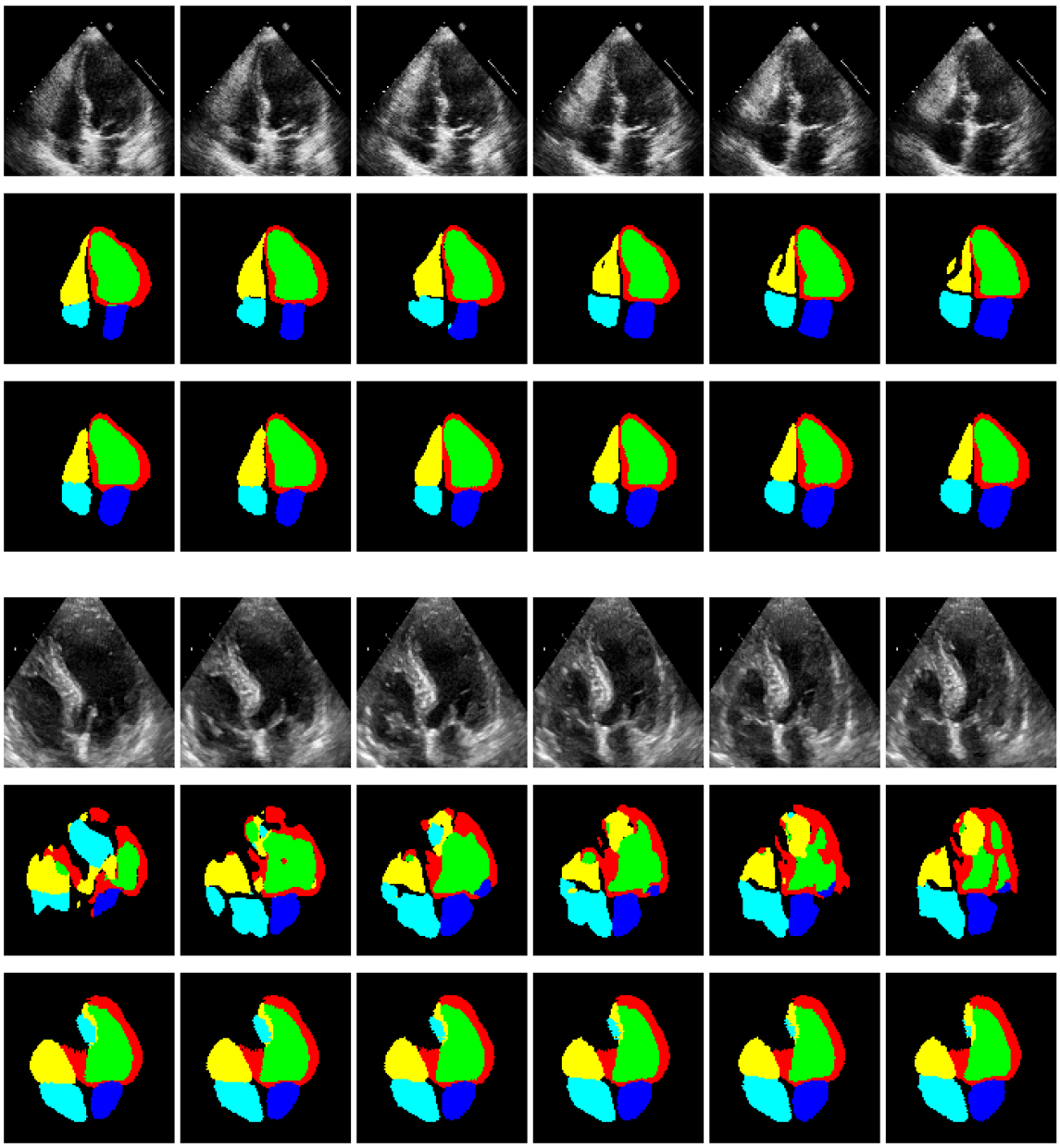

Figure 12: Upper row: Series of consecutive raw cardiac ultrasound frames. Middle row: Semantic segmentation (Zhang et al., 2018). Bottom row: Maximum likelihood reconstruction as given by DeepHeartBeat.

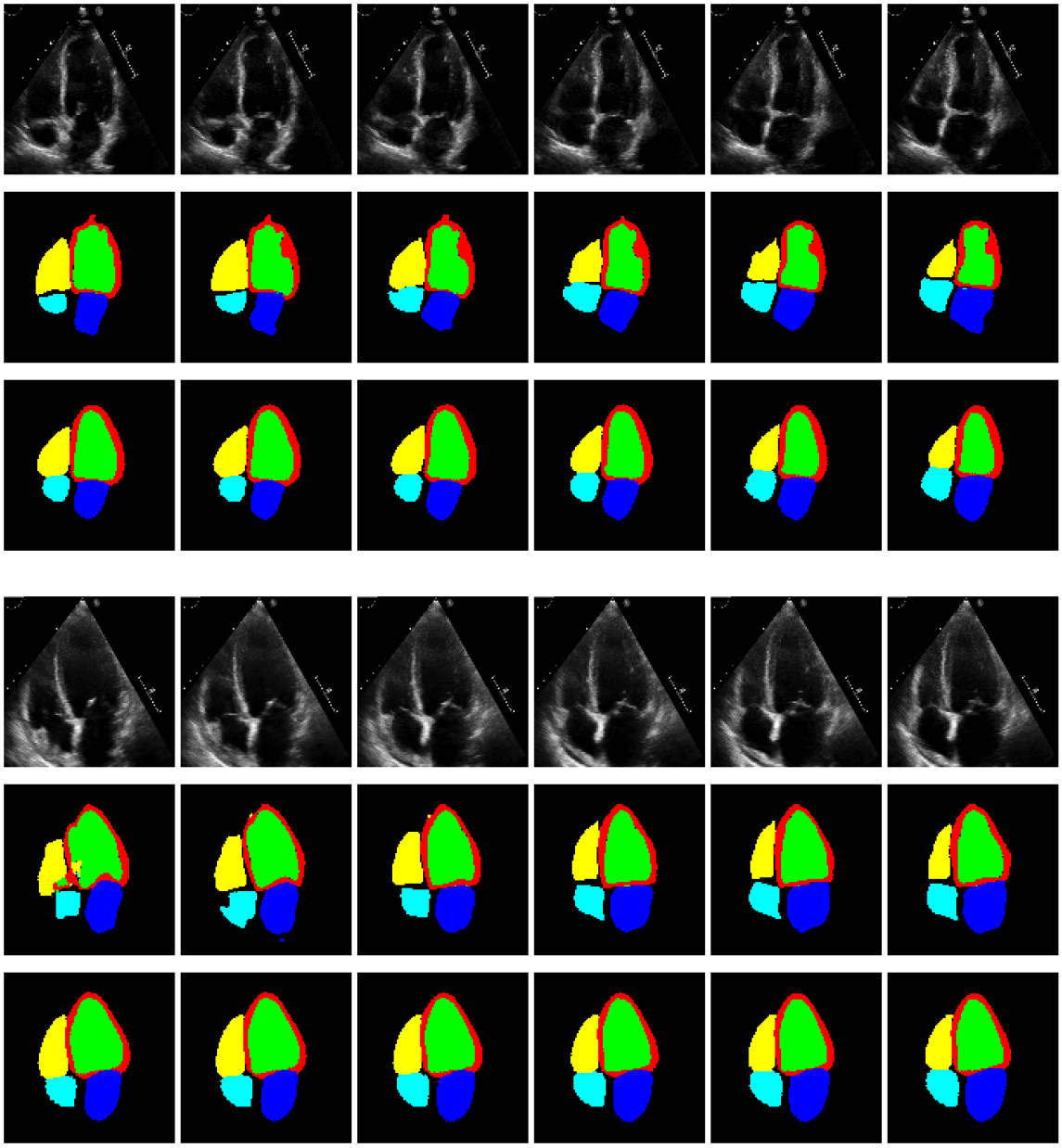

Figure 13: Upper row: Series of consecutive raw cardiac ultrasound frames. Middle row: Semantic segmentation (Zhang et al., 2018). Bottom row: Maximum likelihood reconstruction as given by DeepHeartBeat.

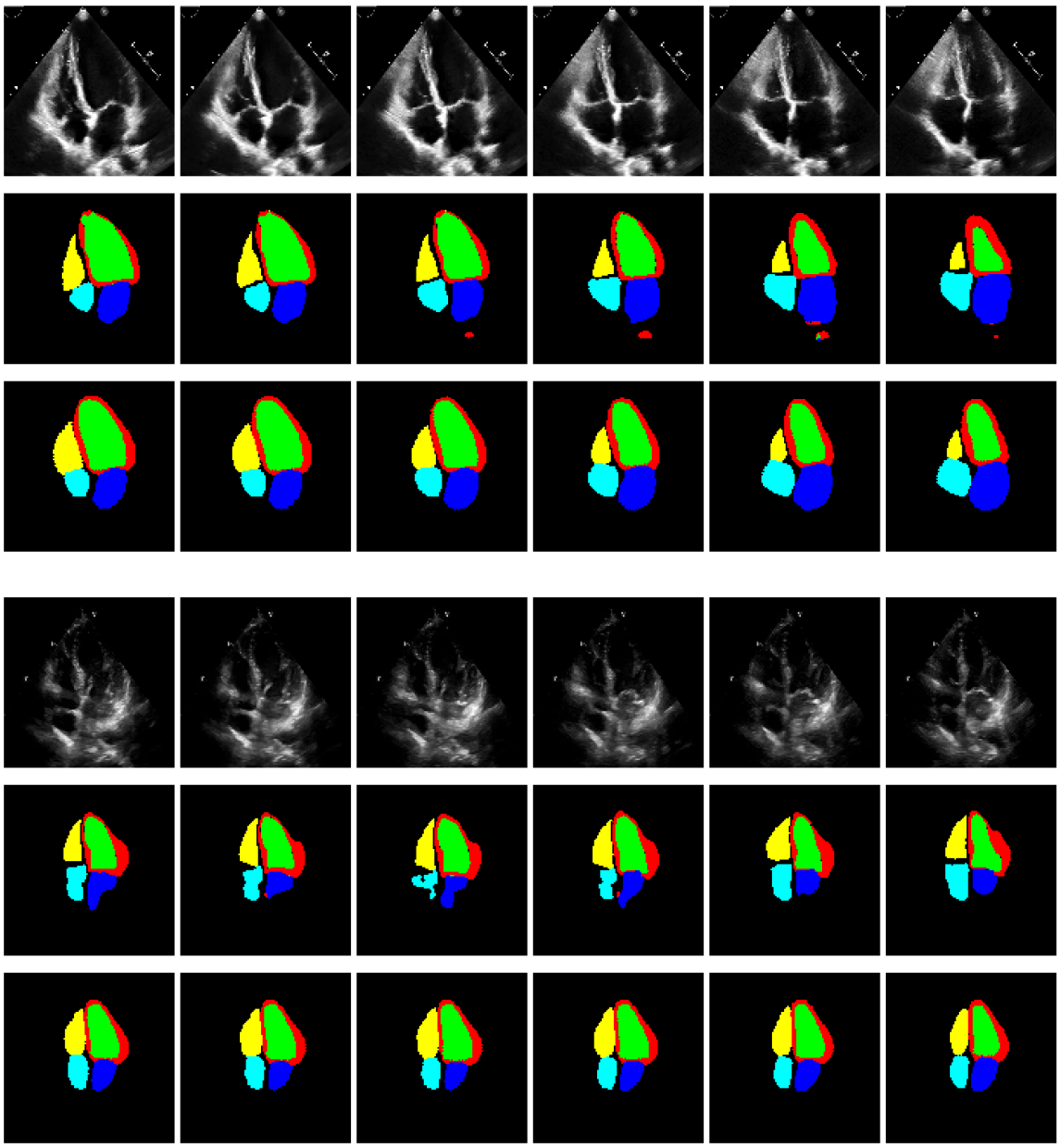

Figure 14: Upper row: Series of consecutive raw cardiac ultrasound frames. Middle row: Semantic segmentation (Zhang et al., 2018). Bottom row: Maximum likelihood reconstruction as given by DeepHeartBeat.

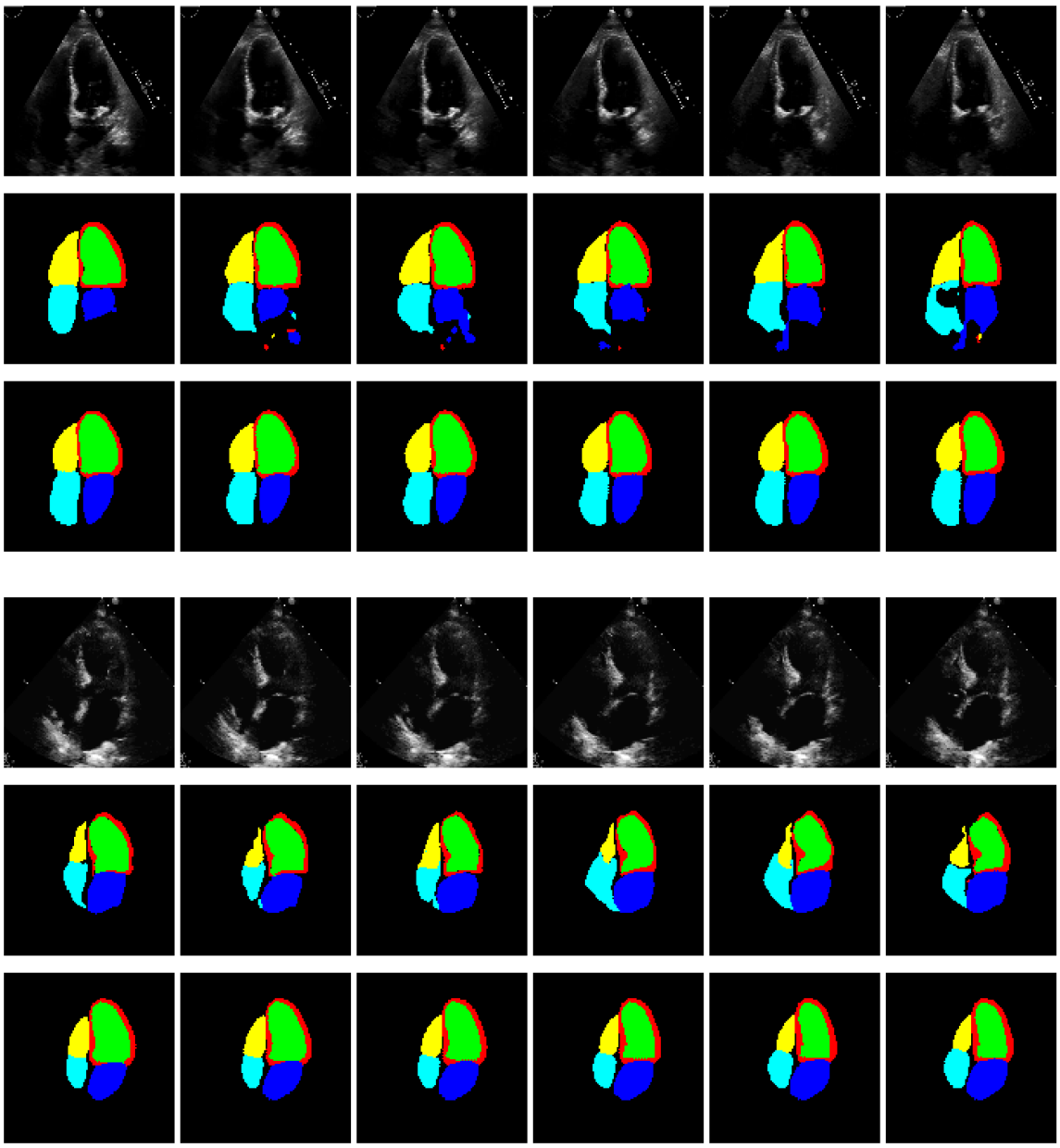

Figure 15: Upper row: Series of consecutive raw cardiac ultrasound frames. Middle row: Semantic segmentation (Zhang et al., 2018). Bottom row: Maximum likelihood reconstruction as given by DeepHeartBeat.