ECHO-Technical Report

Christina Bukas et al.

July 2020

1 Introduction

Echocardiography is an essential tool in modern cardiology, with image visualization and processing being critical for medical diagnosis. Correctly deciphering such images typically requires extensive training and practice, due to their complexity and general uninterpretability. Typically, in standard laboratory procedures dozens of mice are scanned weekly and manual annotations are required for further analysis and exploration. Automating such procedures can lead to big imaging mouse data, thereby providing new possibilities to generate accurate and consistent interpretation of echocardiograms. To this end, we propose a fully automated end to end framework for feature extraction from M-mode transthoracic echocardiography of mice (ATEFEX).

To ensure the method's accuracy, we aim, as a first step, to detect good and bad acquisition regions in echocardiograms. During an acquisition sudden movements or device misplacement may lead to a deterioration of the quality of acquisition and features in the image which do not truly represent the mouse's heart state can appear. Distinguishing between good and bad regions of acquisition leads to inclusion of only the former during automatic feature extraction and thereby ensuring that wrong measurements are excluded. This is done by training a classification network to determine good and bad quality regions and inserting the trained network into the framework.

As mentioned, The daily increasing number of images stretches the limits of manual annotation and thus technicians generally annotate a small section of three consecutive heartbeats in an echocardiogram, according to the recommendations of the American Society of Echocardiography [1]. We propose a method that can automatically extract all features over a complete good recording time, resulting in a generation of large amounts of measurements. This can lead to the discovery of potentially new and yet unexplored cardiac phenotypes and lay the ground for further exploration of vast and currently unexploited data. For this purpose, a second network is trained to segment the Inner Left Ventricle Diameters (LVID) in the echocardiogram. The trained network is then included into the framework and the generated segmentation masks are used for feature extraction.

2 Methods

2.1 Data description

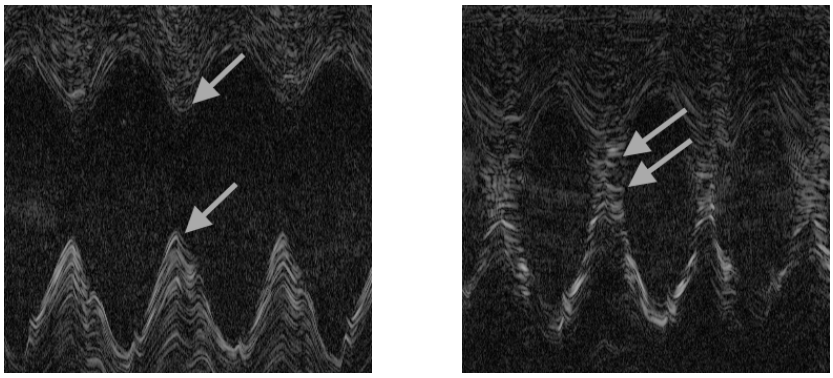
The dataset provided includes a total of 60 mice, consisting of a mixture of males and females, as well as mutants and controls, in order to ensure a large amount of variability in the data. All mice were taken from the ... experiment, had a weight ranging from ... to ... and were at an age of ... at the time of acquisition. All echocardiograms were acquired in M-mode, using a linear transducer, with a focus depth of 0.6 and an imaging frequency of 40 MHz. The images depict the Left Ventricle Inner Diameter of the heart over time. Each echocardiogram is stored as a DICOM file under the Ultrasound Multi-frame Image Storage standard and consists of 49 frames. All acquisitions have a total length of 4.869 seconds with a pixel resolution of 0.833 ms in the x-axis, while the resolution in the y-axis varied for each acquisition.

For all 60 mice, regions of good and bad quality of acquisition were defined and provided by expert technicians. These regions have a minimum length of 0.3 seconds and each acquisition can have one or multiple good and bad regions. In addition, technicians annotated the left ventricle upper and lower trace in all good regions of acquisition, which served as ground truth for the segmentation task. For the task of acquisition quality classification 10 mice were set aside for testing the network's performance, while for the task of segmenting the LVIDs six mice were set aside for evaluation of the network's performance.

2.2 Dataset preparation

In order to make the data accessible for training a neural network several preprocessing steps were required. For training two separate models, one for image classification, the other for segmentation, we created two separate training and testing datasets from the original data. Firstly, the good and bad regions, defined by the technicians, needed to be cropped from the original acquisitions. Since the 49 frames consist of overlapping periods which makes it difficult to handle continuous regions, as a first step we concatenated the frames to acquire one long image array corresponding to the full time of acquisition. This single image made it then easier to crop images corresponding to only good or bad quality acquisition regions. However, since the annotations were of varying length this left us with a rather diverse dataset as some regions could be much longer than others *example images?*. From here on, we refer to images with annotations corresponding to good quality of acquisition as good images and to those with annotations corresponding to bad quality of acquisition as bad images.

After consulting with experts *technicians/cardiologists?* and by looking closer at the images and corresponding annotations, we determined that in good regions a clear range between the upper and lower trace of the LVID can be expected while the heart is in its systole phase. An example of this can be seen in Figure 1a, while an example of a bad image is shown for reasons of comparison



(a) train image good (b) train image bad

Figure 1: Example of two images taken from good and bad quality of acquisition regions of the same echocardiogram.

in Figure 1b. Moreover, technicians explained that during manual annotation of the LVID they define a region as good if it includes at least two *or three?* such heart beats. We assumed that the variation in the images' lengths would make it harder for the network to classify an image as good or bad according to the above-mentioned points *following the same reasoning made by humans?*. Keeping this in mind and wanting the classification network to learn in the same way technicians have learned over the years, we further cropped our data by using a sliding window approach with a fixed *set?* step. By focusing mainly on the interior of the left ventricle and ensuring at least two heart beats per image, we cropped squared images, consisting of approximately 3 heart beats, from the good and bad regions. In this way, for the task of image classification we created a dataset of 850 images (374 good and 476 bad) for train and validation and 307 images (209 good and 98 bad) for testing.

For the dataset required for the image segmentation task we applied a slightly different approach. Here, we want the network to focus on the entire height of the image, rather than just the inner dimension of the left ventricle, as it is exactly this which we wish to segment. We therefore created a dataset of square images where the entire height of the original image is included, while again applying a sliding window approach for long acquisitions. We must note here that only the good acquisition regions were used for creating this dataset. We believe that it is more meaningful for the network to learn to segment only images that truly represent the heart state, since ultimately it is from such regions we wish to extract features. We used the annotated upper and lower traces in the good images to create binary segmentation masks which were used as ground truth during training and testing. In this way, a dataset of 457 images was created for training and validation of the segmentation network and 46 images for testing.

While all the above steps were performed as a preprocessing step to create

two datasets for training classification and segmentation networks, this reasoning also provided the basis for the ATEFEX framework design as can be seen Section 2.3.

2.3 ATEFEX Framework Design

The goal of this work is to create a framework for automatic and accurate feature extraction from M-mode echocardiography mouse data. The framework was designed to accept a single or multiple echocardiograms as input. These are then processed as explained in Section 2.2, resulting in the creation of two mini-batches of images for each acquisition. The sole difference here is that we do not use a sliding window approach but rather crop the images successively. The first batch is fed successively into a trained network which classifies images as good or bad, while the second is fed into another trained network which outputs a segmentation mask of the left ventricle inner diameter. The two models used are described in Section 2.3.1, while the training process for both is outlined in Section 2.3.2. The results from the mini-batches are then concatenated, thus a segmentation mask for the entire acquisition is created as well as good and bad labels for regions of the acquisition. The length of these regions depends on the mouse weight *weight or body mass?*, namely the larger the mouse the longer the regions. The segmentation mask is used to extract features, while the quality labels to determine whether these features should be taken into account. The entire framework along with its input and output is depicted in Figure 2.

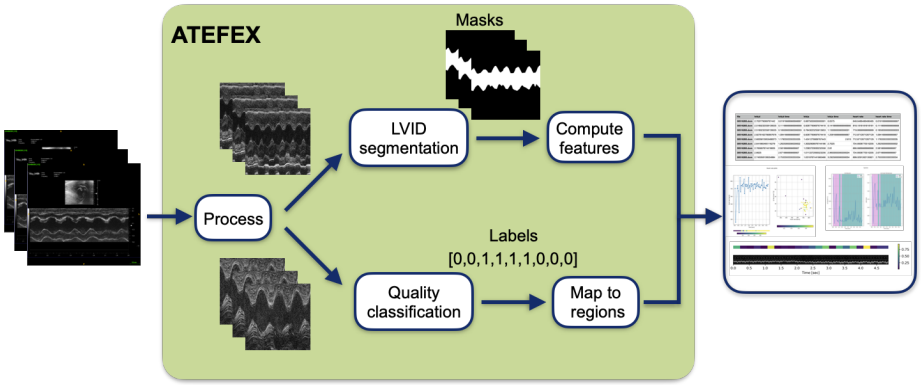


Figure 2: The main steps of the ATEFEX framework.

For extracting features from the echocardiogram the first step is to measure the LVID for each time instance. This corresponds to the number of pixels segmented as inner left ventricle diameter for each column in the image, after mapping it to real world values, in our case mm. Next, by finding local minima and maxima of the LVID measurements we obtain all occurrences *measurements again?* in the systole and diastole phase accordingly, namely LVID;d and LVID;s. The time of their incidence is also computed and stored. With the

LVID measurements we can then easily also compute the Left Ventricle Volume with use of the Teichholz formula [2]:

$$LVVol = \frac{7}{2.4 + LVID} * LVID^3 \tag{1}$$

Finally, by calculating the distance between successive LVID;d measurements we compute the heart rate for each heart beat through the following equation:

$$Heartrate_i = \frac{60}{LVID; d_{i+1} - LVID; d_i} \tag{2}$$

In total for each heart beat we extract a total of five features:

- LVID;d and LVID;s in [mm]
- LV Vol;d and LV Vol;s in [mm^3]
- Heart rate [bpm]

It becomes clear that in this way multiple measurements for each feature are obtained from every acquisition. By only including regions of good acquisition quality we observed an average of 37 measurements per acquisition, as opposed to three measurements that are acquired by manual annotation. This allows us to derive measures such as mean, median, variance etc. for the above-mentioned features and paves the way for further statistical analyses.

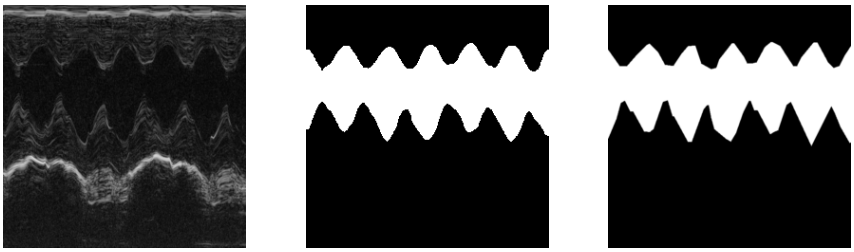
2.3.1 Network architectures

For the task of classifying images according to the quality of acquisition we created a simple classification network. It includes five convolutional blocks (Convolution → ReLU → BatchNorm → MaxPooling) followed by a fully connected layer and a sigmoid function. The convolutional filters and max pooling layers have a kernel size of three and two accordingly. The network is fed images of size 256x256 and outputs a value between 0-1 which is then rounded to get the classification result. The training and evaluation was run a total of ten times and the average results were obtained.

For the segmentation task the QuickNAT model architecture was used [3], with the final layer adjusted for a binary output. QuickNAT follows and extends the U-Net architecture [4], with additional internal skip connections included in each dense block. It consists of four such dense blocks in the encoder and decoder and accepts inputs of any size. It was chosen here due to its inference speed and proven high performance in medical imaging segmentation tasks.

2.3.2 Training the networks

Both the classification and segmentation networks were trained for a total of 20 epochs with a batch size of four and a learning rate of 1e-04. A train-validation split of 90-10% was used.



(a) Input image (b) Prediction (c) Ground truth

Figure 3: Example of segmentation result on test set.

For both models, the binary cross entropy loss was used during training along with Adam as the optimizer [5]. The classification network's performance was evaluated using the accuracy, F1 score, False Positive Rate (FPR) and False Negative Rate (FNR). The segmentation network was evaluated using the Dice Score and the Mean Square Error (MSE). The average performance of both networks on their corresponding test sets can be seen in Tables 1 and 2, while Figure 3 shows an example result of the segmentation network on the test set. As can be seen here, the ground truths masks created from the annotations have unavoidable rough edges caused from interpolating the annotated points (Figure 3c). The network, on the other hand, outputs a much smoother curve (Figure 3b) suggesting that it correctly detects and classifies pixels belonging to the inner diameter of the left ventricle.

accuracy	F1 score	FPR	FNR
96.428	0.976	0.118	0.007

Table 1: Quantitative results of classification network on test set

Dice Score	MSE
0.95	0.02

Table 2: Quantitative results of segmentation network on test set

3 Results

For extensive evaluation of the framework we tested its performance with mice from four different experiments, amounting to a total of 44 mice. *more details on test datasets needs to be added here.* To investigate the robustness of the method we included experiments of mice of various weights and ages, ranging from 10-130 gr and 6-78 weeks accordingly. Manual annotations were provided of the LVID;d, LVID;s and heart rate. The full end to end method takes about

13 seconds to run on a GPU and about 18 seconds to run on a CPU for one echocardiogram. Figure 4 shows an example of the visual outputs of the ATEFEX framework.

From figures 4a, 4b, 4c it becomes clear that in bad quality acquisition regions we have high rises and drops of the measurements. These clearly should not be taken into account in future analyses. This is also clear in 4d, where rather then the binary classification outputs we have included the sigmoid outputs of the network and mapped them to colours. The mapping is explained by the corresponding heatmap, where yellow points correspond to measurements from good quality regions, while purple points correspond to points from bad regions. We see from the graph representing the LV Vol;d over the heart rate that all occurring outliers correspond to measurements in bad acquisition quality regions. This can be explained for example by a sudden increase in the heart rate resulting from the mouse moving, or an imprecise segmentation mask since the network was trained only on good quality regions. In the good regions we see a more reasonable range of measurements, but nevertheless some fluctuation which suggests a variability of data which impossible to detect previously with only three continuous manual annotations per acquisition.

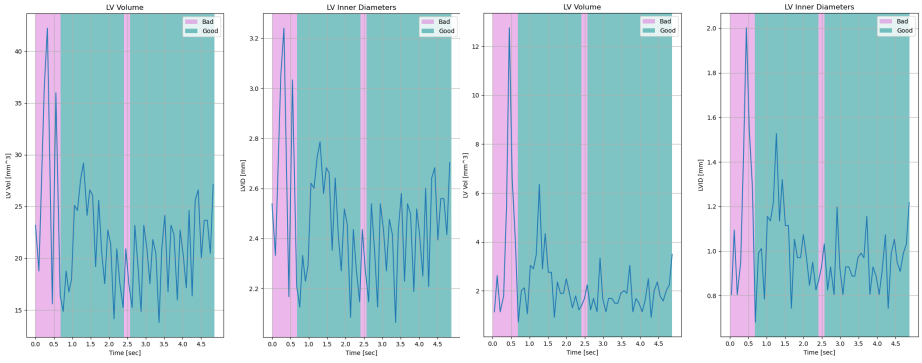
In Figure 5 we compare the manually extracted features with those computed by the ATEFEX framework. Since for each mouse only one manual feature is provided we compared these with the median of all features in good classified regions. As can be seen in the plots for some acquisitions the predicted value is zero, which indicated that all regions in the echocardiogram were classified as bad. We see here a case of our method failing. However, all these acquisitions correspond to mice taken from the *Dummersdorfer* experiment and have an average weight of 130 grams. This is considered an extreme case which very rarely occurs in mouse screening procedures and therefore does not concern us overmuch. *I say the same more or less in results-repetition?* For the rest of the points we see that they follow the ramp function, showing than the predicted values are closely related to their equivalent manual annotations.

After excluding the above mentioned failure cases we calculated the Mean Absolute Error (MAE) between all manually extracted features and the median of the automatically extracted features. The quantitative results can be seen in Table 3.

	LVID;d [mm]	LVID;s [mm]	Heart Rate [bpm]
MAE	0.233	0.256	56.76

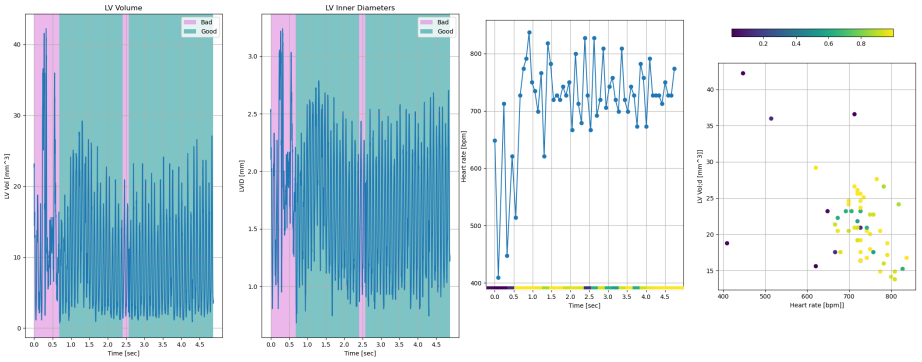
Table 3: MAE between the manual annotations and the median of automatically calculated features in good regions

As discussed, one of the main contributions of the ATEFEX framework is its ability to produce multiple observations where previously only one measurement per feature in an acquisition was extracted. This allows us to now examine a distribution rather than a single point for each feature. In Figure 6 we show boxplots of the total good LVID;d measurements for ten acquisitions of the



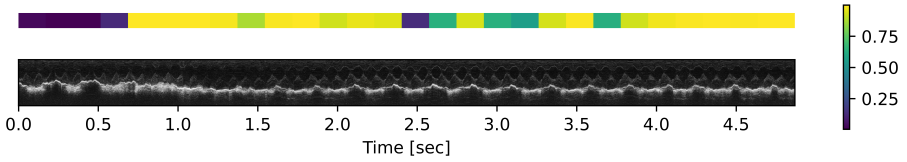
(a) LVID and LV Vol, in diastole

(b) LVID and LV Vol, in systole

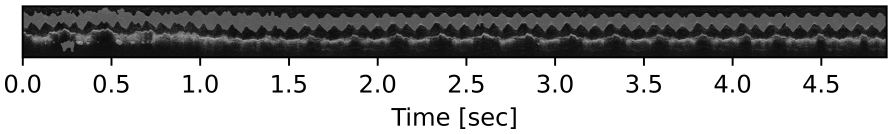


(c) LVID and LV Vol, for each time instance

(d) Heart rate plots

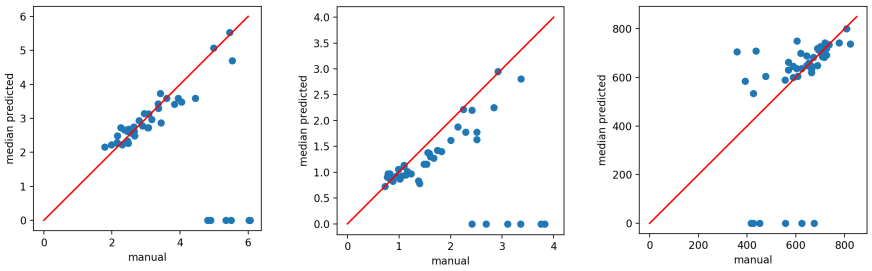


(e) Image of entire acquisition with colors above showing the quality of acquisition



(f) Image and mask of the entire acquisition superimposed

Figure 4: Example of the graphs and images generated by the ATEFEX framework.



(a) LVID;d (b) LVID;s (c) HR

Figure 5: Comparison of manual annotations with the median of automatic features for LVID;d, LVID;s and heart rate.

Golt;1b experiment. The manual annotations are also imposed on each plot to observe whether it falls in the range of automatically calculated measurements, while the boxes show the quartiles of the data. We see only one case where the manual annotation falls outside the data distribution; here, the manual annotation suggests we are dealing with a rather larger heart than normal and which is likely the reason the method is making an underestimation.

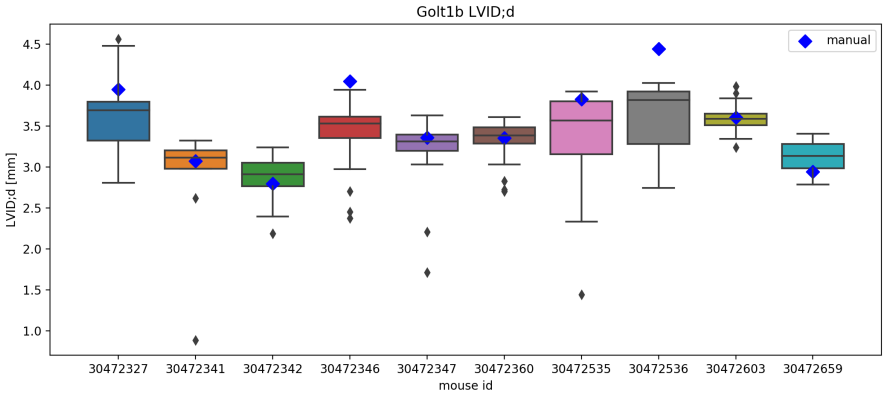
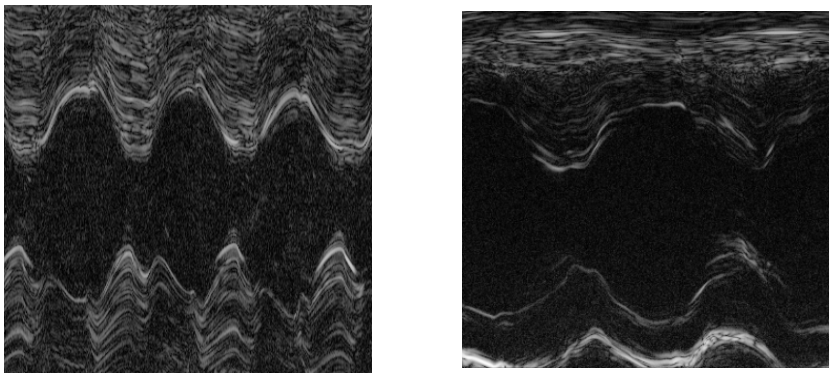


Figure 6: Boxplot of values

4 Discussion

The fact that ATEFEX utilizes two trained networks for automatic feature extraction suggests some limitations. Firstly, both networks were trained on a set of mice belonging to a ... experiment. This means that it is not entirely robust to mouse size. Indeed, this became evident when we tested the framework with the *Dummersdorfer* mice, as previously mentioned in the explanation of 5.



(a) train image good (b) Failure case

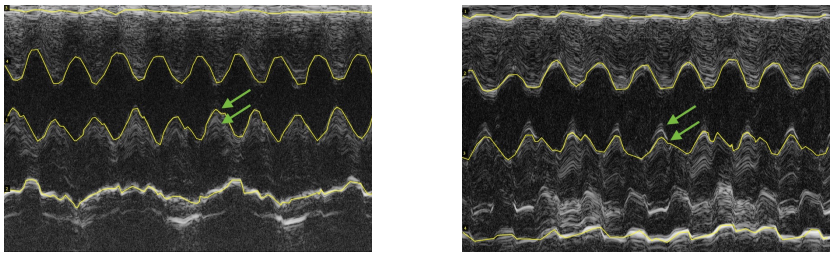
Figure 7: Comparing a good training example with a failure case.

Their body mass is much larger than usual which also corresponds to a much larger heart and slower heart rate. As an example, in Figure 7 we can see two inputs into the classification network, for a train sample and the extreme case of the *Dummersdorfer* mice.

The networks were not trained with such images and it becomes apparent during evaluation that they are not able to perform in these cases. Indeed for the *Dummersdorfer* mice in six out of ten cases the entire echocardiograms were classified as bad (5). We do, however, consider this as an extreme case which will very rarely occur in our screening processes and therefore do not concern ourselves overmuch with these results.

Another point worth mentioning here is that in our estimations of the LVID;s we are in most cases underestimating the annotated value. This becomes also clear by observing Figure 5b, where it is apparent that most points are slightly bellow the ramp function, indicating that the predicted value is smaller than the manual annotation. We believe this is due to the fact that the network is fitting the boundary of the lower trace based on the change of intensity. It was, however, a point of discussion between our technicians whether this point should actually be the boundary of the inner diameter or the second slightly less-intense trace below that. We understand this to be a matter of some discussion with the cardiology community. Indeed, as can be seen in Figure 8 some images were annotated with one convention while others with the other and the decision on where the lower bound should be placed varies from technician to technician. Currently, the network is segmenting all images with the boundary placed at the first, or upper, trace, but if we do decide to go for a different convention in the future we believe we could easily extend the framework and enforce this constraint in future work.

The framework has been developed to extract five features from each echocardiogram. A point for future work, would be to extend this to also extract additional features. For some features, such as the respiratory rate, derivations from



(a) Annotation of first lower trace (b) Annotation of second lower trace

Figure 8: An example of how the annotations change according to technician.

the already computed LVIDs are required, while for others the process would be rather more complex. For example, for the estimation of the Left Ventricle Posterior Wall (LVPW) the segmentation network would need to be extended from a binary segmentation method to a multi-class segmentation approach. In any case, the ATEFEX framework provides a baseline for extracting multiple features in a fully automatic, unbiased and efficient way.

References

[1] DAVID J Sahn, ANTHONY DeMaria, JOSEPH Kisslo, and A ft Weyman. Recommendations regarding quantitation in m-mode echocardiography: results of a survey of echocardiographic measurements. *circulation*, 58(6):1072–1083, 1978.

[2] Garima Arora, Alexander M Morss, Gregory Piazza, Jason W Ryan, Danya L Dinwoodey, Neil M Rofsky, Warren J Manning, and Michael L Chuang. Differences in left ventricular ejection fraction using teichholz formula and volumetric methods by cmr: implications for patient stratification and selection of therapy. *Journal of Cardiovascular Magnetic Resonance*, 12(1):1–2, 2010.

[3] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, Christian Wachinger, Alzheimer's Disease Neuroimaging Initiative, et al. Quicknat: A fully convolutional network for quick and accurate segmentation of neuroanatomy. *NeuroImage*, 186:713–727, 2019.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.