Segmentation and Quantification of the Left Ventricle to Assess the Ventricular Remodeling post Myocardial Infarction

A Technical Report submitted to the Department of Biomedical Engineering

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Meenoti Thakore

Spring, 2023

Technical Project Team Members

Rose Eluvathingal Muttikkal

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

John A. Hossack, Department of Biomedical Engineering

Yanjun Xie, Department of Biomedical Engineering

Yi Huang, Department of Biomedical Engineering

Segmentation and Quantification of the Left Ventricle to Assess the Ventricular Remodeling post Myocardial Infarction

Rose Eluvathingal Muttikkal
Meenoti Thakore
Team 32

Word Count:
Number of Figures and Tables: 8
Number of Equations: 2
Number of Supplements: 4
Number of References: 20

Signed ____ Date _______________
Dr. John Hossack

John A. Hossack Digitally signed by John A. Hossack Date: 2023.05.07 12:14:56 -04'00'

Approved ____ Date _______________
Dr. John Hossack, Department of Biomedical Engineering

Segmentation and Quantification of the Left Ventricle to Assess the Ventricular Remodeling post Myocardial Infarction

Meenoti Thakore[a,1,] Rose Eluvathingal Muttikkal[b], John Hossack[c], Yanjun Xie[d], Yi Haung[e]

[a] Fourth Year Biomedical Engineering Undergraduate, University of Virginia
[b] Fourth Year Biomedical Engineering Undergraduate, University of Virginia
[c] Department of Biomedical Engineering, University of Virginia
[d] Department of Biomedical Engineering, University of Virginia
[e] Department of Biomedical Engineering, University of Virginia
[1]Correspondence: mt5wjv@virginia.edu, 571-269-0408

Abstract

Automatic cardiac image segmentation has the potential to extract information from large amounts of medical image data. Previously developed methods using deep-learning models have been used to classify heart failure based on ejection fraction, but these have a significant error rate and do not use all relevant physiological metrics. Thus, this project aims to develop a deep learning model for left ventricle (LV) image segmentation. The model will be optimized and applied to quantify heart dysfunction post myocardial infarction (MI) in mice. Previously collected mice ultrasound images were transformed to meet model requirements. U-net architecture with Pytorch framework was used for development, and this process included data loading, training, optimization, and testing to make predictions. Accuracy, area under the curve, and F1 score was calculated to assess model performance. The results show an average .90 accuracy, 0.83 F1 score, and 0.96 area under the curve. Model and researcher segmented images were utilized to calculate area and volume and quantify the LV. After comparison, statistically significant differences were observed between model and researcher segmented images which can be attributed to interobserver variability in annotations. Compared to previously developed segmentation models for human echocardiography, our model performed moderately well. The human model had a F1 score of 92% compared to our 83%. The developed model may be applicable for researchers conducting image segmentation on mice echocardiograms and the segmentation can be used to quantify other metrics of heart function quickly and accurately for classification.

Keywords: Segmentation, Deep Learning, Murine Echocardiography

Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally according to the World Health Organization. In the United States, a heart attack occurs every 40 seconds, with about 1.5 million heart attacks and strokes occurring every year[1],[2]. Imaging techniques such as magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound are widely used to non-invasively assess cardiac structures and functions[3]. Cardiac image segmentation is an important step in image analysis and involves the partitioning of image data into specific regions of interest[3]. Numerous image segmentation methods have been developed in the past, including manual (slice by slice) and semi-automatic segmentation[4]. Manual segmentation is often performed by a radiologist or specialized clinician annotating the region of interest in a slice-by-slice manner. Although expert knowledge is utilized, this method is very time consuming, labor intensive, and prone to intra and interobserver variability[5]. Manual cardiac segmentation is challenging since the structural characteristics of the ventricles make it difficult to segment compared to other organs, such as the liver or kidney[6]. Semi-automatic segmentation uses algorithms to assist the process and eliminates the need for slice-by-slice segmentation[7]. Although the time and effort required from the user can be reduced, semi-automatic segmentation is still highly variable[5].

Automated cardiac image segmentation using deep learning is becoming increasingly prevalent because it is more reliable, accurate, and quicker compared to previously used methods. Previously developed deep learning models have achieved dice similarity coefficients between 82-92% when segmenting the LV[8]. This represents an 82-92% overlap between automatic and manual segmentation results. Certain deep learning models have also been developed to classify heart failure in humans based on ejection fraction[9]. The proposed project will develop a deep learning model to automate image segmentation of the LV from murine echocardiography videos. Analyzing the LV using 2D echocardiographic images is a common medical procedure for patients with cardiac issues and is used to distinguish between diseased and non-diseased states. Echocardiography is the gold standard in diagnostic imaging of the heart since it is a non-invasive and low-cost technique[6]. Thus, the analysis of murine echocardiography videos prior to and after induction of a MI may be relevant to clinically used imaging modalities. Ultimately, the proposed project is seeking to overcome current limitations in the robustness of deep learning-based models for LV quantification. To this end, we propose to develop a deep learning model for automated image segmentation of left ventricle heart ultrasound images (Aim 1). We will then optimize and apply the model to quantify heart dysfunction post myocardial MI in mice (Aim 2).

Results

Model Performance

This project aimed to develop a deep learning model for LV image segmentation. After development, the model was assessed both quantitatively and qualitatively. Figure 1 shows the model segmentation output, comparing a model prediction output to the segmentation target[10]. The prediction output indicates that the model learned the features of the LV.

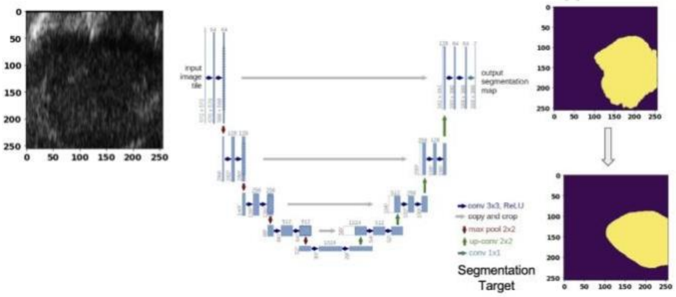

Fig. 1. Model Segmentation Output Results. The input (top left) was put into the model and the prediction output (top right) was compared to the corresponding target image (bottom right).

Accuracy, F1 score, and area under the curve (AUC) metrics were calculated to evaluate performance. 356 images from the validation dataset were used for analysis as shown in Table 1 below.

Table. 1. Validation data was run through the model and performance metrics comparing the model predictions and target images were calculated for the entire dataset.

Metric # of Images	Accuracy	F1 Score	Area Under the Curve
356	0.8971	0.8313	0.9582

Training and validation loss values were also found after each epoch and average values are presented in Figure 2. Validation loss decreased by 40% to 0.297 and training loss decreased by 20% to 0.076. Learning rate was also calculated, and it stayed constant at 0.1 throughout training as seen in Figure S1.

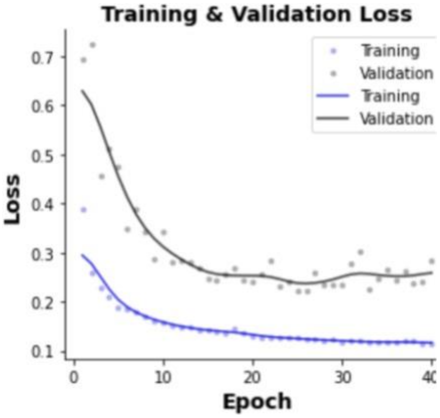

Fig. 2. Training and validation loss curve after training.

LV Quantification

This project aimed to quantify and classify heart dysfunction post MI in mice. LV area and volume metrics were calculated for model and researcher segmented images to identify differences between the LV pre and post MI. Average LV area for 3 baseline videos and 3 post iNOS videos are shown in Table 2.

Table. 2. Comparison of average LV area from researcher and model segmented images from 3 baseline and 3 post iNOS videos.

Metric Image Type	Model Segmented Area (mm^2)	Researcher Segmented Area (mm^2)
Baseline	8.889 +/- 5.255	7.030 +/- 3.127
28 Days Post iNOS	20.976 +/- 4.902	11.216 +/- 2.365

Average LV volume for 6 videos is shown in Table 3.

Table. 3. Comparison of average LV volume pre and post MI for 6 videos.

Metric Image Type	Model Segmented Volume (μL)	Researcher Segmented Volume (μL)
Baseline 1	139.347 +/- 126.894	49.710 +/- 38.393
Baseline 2	142.537 +/- 122.560	56.831 +/- 44.600
Baseline 3	84.650 +/- 86.576	97.857 +/- 64.655
28 Days Post MI 1	474.402 +/- 223.250	188.030 +/- 59.666
28 Days Post MI 2	618.218 +/- 222.202	185.447 +/- 64.838
28 Days Post MI 3	629.946 +/- 318.860	171.945 +/- 49.791

Model versus researcher segmented values for all area and volume measurements were found to be significantly different based on t-tests run, with p-values less than 0.05 as shown in Table S1. The null hypothesis, which stated that the researcher and model values for area and volume were similar, was rejected. Statistically significant differences were also found between all area and volume measurements for baseline versus iNOS mice. This leads to the rejection of the null hypothesis, that baseline and iNOS mice would have similar area and volume measurements. P-values are all less than 0.05 and are presented in Table S2. Since model and researcher segmented images were based on different annotation labels, the differences between annotations were calculated to find interobserver variability. Differences between researcher and model-trained annotations were significant for both baseline and iNOS mice, with p-values less than 0.05. P-values are shown in Table S3. Due to time constraints and unexpected setbacks, the rest of aim 2, classification of heart dysfunction post MI, was not completed.

Discussion

Based on the two aims presented at the beginning of this project, Aim 1's results demonstrated that our developed model performed moderately well compared to previously developed segmentation models for human echocardiography[9]. Table 4 compares F1 score, accuracy, and AUC for the developed mice model and existing human model. The differences most notably for F1 score and accuracy may be attributed to the size of the training dataset. The human model utilized a large dataset of 10,030 annotated echocardiogram videos while the mice model used 2841 images for training. Larger training datasets allow the model to identify correlations and trends more effectively than a smaller sample[9]. Future work should include annotating the remaining mice ultrasound videos to use for training and validation. In addition, further optimizations such as modifying the loss function or data augmentation could be implemented to improve model performance.

Table. 4. Comparison of model performance metrics between an existing human model and the developed mice model.

	Accuracy	F1 Score	Area Under the Curve
Human Model	0.96	0.92	0.96
Mice Model	0.90	0.83	0.96

Based on the results seen in Figure 3 the developed model demonstrated no signs of overfitting or underfitting. Overfitting occurs when the model gives accurate predictions for the training data, but not for new data. This characteristic is seen when the validation loss curve increases during the end of training[11]. Underfitting occurs when the model cannot accurately predict both training and new data. This characteristic is represented on the graph when the validation curve does not decrease throughout training[11]. As seen in the training and validation loss curve, overfitting and underfitting is not present, resulting in a model that is able to learn LV features. It can also be observed from this graph that training loss is lower than validation loss. This behavior is expected since the model is learning from the training data and should be able to predict it better than the validation set[12].

Figure S1 shows that the learning rate stayed constant throughout training. The learning rate was set to decrease if validation loss stayed constant for 10 epochs, however this was not observed in the model. If this was observed, the learning rate would increase resulting in faster training once loss is low and stable. Further optimizations could be made to train the model for more than 40 epochs to achieve stable validation loss values and learning rate decay.

Aim 2's results demonstrated that LV area and volume based on model and researcher segmentation were significantly different, when they were expected to be similar. These values were used to compare the model segmentation output to images not trained on the model. The metrics were calculated based on the assumption that researcher segmented area and volume is the expected value. In this case, the researcher refers to our capstone graduate advisors. These differences were due to significant interobserver variability in annotations since the model was trained on annotations we made. Overall, the differences between researcher and model segmented area and volume may not correlate with model accuracy. The discrepancy could be attributed to differences in annotations and the low sample utilized for the calculations. Taking the average across each condition may have also contributed to the large differences. Volume measurements may be more accurate if end diastolic and end diastolic values are calculated based on the smallest and largest area instead of average area. These discrepancies suggest that researcher annotated segmentations are better suited for model training. A model with increased accuracy may also reduce differences in the results presented.

Calculating metrics of LV area and volume can also be used to classify images pre and post MI. Based on the results, both of these metrics were significantly larger post MI since an enlarged heart can be caused by damage to the heart muscle. The iNOS knockout in mice post MI slows down the healing process which contributes to the larger area and volume values. This trend is evident in both Tables 2 and 3. There is no standard range for the area and volume of a mice LV, however the calculated values can be used to compare between the same mouse. Future work should include defining thresholds for the LV area and volume to classify ultrasound images into pre or post MI groups.

Currently, this model may be applicable for researchers to conduct image segmentation on mice echocardiograms. Future work should involve optimizing model performance to improve accuracy. The segmentation should also be used to quantify other metrics of heart dysfunction such as ejection fraction, wall thickness, and cardiac contractility. These measurements can be utilized for classification of heart dysfunction. Then a GUI software should be developed to improve the accessibility of the model. Lastly, the viability of this model for human echocardiography should be explored in order to apply the developed model in a clinical setting.

Materials and Methods

This project involved various steps that were completed in parallel to develop a deep learning model for LV segmentation and quantification as laid out in the two aims previously mentioned. Each step is described in the following sections and laid out in Figure 3 below.

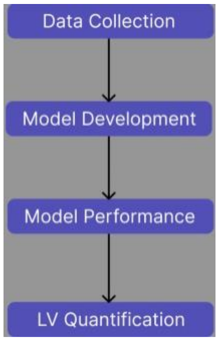

Fig. 3. High-level overview of the steps taken to achieve our project aims.

Data Collection

2D B-mode videos of murine echocardiography were previously collected by Dr. Brent French's lab for baseline and iNOS knockout mice pre and post MI. The baseline mice were treated as the healthy control group, while iNOS mice were the experimental group. The iNOS gene knockout negatively impacts and slows down the healing process post MI, resulting in larger differences in LV characteristics after a heart attack. The data was collected with a Vevo 2100 scanner using a MS 400 transducer and center frequency of 30 MHz and exported in DICOM format. There were 5 mice at 5 different imaging sessions: baseline (pre-MI), 7-, 14-, 21-, and 28-days post MI. Each video had approximately 120 frames.

Model Development

Prior to working with the mice ultrasound data, a framework for the model was developed using a publicly available dataset of liver CT images. A pre-trained U-Net model for abnormality segmentation of brain MRI volumes was utilized in Pytorch for training. The pre-trained model requires 3 input channels, 1 output channel, and 32 features in the first layer. The U-Net is a convolution neural network architecture which is commonly used for fast and precise segmentation of images[13]. While the framework was being developed, data preparation of the murine echocardiography videos was also completed. This involved manual ground truth labeling of the videos to identify the LV using 3D slicer software. Ground truth labeling refers to the actual model target or segmentation maps and can be used as a comparison to model predictions made on unlabeled, input data[14].

Transformation and Data Loading

The videos and corresponding labels were then loaded into the model using a data loader, which batches the data into input-target pairs as seen in Figure 4, and performs transformations. This included resizing and reshaping the data into tensors to meet model requirements. A batch size of 16 was utilized for training as models trained with small batch sizes generalize well on the validation dataset used for testing. However, larger batch sizes take less time to train but are less accurate, thus 16 provided higher accuracy with reasonable training time[15]. Augmentations including rotating the training images were implemented in the data loader for more robust training. The probability of a rotated image was p = 0.5 which eventually resulted in batches with rotated input-target pairs[15].

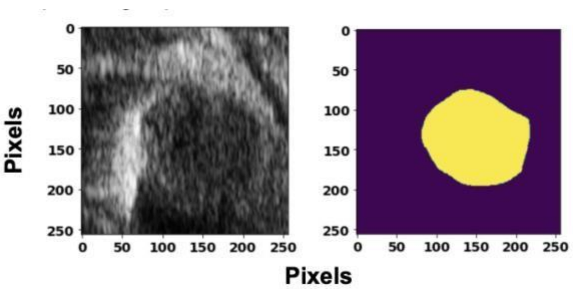

Fig. 4. Input (left) and target (right) pairs were transformed to meet model requirements and loaded for training.

Training

The loaded data was split 80:20 for training and validation respectively which resulted in 2841 images for training and 356 images for validation. The model was trained for 40 epochs using Rivanna, UVA's high performance computing system. A total of 16 videos of a mouse taken at baseline (14 B-mode from the short axis and 2 B-mode from the long axis) and a total of 12 videos of 4 mice taken 28 days post MI were used for training. The implemented train function iterated through the training data loader and sent the batches through the network. The output, along with the corresponding target, was used to compute the loss using a Binary Cross Entropy (BCE) loss function. BCE tracks incorrect labeling of the data which contributes to increased accuracy[16]. Based on the computed gradients, a backward pass and step with the stochastic gradient descent (SGD) optimizer is performed to update the model parameters. Learning rate decay was also implemented as an optimization technique. The initial learning rate was kept as the default, 0.1, which means that the weights in the network are by 0.1 * estimated weight error. 0.1 was picked since a very low learning rate results in slow training, while a high learning rate causes divergent behavior in the loss function[17]. The learning rate scheduler in the model increased the learning rate by 0.01 if the validation loss stays constant for 10 epochs. This allows for faster training without affecting loss values.

Testing

Input images from the validation data were run through the model to assess overall performance qualitatively and quantitatively. This resulted in predictions that can be compared to the corresponding targets from the validation dataset. To quantitatively assess performance, accuracy, F1 score, and area under the curve (AUC) metrics were calculated using the torchmetrics function to assess model performance. Accuracy refers to the images correctly predicted across the entire dataset, F1 score measures accuracy through a combination of precision and recall, and AUC calculates the probability of making a correct prediction[18].

LV Quantification

Area and volume metrics on model and researcher segmented images were used to quantify the LV. 3 baseline and 3 iNOS videos were separately loaded into the model and prediction images were found using the trained model. The model segmented area was calculated by multiplying the sum of all the elements in the input image tensor by the area of one pixel. The area of one pixel was about 0.001 mm^2 for videos taken in all imaging sessions. The average area across all slices of all 3 videos for baseline and iNOS was separately calculated for the final areas based on model segmentation. A similar process was used to find the areas for researcher segmented images. Our capstone advisors provided us with segmentation for the same 3 baseline and iNOS videos. The same formula was utilized to find the researcher's segmented areas.

LV volume was calculated based on the area values found for each video slice from the process described previously. LV volume was calculated using Equation 1 shown below, where A is the LV area, and L is the long axis length measured from the LV apex to mitral valve[19]. The long axis lengths were measured for each video by taking the average of about 20 frames per video around the minimum and maximum areas, which represent end diastolic and end systolic volume. The long axis lengths ranged from about 6.73 mm to 7.58 mm depending on the imaging session. The values of volume per video slice were average for each video, resulting in the table previously presented.

$$V = \frac{\frac{8\pi}{3} x A^2}{L} \qquad [1]$$

Statistical analysis was done to compare between area and volume values for model and researcher segmented images. A two tailed type 1 t-test was used to determine if there were no significant differences between the values. Additional analysis was performed to compare area and volume measurements pre and post MI. A two tailed type 1 t-test was run to determine if there were significant differences between the values. To account for the large differences observed between model and researcher segmented image, interobserver variability was calculated. This refers to the difference in annotations between us and our capstone advisors. Interobserver variability was calculated using Equation 2, where A is student annotation and B is researcher annotation[20]. A two tailed type 1 t-test was run to determine if differences between the two annotations were significant.

$$|A - B| \qquad [2]$$

End Matter

Author Contributions and Notes

J.H., Y.X. and Y.H. designed research, R.E.M. and M.T. performed research, R.E.M. and M.T. wrote model software, R.E.M. and M.T. analyzed data; and R.E.M. and M.T. wrote the paper.

The authors declare no conflict of interest.

Acknowledgments
We would like to acknowledge Dr. Brent French's lab for collecting the murine echocardiography data.

References

1. CDC. (2022, July 15). Heart Disease Facts | cdc.gov. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/facts.htm

2. CDC. (2021, February 26). Million Hearts® Costs & Consequences. Centers for Disease Control and Prevention. https://millionhearts.hhs.gov/learn-prevent/cost-consequences.html

3. Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., & Rueckert, D. (2020). Deep Learning for Cardiac Image Segmentation: A Review. Frontiers in Cardiovascular Medicine, 7. https://www.frontiersin.org/articles/10.3389/fcvm.2020.00025

4. Fasihi, M. S., & Mikhael, W. B. (2016). Overview of Current Biomedical Image Segmentation Methods. 2016 International Conference on Computational Science and Computational Intelligence (CSCI), 803–808. https://doi.org/10.1109/CSCI.2016.0156

5. Manual Segmentation—An overview | ScienceDirect Topics. (n.d.). Retrieved November 11, 2022, from https://www.sciencedirect.com/topics/computer-science/manual-segmentation.

6. Shoaib, M. A., Lai, K. W., Chuah, J. H., Hum, Y. C., Ali, R., Dhanalakshmi, S., Wang, H., & Wu, X. (2022). Comparative studies of deep learning segmentation models for left ventricle segmentation. Frontiers in Public Health, 10. https://www.frontiersin.org/articles/10.3389/fpubh.2022.981019

7. Kim, Y. J., Lee, S. H., Park, C. M., & Kim, K. G. (2016). Evaluation of Semi-automatic Segmentation Methods for Persistent Ground Glass Nodules on Thin-Section CT scans. Healthcare Informatics Research, 22(4), 305–315. https://doi.org/10.4258/hir.2016.22.4.305

8. Hsu, W.-Y. (2019). Automatic Left Ventricle Recognition, Segmentation and Tracking in Cardiac Ultrasound Image Sequences. IEEE Access, 7, 140524–140533. https://doi.org/10.1109/ACCESS.2019.2920957

9. Ouyang, D., He, B., Ghorbani, A. et al. Video-based AI for beat-to-beat assessment of cardiac function.Nature 580, 252–256 (2020). https://doi.org/10.1038/s41586-020-2145-8

10. Milesial. (2023). U-Net: Semantic segmentation with PyTorch [Python]. https://github.com/milesial/Pytorch-UNet (Original work published 2017)

11. Koehrsen, W. (2018, January 28). Overfitting vs. Underfitting: A Complete Example. Medium. https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765

12. Baeldung. (2022, February 19). Training and Validation Loss in Deep Learning | Baeldung on Computer Science. https://www.baeldung.com/cs/training-validation-loss-deep-learning

13. PyTorch. (n.d.). Retrieved May 4, 2023, from https://www.pytorch.org.

14. What is Ground Truth in Machine Learning? | Domino Data Lab. (n.d.). Retrieved May 4, 2023, from https://www.dominodatalab.com/data-science-dictionary/ground-truth.

15. Weights & Biases. (n.d.). W&B. Retrieved May 4, 2023, from https://wandb.ai/ayush-thakur/dl-question-bank/reports/What-s-the-Optimal-Batch-Size-to-Train-a-Neural-Network---VmlldzoyMDkyNDU.

16. Ho, Y., & Wookey, S. (2020). The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling. IEEE Access, 8, 4806–4813. https://doi.org/10.1109/ACCESS.2019.2962617

17. Haswani, V. (2021, May 30). Learning Rate Decay and methods in Deep Learning. Analytics Vidhya.https://medium.com/analytics-vidhya/learning-rate-decay-and-methods-in-deep-learning-2cee564f910b

18. Huilgol, P. (2019, August 24). Accuracy vs. F1-Score. Analytics Vidhya. https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2

19. Duan, C., Montgomery, M. K., Chen, X., Ullas, S., Stansfield, J., McElhanon, K., & Hirenallur-Shanthappa, D. (2022). Fully automated mouse echocardiography analysis using deep convolutional neural networks. American Journal of Physiology-Heart and Circulatory Physiology, 323(4), H628–H639. https://doi.org/10.1152/ajpheart.00208.2022

20. Popović, Z. B., & Thomas, J. D. (2017). Assessing observer variability: A user's guide. *Cardiovascular Diagnosis and Therapy*, *7*(3), 317–324. https://doi.org/10.21037/cdt.2017.03.12

Supplemental Information

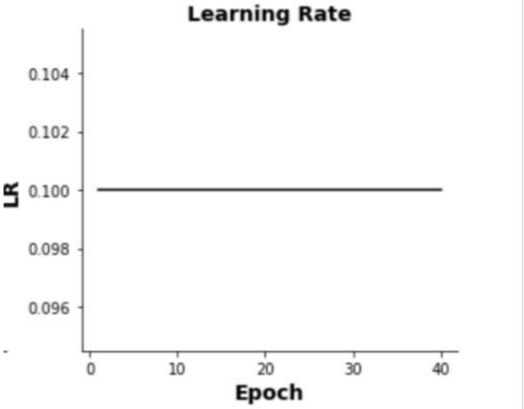

Fig. S1. Learning rate graph after training was run for 40 epochs. The value stayed constant at 0.1.

Table. S1. P-values were less than 0.05 signifying that the difference between researcher and model segmented areas are significant.

Baseline Area T-Test	5.92E-19
Segmented Area T-Test	5.23E-116

Table. S2. P-values were less than 0.05 signifying that the difference between researcher and model segmented volumes are significant..

Baseline 1 T-Test	2.07E-18
Baseline 2 T-Test	1.89E-19
Baseline 3 T-Test	1.80E-02
Segmented 1 T-Test	3.95E-25
Segmented 2 T-Test	5.28E-40

Segmented 3 T-Test	2.41E-35

Table. S3. P-values were less than 0.05 signifying that the difference between researcher and model-trained annotations are significant.

Baseline T-Test	6.47E-41
Segmented T-Test	2.77E-11