Predicting Left Ventricular Ejection Fraction Using Cardiac Echocardiograms

Katelyn Bechler  
Department of Biomedical Data Science  
Stanford University  
Stanford, CA 94305  
kkechler@stanford.edu

Raghav Garg  
Stanford University  
Stanford, CA 94305  
rmgarg@stanford.edu

Vy Ho  
Division of Vascular Surgery, Department of Surgery  
Stanford University  
Stanford, CA 94305  
vivianho@stanford.edu 

December 11, 2021

1 Abstract

Left ventricular ejection fraction (LVEF) is a valuable but resource-intensive measurement of cardiac function. Using a publicly available database of 10,024 expertly labeled echocardiograms, we trained and evaluated deep learning classifiers to predict abnormal LVEF with and without intermediate segmentation of the left ventricle. The best performing model utilized the VGG-16 architecture without intermediate segmentation to predict a binary outcome of normal versus abnormal ejection fraction with an accuracy of 0.85 and area under the receiver operating curve (auROC) of 0.85. Conversely, a segmentation model based on the two-dimensional U-Net achieved mean intersection over union of 0.85, training a model solely with segmentation masks produced a positive-only classifier with auROC 0.51.
2 Introduction

Left ventricular ejection fraction (LVEF) is a measure of cardiac function that is important for diagnosis, risk stratification, and treatment of many cardiovascular diseases (1,2). Defined as the percentage change in volume of the left ventricle from diastole to systole, LVEF is abnormally low in pathologies such as congestive heart failure and abnormally high in hypertrophic cardiomyopathy (1).

Traditionally, LVEF can be calculated from echocardiograms by performing at least 20 two-dimensional tracings of the left ventricle over a linear plane, then summing the volumes in a method of disks (3). These calculations require dedicated volumetric software and cardiologist supervision, thus limiting LVEF access to high-resource healthcare settings with cardiac laboratories and sonographers.

As such, investigators have developed deep learning models aimed at predicting LVEF as a continuous variable. These models demonstrated acceptable accuracy when trained on 10,000 to 50,000 expertly-labeled echocardiograms, but ultimately relied on extensive hospital and physician resources to produce data. Additionally, while granular LVEF measurement is needed in patients with chronic cardiovascular disease or those undergoing evaluation for advanced therapies such as transplant, categorical predictions of LVEF as normal or abnormal may be useful and more interpretable in outpatient or low-resource clinical settings. A provider without access to cardiology care could use a categorical LVEF measurement to rule in or rule out a specific disease process during differential diagnosis, or as an initial screening test prior to sending out a referral for formal echocardiogram.

We explore methods of developing deep learning classifier with and without left ventricular segmentation to enable less resource-intensive methods for automated LVEF estimation.

3 Related Work

Automated LVEF estimation was initially proposed in cardiac magnetic resonance imaging (MRI), which is the gold standard modality for measuring left ventricular volumes (4). Early technologies sought to assist radiologists in tracing chamber boundaries and measuring the left ventricular outflow tract using motion maps and fluid simulations (5,6). Lee at al demonstrated that a region growth method with iterative thresholding outperformed classical tracing methods for segmenting the left ventricular chamber, while Abdelmaguid et al utilized U-net models to segment the left ventricle with Dice coefficient greater than 0.95 for the region of interest, and a corresponding 0.094 root mean squared error for the ejection fraction calculation (7,8).

This led to corresponding interest in estimating LVEF in echocardiograms, which are more easily obtained due to reduced cost, wider availability of imaging devices, and fewer medical contraindications. Ouyang et al trained a ResNet-based deep learning model, Echonet-Dynamic, on a corpus of 10024 expertly
annotated echocardiograms by performing ventricular segmentations to identify frames at systole and diastole. The resultant frames were fed to a ResNet-based CNN which predicted ejection fraction values within a mean absolute error of 6.0 percent (9). Asch et al sought to predict EF directly without intermediate volumetric segmentations, and the resultant deep learning model Auto-EF had a 2.9 percent mean absolute deviation in ejection fraction estimation from reference values (10). While both models were trained on professionally obtained echocardiograms, which provide increased clarity and additional chamber views for volumetric calculations, fully convolutional neural networks have recently been utilized to segment videos using novel methods that incorporate temporal data. Valipour et al explores using recurrent units in addition to a fully convolutional network. Specifically, using a conv-GRU as the recurrent unit can help enable the extraction of temporal information from feature maps due to the reduced parameter set (11). In all completed experiments and for all metrics measured, the results proved the recurrent version performed comparable or superior to the baseline fully convolutional network, suggesting incorporating temporal data in this way can be advantageous. Similarly, Yang et al tackles automatic segmentation in ultrasound images for prostate cancer (12). They propose using RNNs to explore shape priors from hidden states sequentially, merging shape predictions from different perspectives into one comprehensive predictions and then applying refinement techniques.

Like Valipour et al and Yang et al we will explore ways in which we can segment the left ventricle in our dataset (11,12). However, we also want to focus on applying video segmentation of the left ventricle using semi-supervised methods, as being able to successfully perform this using partially labeled data could be extremely beneficial. Akin to Ouyang et al, we will be using the Echonet-Dynamic dataset of annotated echocardiograms to develop a deep learning model for LVEF prediction (9). Yet our model will ultimately make categorical predictions of LVEF (as opposed to continuous) and also explore semi-supervised methods as we believe these partially labeled data scenarios are more common and would achieve higher clinical utility as they would support less resource-intensive methods for automated LVEF estimation.

4 Data

The Stanford Echonet-Dynamic Database is a deidentified dataset of 10,024 echocardiography videos with expert-labeled ejection fraction, stroke volume, end-diastolic volume, and segmentations of the left ventricle (9). Videos are formatted in .avi format, and two frames in each video are labelled with end systolic and end diastolic volumes. A sample data schematic is provided in Figure 1, and variables are displayed in Figure 2.

In terms of results pertaining to our data preprocessing, we were successfully able produce the code that converts the hand drawn "mesh" of lines drawn by an echocardiographer on the left ventricle into a full ventricle mask. We did so by using the PIL library to first "draw these lines" onto an array and then
running a alpha shape algorithm with alpha = 0.3 to draw the boundary around the mesh. Though the alpha shape method is not too computationally expensive, but we were not sure if there was another simpler way to get the boundary of the mesh by just using the start and end points of the lines. However, after completing the preprocessing, we were able to visually confirm that the images, masks, and sample weights correctly lined up for all three splits of the data. Figure 3 below shows the an example of this conversion.

Figure 1: Data example of frame-by-frame echocardiogram with annotated end systolic and end diastolic masks

Figure 2: Dataset variables provided in the EchoNet Dynamic dataset
5 Approach

The Ouyang et al. paper originally published the EchoNet dataset and modeled it using a ResNet Base (9). Their findings on the effect of frame rate sampling and clip length informed our video preprocessing. We complemented their work by exploring a UNet architecture and fully convolutional networks to perform a similar prediction task. Contrary to Ouyang et al who generated continuous predictions for LVEF, our task predicted LVEF and classified into two clinically relevant categories, "abnormal" and "normal". Prior to modeling, we annotated the labels and classified into their clinically relevant categories, which then fed into our various deep learning models. We classified LVEF based on the calculated ejection fraction provided in the original dataset, with a "normal" LVEF being 50-75% and an "abnormal" LVEF being <50% or >75%.

We aimed to train and evaluate several different models in predicting LVEF from the Echonet Dynamic dataset through two different approaches. The first approach was to build a CNN model taking two frames as input and output a prediction of normal or abnormal LVEF. The second approach was to build a 2D-UNet taking two image frames as input to produce segmentation masks of the left ventricle, which were then used to predict normal or abnormal LVEF.

Model design and training was done in Python 3.8 using the keras deep learning library. For the CNN-only approach, two models were trained: a simple model using 2 convolutional, 1 max pooling, and 1 dense layer, and a model using pre-trained weights from the VGG-16 architecture. All models were trained on a 0.6/0.2/0.2 training, validation, and test split with a 23.2% class prevalence in each partition.

Figure 3: Echo with corresponding drawn mesh and computed mask
We also built a 2D-UNet model to segment the left ventricle in a video frame. The ground truth masks were drawn using the manual tracings from the image frames and the alpha shape algorithm, with data augmentation as appropriate.

Finally, we built a two-step model to predict "normal" or "abnormal" LVEF using the 2D-UNet to segment the left ventricle and use the original images to generate masks, which we then fed into the fully convolutional network classifier to predict LVEF status.

### Table 1: Distribution of features in train, validation, and test datasets

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abnormal Ejection Fraction (N, %)</td>
<td>1398 (23.2%)</td>
<td>466 (23.2%)</td>
<td>466 (23.2%)</td>
</tr>
<tr>
<td>End Systolic Volume (ESV)</td>
<td>43.8</td>
<td>43.2</td>
<td>42.4</td>
</tr>
<tr>
<td>End Diastolic Volume (EDV)</td>
<td>91.8</td>
<td>91.2</td>
<td>89.9</td>
</tr>
<tr>
<td>Number of Frames</td>
<td>176.5</td>
<td>187.1</td>
<td>174.9</td>
</tr>
</tbody>
</table>

6 Experiments

**Initial Data Processing:** The video files in the EchoNet Dynamic database consisted of 10,030 total videos, 10,024 of frame size 112x112 and 6 of frame size 768 x 1024. In order to reduce heterogeneity in frame size, the latter videos were excluded. Annotated systolic and diastolic frames from each echocardiogram were then extracted using the OpenCv2 package, and the continuous LVEF values provided were categorized into normal and abnormal bins (13). The resultant corpus was split into 60% train, 20% validation, and 20% held-out test set using the sklearn package (14). The distribution of labels and other data features is described in Table 1.

**Fully Convolutional Network Experiments:**

The first model developed was a simple convolutional neural network. The systolic and diastolic images were concatenated prior to undergoing two 2-dimensional convolutional layers, followed by 1 2-dimensional max pooling layer, 1 2-dimensional global average pooling layer, and 1 dense layer with sigmoid activation. A visual schematic of the CNN is provided in Figure 4. The second model leveraged the VGG-16 model, which is pretrained on the ImageNet corpus and has demonstrated accuracy in other cardiovascular imaging tasks including electrocardiograms (15,16). A visual schematic of this CNN is provided in Figure 5.

Both models were tuned on the training and validation with a range of learning rates [0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001] and the ADAM optimizer for best accuracy. Once the best learning rate was identified, the model was run on the held-out test set. The operating point was selected using Youden’s index.
<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>auROC</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple CNN</td>
<td>0.83</td>
<td>0.84</td>
<td>0.72</td>
<td>0.82</td>
</tr>
<tr>
<td>VGG-16 CNN</td>
<td>0.85</td>
<td>0.85</td>
<td>0.77</td>
<td>0.80</td>
</tr>
<tr>
<td>2D-Unet + Simple CNN</td>
<td>0.77</td>
<td>0.51</td>
<td>1.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 2: Accuracy, auROC, Sensitivity, and Specificity of all reported models

The resultant performance parameters are displayed in Table 2, while the receiver operating characteristic and precision-recall curves are displayed in Figure 6. Addition of the VGG-16 CNN to the model resulted in improved accuracy to 0.846 and sensitivity to 0.77.

In stratifying cases by the continuous LVEF measurements provided in the EchoNet Dynamic datasets, it appeared that the model had particular difficulty identifying abnormal EF that were slightly below the normal 50% threshold - false negative cases had a mean LVEF of 45.5 compared to true positive cases with a mean of 37.2 (Student’s T-test, p-value = 3.3e-22). Figure 7 illustrates the distributions of continuous LVEF, end-systolic volume (ESV), and end-diastolic volume (EDV) by classifier output.
Figure 4: Schematic for the Simple CNN
Figure 5: Schematic for the VGG-16 transfer learning CNN
Figure 6: Receiver operating and precision-recall curves for the simple (A,B) and VGG-16 CNN (C,D) models, respectively.
Figure 7: Distribution of expertly-annotated continuous left-ventricular ejection fraction (A), end-systolic volume (B), and end-diastolic volume (C) by VGG-16 classifier output.
2D-UNet Segmentation Experiments:
As we believed that a 2D segmentation of the left ventricle could add to the base video input, we developed a 2D UNet model that used the alpha-shape algorithm transformed masks as output labels, and the original video frames as input images. The model performed reasonably well with an IoU score of 0.85, and visually comparable results. The largest difference between the model’s masks and the true masks were that the model more conservatively segmented the left ventricle, whereas the echocardiographer annotations often went over the bordering tissue. Figure 8 compares the two segmentations of a given frame below.

![Figure 8: 2D-UNet Segmentation Test Mask vs Prediction](image)

3D-UNet Segmentation Experiments:
Our results indicated that a 3D UNet model was not be able to effectively segment the left ventricle due to the sparsity of the labeled frames and because all labeled data is only along 2 dimensions (width and depth). Our binary accuracy for the model was 0.1, and the average IOU was similarly 0.1. Looking at the predictions versus the mask, it seems like the model is able to identify the left ventricle reasonable confidently, but is also picking up other cavities filled with fluid, even those outside heart. Figure 9 below shows 3 original echo frames with their corresponding masks and predictions.
Figure 9: 3D-UNet Segmentation Test Mask vs Prediction
2D-UNet Segmentation + Convolutional Network Experiments:

Using the most successful 2D-UNet Segmentation model (mean IoU 85%) and the simple CNN model, we developed a model to classify LVEF as "normal" or "abnormal" using the output masks of the 2D-UNet as the input for the fully convolutional network model. We believed using a combination of these two highly accurate models would produce a relatively good model for predicting LVEF and using the masks as input (as opposed to the raw images) might provide additional insight that could improve accuracy. However, the results of this model were not exemplary. The auROC score was 0.50, which is no better than random classification, with a 1.00 sensitivity, and 0.00 specificity, see Figure 10. We also tried the prediction classification task with VGG16 convolutional model using the output masks from the 2D-UNet segmentation model to compare. Unfortunately, this yielded similar results, with a auROC of 0.50 as well.
7 Conclusion

In conclusion, our efforts to develop a deep learning classifier for prediction of abnormal LVEF from echocardiography yielded a best-performing model based on the VGG-16 architecture, with an accuracy of 0.85, area under the receiver operating curve (auROC) of 0.85, and sensitivity of 0.77. While this model’s performance was inferior to that of the original EchoNet Dynamic model, which utilized ResNet architecture with a 0.92 auROC for reduced ejection fraction classification, we demonstrate acceptable performance tradeoff for a model that is also able to detect abnormally high EF (9). That we were unable to train a model purely on left ventricular segmentations alone suggests that while left ventricular volume is the only metric used to traditionally calculate LVEF, other features are being incorporated into the model in this deep learning approach. While class activation maps were not informative for in our experiments, further evaluation with ablation studies may better elucidate relevant features.

In addition to further model evaluation, we plan to continue experimenting with the 3D-Unet to enable segmentation of multiple contiguous frames, which may improve prediction by capturing the temporal relationship between beat-to-beat left ventricular volumes. In order to improve the clinical relevancy of the classifier, we will also explore a multiclass classification schema with low, normal, and high LVEF classes and evaluate whether there is a tradeoff in performance. While adding classes requires more discriminative model performance for equivalent auROC, this may lead to more accurate classification as there are distinctly
different pathologies that lead to low EF versus high EF.

8 Contributions

Katelyn Bechler: Literature review, 2D-UNet segmentation +CNN experiments, Results discussion, Wrote paper

Raghav Garg: Literature review, 3D-UNet segmentation experiments, 2D-Net segmentation experiments, Results discussion, Wrote paper

Vy Ho: Literature review, simple CNN and VGG-16-CNN training and evaluation experiments, Results discussion, Wrote paper

9 References