Interpretable deep learning prediction of 3d assessment of cardiac function*

Grant Duffy[1], Ishan Jain[1], Bryan He[2] and David Ouyang[1]

1. *Department of Cardiology, Smidt Heart Institute, Cedars-Sinai Medical Center*

127 S San Vicente Blvd A3600, Los Angeles, CA 90048

2. *Department of Computer Science, Stanford University*

353 Jane Stanford Way, Stanford, CA 94305

Email: David.Ouyang@cshs.org

As deep learning plays an increasing role in making medical decisions, explainability is playing an increasing role in satisfying regulatory requirements and facilitating trust and transparency in deep learning approaches[1]. In cardiac imaging, the task of accurately assessing left-ventricular function is crucial for evaluating patient risk, diagnosing cardiovascular disease, and clinical decision making[2,3]. Previous video based methods to predict ejection fraction yield high accuracy but at the expense of explainability[4] and did not utilize the standard clinical workflow. More explainable methods that match the clinical workflow, using 2D semantic segmentation, have been explored but found to have lower accuracy[5]. To simultaneously increase accuracy and utilize an approach that matches the standard clinical workflow, we propose a frame-by-frame 3D depth-map approach that is both accurate (mean absolute error of 6.5%) and explainable, utilizing the conventional clinical workflow with method of discs evaluation of left ventricular volume. This method is more reproducible than human evaluation and generates volume predictions that can be interpreted by clinicians and provide the opportunity to intervene and adjust the deep learning prediction.

Keywords: Machine Learning; Echocardiology; Explainable AI; Depth Map.

1. Introduction

There have been significant advances in the application of artificial intelligence in medical contexts, with deep learning models applied to dermatology[6], radiology[7], cardiology[4], and many other domains of medical data interpretation. But for medical applications, accuracy on retrospective datasets is not enough to be accepted into practice, with open questions regarding how to understand deep learning model predictions and how humans can interact and adjust model output. Many machine learning algorithms are 'black box' solutions that lack an explanation of their predictions[8]. When model predictions are incorrect, its often difficult to understand why or how the model fails, leading to distrust in predictions by physicians and putting into question the role of deep learning when making critical medical decisions. Black box algorithms can also hide biases and pitfalls that would otherwise be apparent in more explainable methods. Deep learning algorithms can reinforce hidden biases and disparities if not well understood or fully explored[9]. Regulations for using AI in medicine are in flux but could potentially require explainability before a method can become

This work is supported by NIH K99 HL157421-01

standard practice in medicine or used in medical devices[1,10]. The goal of explainability in AI and deep learning is to produce models that can be interpreted in a way that allows humans to understand and validate predictions. This allows a better understanding between not only the statistical relationships between attributes in a dataset, but also causal and mechanistic relationships. Explainable AI aims to overcome the pitfalls of 'black box' algorithms and is important for widespread adoption of AI if it is to reach its potential in medicine and beyond.

The impact of cardiac disfunction on overall health and wellbeing ranges from exercise intolerance and fatigue to higher risk of mortality. Cardiovascular disease is the leading cause of death in the United states and projected cost of care for cardiovascular disease is projected to increase by 61% to $276 billion between 2010 and 2030[11]. The need for low-cost, accurate assessment of cardiac function is essential to providing life-prolonging treatment and detecting at-risk patients. Ejection fraction (EF) is typically measured using cardiac ultrasound, or echocardiography, because of its high temporal resolution, low cost, and availability.

$$EF\% = 100 * \frac{EDV - ESV}{EDV} \quad (1)$$

Left ventricular (LV) ejection fraction (EF) is one important metric for evaluating cardiac function. EF is defined by the percent volume change between the end diastolic volume (EDV) and the end systolic volume (ESV) as shown in Eq. 1 and describes what percent of the left ventricle volume is pumped out every heartbeat and is interpreted as a measure of the heart's efficiency. Low EF, even marginally reduced EF, has been shown to be highly indicative of heart disease[12–14]. While traditional echocardiography utilizes 2D image acquisition, the heart is a 3D structure, requiring assumptions and approximations to estimate volume from the area measured in standard views. The method of discs (MOD) is the standard method for calculating LV volume from a single view. This method assumes that cross sections of the LV perpendicular to the major axis are circular, as shown in figure 1. The volume is then calculated using Eq. 2 where dx is the perpendicular distance between cross sections and d_i is the length of the cross section i or the diameter of the circular cross section.

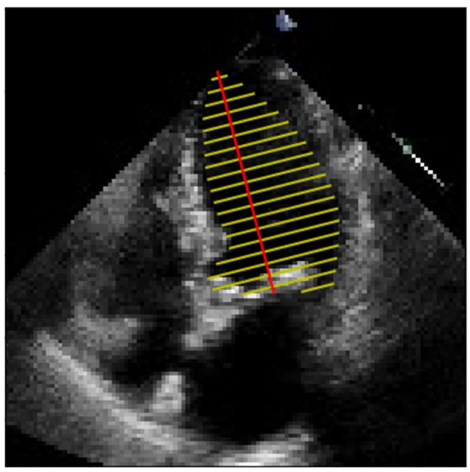

Fig. 1. Example human method of discs label

$$V = \sum \frac{\pi {d_i}^2}{4} dx \quad (2)$$

There have been previous attempts to evaluate ejection fraction using artificial neural networks, and although these attempts address explainability, none of them achieve a high level of accuracy while proving that the method is basing its predictions on the physical quantities that define EF. There is an inherent tradeoff between interpretability and accuracy, requiring additional innovations to optimize both attributes of an AI model, as portrayed in Figure 2. Our method addresses this problem by directly predicting the 3D geometry of the left ventricle in the form of a depth map which can be directly evaluated to find volume and EF. By using a depth map of the LV as an intermediate step in calculating EF, we provide a method to directly interpret the model's performance and prove that the physical quantities used to calculate EF are being used.

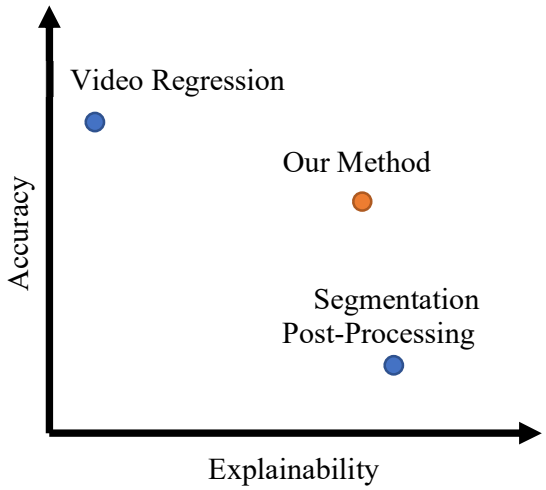

Fig. 2. Hypothetical tradeoff between accuracy and explainability

2. Related Works

Previous ML solutions to predict EF closely align with solutions to traditional ML tasks, namely regression and semantic segmentation. The our first approach[5] utilized an image based CNN regression model to predict several phenotypes and cardiovascular function, including ejection fraction, from a single frame of echocardiogram videos. This method was able to predict EF with an R^2 value of 0.5. Explainability for this model was addressed by using a gradient-based sensitivity map to highlight regions of interest to the model. Although this method gives physicians some insight into where in the image the model is extracting the most information, it does not prove that the model is finding systolic and diastolic volumes to calculate EF. Because the input to the model is only one frame, it cannot be precise in estimating systolic and diastolic volumes.

Subsequently, we proposed EchoNet-Dynamic[4], which uses a 3D video-based CNN model with a regression task to predict ejection fraction. This performs at a much higher level, achieving an R^2 of 0.81. Explainability is then addressed with a separate segmentation model. The segmentation model is used to accurately predict the left ventricle region. This model produces predictions that

are easily interpretable by physician. Although these predictions may give physicians confidence, using them as explainability for the ejection fraction is problematic because it is done with a different model. There could potentially be examples where the regression model predicts a bad EF, but the segmentation model performs well. Interpreting the performance of the regression model using the segmentation model could result in having high confidence in bad predictions or vice versa.

One solution to this problem is to calculate the ejection fraction directly from the segmentation model output[5]. This process involves predicting the segmentation for each frame, predicting the major axis direction, and creating perpendicular discs to use for volume calculation. The benefit of this approach is that the model predictions are directly used for volume calculation and is done so in a method that is already standard practice. This means that the predictions are easily interpreted and could even be corrected by physicians. The drawback of this method is that every step of the post-processing adds error to the volume calculation. On top of the model inaccuracy, picking the major axis and disc cross section both introduce more error. These factors contribute to a low overall R^2 of 0.49.

3. Method

The method proposed here improves apron the segmentation approach by predicting the z-depth of the LV at each pixel. Any pixels outside the LV are zero. Figure 3 visualizes an example LV depth map. Depth maps are predicted for every frame of an echocardiogram video. The volume for each frame is calculated by summing the pixel depths. From the frame volumes, end systole and end diastole are selected and used to calculate EF. These depth maps can be easily visualized as a 3D surface over the LV. This approach is analogous to doing a regression to depth for each pixel instead of classification for each pixel as seen in a segmentation model. CNNs have previously been used to predict depth information and is an approach that has become commonplace in computer vision and autonomous AI[15]. To calculate the volume for a given prediction, the pixel depths are simply summed together. This allows more of the work to be done by the model but is still closely aligned with the human method and therefore is easily interpreted.

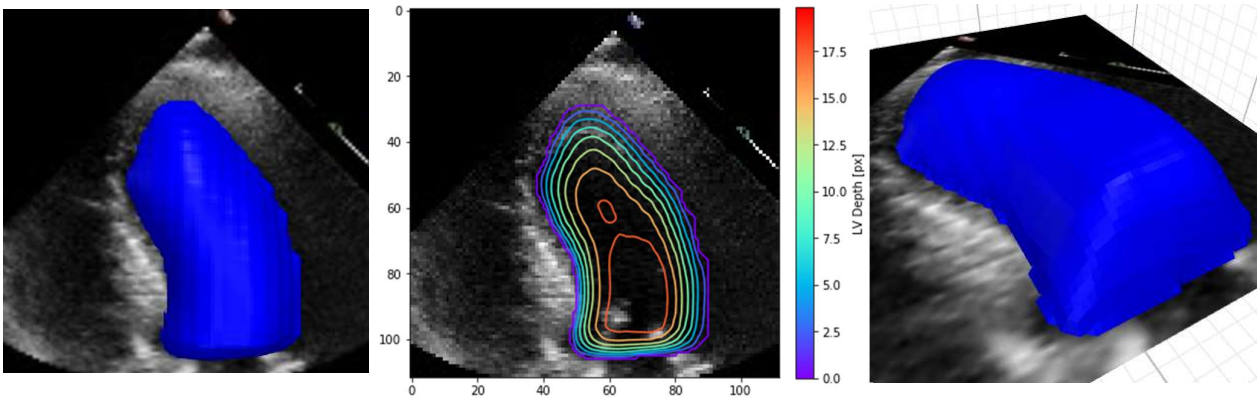

Fig. 3. Example depth map prediction shown in different perspectives and with contours to show geometry

3.1. *Model Architecture*

The model architecture used is based on a DeepLabV3 with a ResNet-50 backbone. The Torchvision[16] implementation, as shown in Figure 4 a), is composed of roughly three parts, the ResNet-50[17] backbone, a DeepLabV3[18] head, and a convolutional classifier. The output of the DeepLab head and classifier are the same resolution as the output layer of the ResNet[17] backbone. For 112x112 images, this resolution is 14x14. The output of the classifier is upsampled using simple bilinear interpolation leading to interpolation artifacts seen in Figure 5 (center). This is solved by replacing the classifier and upsampling with a transposed convolution with kernel size 8x8 and stride 8x8, shown in Figure 4 b). This layer acts as a learnable upsampling layer that allows finer detail in the predictions.

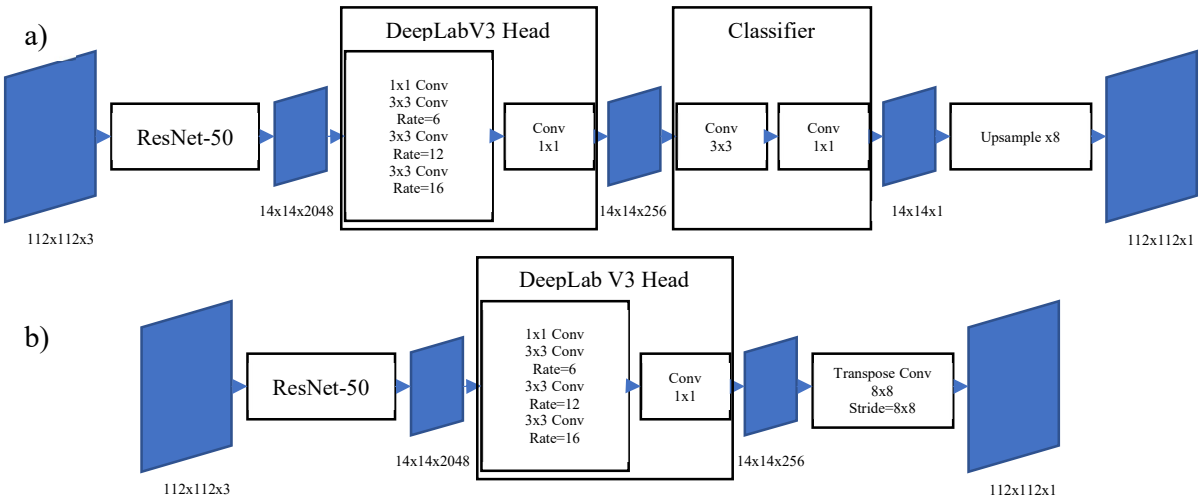

Fig. 4. a) Original DeepLabV3 architecture b) architecture used here

Figure 5 shows the ground truth (left) label and two depth map predictions; the middle is a prediction from an unmodified DeepLabV3 model, the right was predicted by the modified DeepLabV3 that includes a trainable upsampling layer. The output of the DeepLab head on the unmodified model is of resolution 14x14 pixels and is upsampled using linear interpolation to the full 112x112. This creates interpolation artifacts that appear as 8x8 square surface facets that are clearly not physiological. Although the overall performance metrics were similar between the two

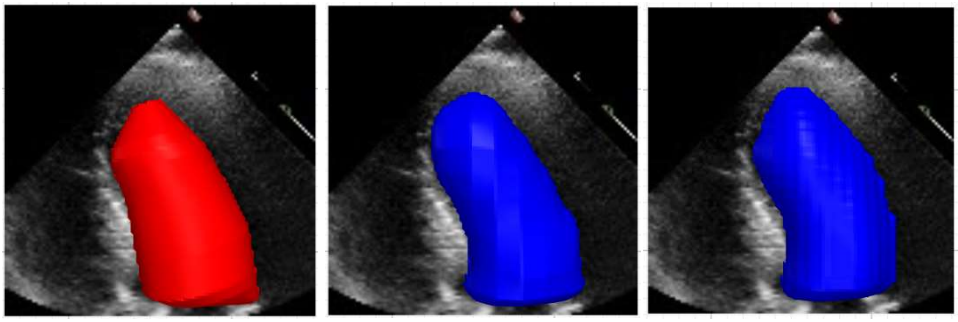

Fig. 5. Example ground truth label, original DeepLabV3 inference, our modified model inference from left to right

models, it is clear that the modified model is capable of predicting much higher detail and follows the curvature of the LV more closely.

3.2. *Ground Truth Depth Map Labels*

Traditionally, when calculating LV volume, humans trace the outline of the LV and label the major axis. Cross sections are then generated perpendicularly to the major axis. Volume is calculated using the method of discs, whereby each cross section is considered to be the diameter of a cylinder with a length equal to the spacing between the cross sections.

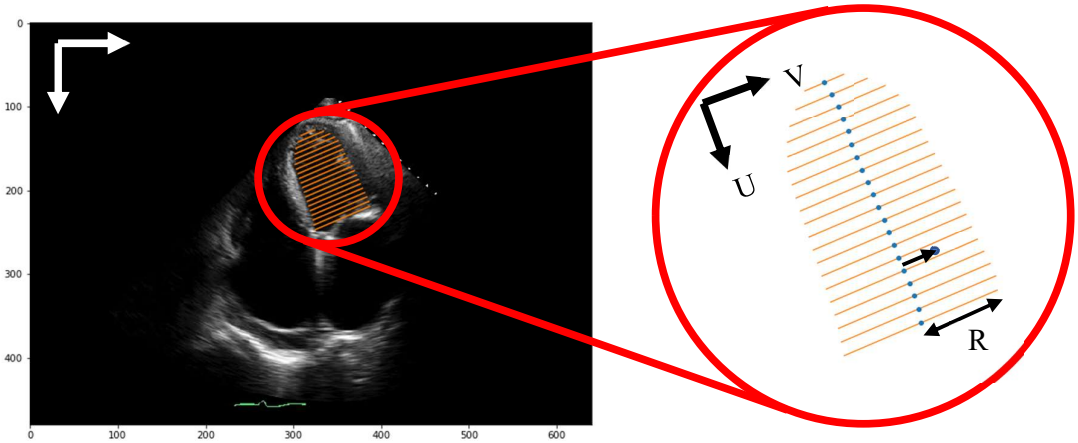

Fig. 6. Depth map label diagram

For the depth map approach, the value of each pixel corresponds to the depth of the LV perpendicular to the image at that pixel. Human labels in the form of MOD cross sections are used to generate depth map labels. The general approach used to do this is to find function of pixel location x and y and returns the depth at that location given a human label.

$$z = f(x, y) \tag{3}$$

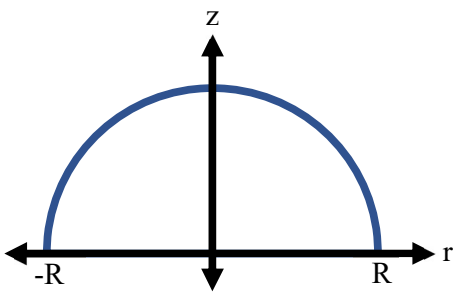

Fig. 7. Semicircle relationship between depth and radius

If we consider each cross section to be the diameter of a disc, then the depth can be modelled as semi-circle function where R is the radius of the cross section and r is the radius of the point of interest to the center of the disc.

$$z = \sqrt[2]{R^2 - r^2} \tag{4}$$

To find values of R and r for every pixel, the coordinate system is interpolated such that all the cross-sections are vertical. This new coordinate system is described in terms of u and v.

$$\begin{bmatrix} u \\ v \end{bmatrix} = [M_{rot}] \begin{bmatrix} x \\ y \end{bmatrix} \tag{5}$$

Because all the cross-sections are parallel and perpendicular to the long axis, their radii and centers can be interpolated as a function of u. This is done for every pixel giving values of R and r for every pixel. Eq. 2 is then used to approximate the depth at each pixel from these values. For pixels where r > R, z is set to zero corresponding to a depth of zero outside the LV.

3.3. *Dataset*

Data was processed from standard echocardiogram studies. Apical four chamber views with human labelled LV traces were selected and any videos with color doppler or bad labels were removed. The dataset consists of 10,030 unique videos corresponding to unique patients split into subsets of 7,465, 1,277, and 1,288 videos for training validation and testing respectively. This is the same dataset used for training and evaluating EchoNet-Dynamic and segmentation post-processing method. Typically, two frames per video are annotated, end diastole and end systole. Therefore, there are 19,986 frames with annotated LV traces which were processed to produce depth map labels.

3.4. *Loss Function and Training*

The primary loss function used for training is MSE loss for each pixel. This loss is augmented by MSE loss for volume as calculated as the sum of pixels for each frame. This is described by equation X where y and $\hat{y}$ are the pixel values of the ground truth and predicted depth maps respectively. w is the weighting term applied to the volume loss. Experimentation found $w = 10$ to perform well.

$$l = \frac{1}{n}\sum(y - \hat{y})^2 + w(\sum y - \sum \hat{y})^2 \tag{6}$$

An Adam optimizer was used with a learning rate of 0.01. The model was trained with a batch size of 128 for 100 epochs. The epoch with the lowest validation loss was selected for evaluation.

3.5. *Beat-to-Beat EF Calculation*

The model predicts a depth map from which a volume can be calculated for every frame of a video. To calculate ejection fraction, end diastolic and end systolic volumes need to be found. The *find_peaks*[19] function is used to find all of the end systole and end diastole frames in the video by finding local maxima in calculated volume across the video. Several hyperparameters including prominence and distance were tuned for this application. Sequential pairs of systole and diastole are used to calculate EF for every beat of the video. The median EF is selected from these beats.

4. Results

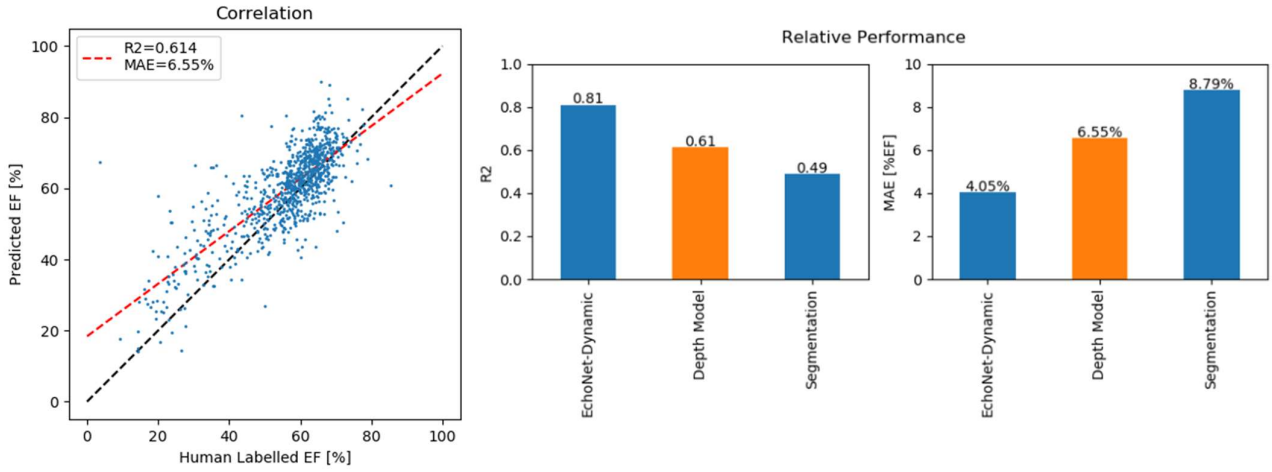

Fig. 8. Performance of depth map model and comparison to prior works

4.1. *Performance on Test Set*

Performance analysis was done using a test set not yet seen by the model during training. For these patients, inference was run for the entire clip. Volumes were calculated for each frame and the beat-to-beat method was used to calculate ESV, EDV and EF. When comparing the EF predicted by this method to the EF calculated by humans, there is a correlation with an R^2 value of 0.61 and MAE of 6.55%. Although the performance did not match the video regression method, it did perform better than the segmentation post-processing method. It is known that the inter-observer variability can be anywhere from 7.6% to 13.9%[20–22]. Because this variability applies to the 'ground truth' labels in the test set, it is likely that the true performance of all three methods exceeds the performance on the test set.

Table 1. Relative performance compared to prior work.

Method	R2	MAE
EchoNet	0.50	7.00%
EchoNet-Dynamic	0.82	4.05%
Segmentation Processing	0.49	8.79%
Depth-Map	0.61	6.55%

When comparing the performance of this method to EchoNet-Dynamic, there are several limitations that may be contributing to the performance gap between the methods. EchoNet-Dynamic is a true end-to-end deep learning approach that does not rely on any heuristics or assumptions about the data or predictions. This end-to-end regression task, even with tens of thousands of training examples, might be prone to overfitting and cause limited generalizability. In contrast, this approach is limited to predicting the LV geometry for each frame of the video. While the absolute performance has a slight decrease in performance, the workflow of using selection of systole and diastole and downstream calculation of EF provides a more robust, likely generalizable, and internally consistent interpretation. Depth Map Interpretation

4.2. *Volume Tracing Interpretation*

In practice, the video regression method would only give the clinician the EF number with no additional information. For the depth map method, a clinician can look at the depth map and visually inspect wether or not the volume appears to be correct. This is demonstrated in figure 9 where a) shows a good depth map prediction while b) shows a prediction where the model failed. This means that when implemented with a clinician in-the-loop, assessing the quality of predictions, this method will perform better than by itself. If clinicians are able to identify and correct predictions that are not biologically plausable and whose errors are greater than 20% (which accounts for only the worst 2.9% of cases in the test set), the R2 performance of this method increases to 0.75 (MAE to 5.36%). Volume Trace Interpretation.

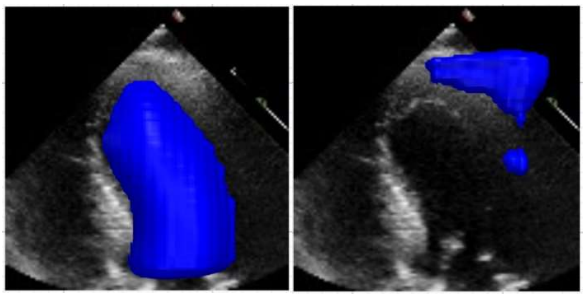

Fig. 9. Example 'good' and 'bad' depth map predictions

Another advantage of this method over EchoNet-Dynamic and traditional human MOD is that it creates volume predictions for every frame in a video. This would be overwhelmingly time consuming and tedious for a human to produce. Figure 10 shows an example predicted volume for an entire video. This is done by summing the pixel values for each frame's predicted depth map. Green and orange points represent frames identified as end diastole and end systole respectively as calculated using the scipy *find_peaks* function. The red X's show which frames were selected by the human for calculating ejection fraction. These volume traces may give physicians additional information about cardiac function that may otherwise go unnoticed.

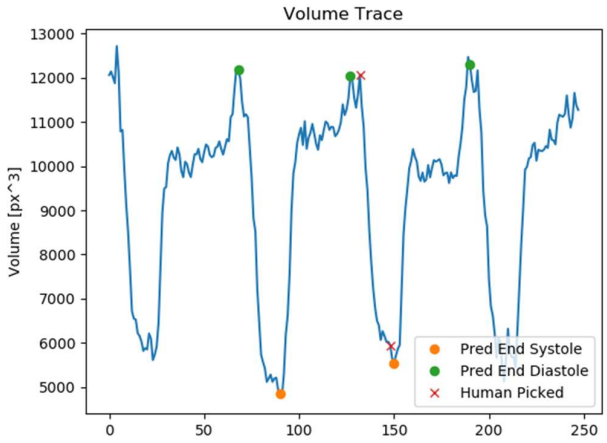

Fig. 10. Example predicted volume tracing

5. Discussion

Our method is a depth map-based approach to predicting LV volume. The predicted volumes are used to estimate cardiac function in an explainable way. A modified DeepLabV3 architecture was used to improve the performance on small images. Although this method does not exceed the accuracy of the state-of-the-art EchoNet-Dynamic approach of video regression, it outperforms human inter-observer variability while achieving a high level of explainability.

This approach could allow clinicians to evaluate cardiac function quickly while giving them the autonomy to interpret the model's performance and use their own judgment when considering the model's predictions in making decisions. Because the approach is easily interpreted, the method will more likely be trusted and accepted into practice.

Although regulations around the world regarding the use of AI in medical devices are in flux, guidelines for best practices and interpretations of current regulations and guidelines suggest that regulations will be made requiring interpretability along with responsible data sourcing and performance benchmarking. Currently, the U.S. Food and Drug Administration (FDA) has proposed best practices and protocols for managing medical devices based on AI/ML, but they have not been solidified. FDA approval, for now, continues on a case-to-case basis. More widespread trust and acceptance in the future will likely include requirements for explainability[23].

By framing ML problems in ways that align with human understanding, explainability can be achieved in without sacrificing accuracy. Future work could further improve performance and explainability by applying this ideology to a greater scope, for example to include the prediction of end systole/diastole or prediction of image quality.

6. Acknowledgements

D.O. is supported by the National Institutes of Health grant K99-HL157421

7. Bibliography

1. 1.Holzinger, A., Biemann, C., Pattichis, C. S. & Kell, D. B. What do we need to build explainable AI systems for the medical domain? *arXiv [cs.AI]* (2017).
2. 2.Ziaeian, B. & Fonarow, G. C. Epidemiology and aetiology of heart failure. *Nat. Rev. Cardiol.* **13**, 368–378 (2016).
3. 3.Dellinger, R. P. *et al.* Surviving Sepsis Campaign: international guidelines for management of severe sepsis and septic shock, 2012. *Intensive Care Med.* **39**, 165–228 (2013).
4. 4.Ouyang, D. *et al.* Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).
5. 5.Yuan Neal *et al.* Systematic Quantification of Sources of Variation in Ejection Fraction Calculation Using Deep Learning. *JACC Cardiovasc. Imaging* **0**,.
6. 6.Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
7. 7.Sarker, L., Islam, M. M., Hannan, T. & Ahmed, Z. COVID-DenseNet: a deep learning architecture to detect COVID-19 from chest radiology images. (2020).
8. 8.Adadi, A. & Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **6**, 52138–52160 (2018).

9. 9.Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
10. 10.Center for Devices & Radiological Health. Artificial Intelligence and Machine Learning in Software. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.
11. 11.Heidenreich, P. A. *et al.* Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. *Circulation* **123**, 933–944 (2011).
12. 12.Chioncel, O. *et al.* Epidemiology and one-year outcomes in patients with chronic heart failure and preserved, mid-range and reduced ejection fraction: an analysis of the ESC Heart Failure Long-Term Registry. *Eur. J. Heart Fail.* **19**, 1574–1585 (2017).
13. 13.Shah, K. S. *et al.* Heart Failure With Preserved, Borderline, and Reduced Ejection Fraction: 5-Year Outcomes. *J. Am. Coll. Cardiol.* **70**, 2476–2486 (2017).
14. 14.Papolos, A., Narula, J., Bavishi, C., Chaudhry, F. A. & Sengupta, P. P. U.S. Hospital Use of Echocardiography: Insights From the Nationwide Inpatient Sample. *J. Am. Coll. Cardiol.* **67**, 502–511 (2016).
15. 15.Eigen, D., Puhrsch, C. & Fergus, R. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. *arXiv [cs.CV]* (2014).
16. 16.vision: Datasets, Transforms and Models specific to Computer Vision. (Github).
17. 17.He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *arXiv [cs.CV]* (2015).
18. 18.Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv [cs.CV]* (2017).
19. 19._peak_finding.py at v1.7.0 · scipy/scipy. (Github).
20. 20.Malm, S., Frigstad, S., Sagberg, E., Larsson, H. & Skjaerpe, T. Accurate and reproducible measurement of left ventricular volume and ejection fraction by contrast echocardiography: a comparison with magnetic resonance imaging. *J. Am. Coll. Cardiol.* **44**, 1030–1035 (2004).
21. 21.Cole, G. D. *et al.* 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. *Int. J. Cardiovasc. Imaging* **31**, 1303–1314 (2015).
22. 22.Koh, A. S. *et al.* A comprehensive population-based characterization of heart failure with mid-range ejection fraction. *Eur. J. Heart Fail.* **19**, 1624–1634 (2017).
23. 23.Gerke, S., Babic, B., Evgeniou, T. & Cohen, I. G. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *NPJ Digit Med* **3**, 53 (2020).