Leveraging 3D Echocardiograms to Evaluate AI Model Performance in Predicting Cardiac Function on Out-of-Distribution Data*

Grant Duffy, Kai Christensen and David Ouyang
Department of Cardiology, Smidt Heart Institute, Cedars-Sinai Medical Center,
127 S San Vicente Blvd A3600
Los Angeles, CA 90048
Email: David.Ouyang@cshs.org

Advancements in medical imaging and artificial intelligence (AI) have revolutionized the field of cardiac diagnostics, providing accurate and efficient tools for assessing cardiac function. AI diagnostics claims to improve upon the human-to-human variation that is known to be significant[1–3]. However, when put in practice, for cardiac ultrasound, AI models are being run on images acquired by human sonographers whose quality and consistency may vary. With more variation than other medical imaging modalities[4], variation in image acquisition may lead to out-of-distribution (OOD) data and unpredictable performance of the AI tools. Recent advances in ultrasound technology has allowed the acquisition of both 3D as well as 2D data, however 3D has more limited temporal and spatial resolution and is still not routinely acquired[5]. Because the training datasets used when developing AI algorithms are mostly developed using 2D images, it is difficult to determine the impact of human variation on the performance of AI tools in the real world. The objective of this project is to leverage 3D echos to simulate realistic human variation of image acquisition and better understand the OOD performance of a previously validated AI model[2]. In doing so, we develop tools for interpreting 3D echo data and quantifiably recreating common variation in image acquisition between sonographers. We also developed a technique for finding good standard 2D views in 3D echo volumes. We found the performance of the AI model we evaluated to be as expected when the view is good, but variations in acquisition position degraded AI model performance. Performance on far from ideal views was poor, but still better than random, suggesting that there is some information being used that permeates the whole volume, not just a quality view. Additionally, we found that variations in foreshortening didn't result in the same errors that a human would make.

Keywords: 3D Echo; AI; Machine Learning; Echocardiology.

*This work is supported by NIH R00 HL157421-01

1. Introduction

Echocardiography, or cardiac ultrasound, is the most prevalent imaging modality[6]. Cardiac ultrasound is able to provide an accurate, noninvasive views of the heart in real time with limited equipment and with high temporal resolution[7]. In traditional transthoracic echocardiology, a sonographer will acquire 2D images and videos of the heart in standard orientations or views. Two standard views are the apical four chamber (A4C) and apical two chamber (A2C) views which are both views taken along the major axis of the heart from its apex. These views are crucial for assessing cardiac function, diagnosing heart failure and cardiac hypertrophy[1,6,8–15]. These two views are in theory only separated by a probe rotation of roughly 60 degrees, however this depends on sonographer judgement for the view quality and probe placement.

Recent advances in ultrasound technology have increased the temporal and spatial resolution of images acquired. Wide field of view allows for 3D images to be acquired with the same probes and hardware, however at lower resolution[7,8]. In addition to the standard TTE views, sometimes additional 3D images are acquired to better characterize complex cardiac structures and provide holistic evaluates of cardiac form and function. Focused images of the heart valves as well as the left ventricle can be used to accurately assess metrics that might be challenging to measure in 2D images.

One example of acquisition error in 2D images is foreshortening, where inappropriate or suboptimal images of the left ventricle can cause overestimation of the cardiac function[16,17]. Apical views depend on being placed near the apex of the left ventricle, which should not contract in, however off-axis foreshortened views will show contraction of the left ventricle that exaggerate the left ventricular function. The result of this error is the underestimate of LV volume at systole and ultimately an overestimate of ejection fraction[17]. Although foreshortening is known to be a common source of measurement error, it is difficult to know how prevalent it is because it is difficult to quantify foreshortening in 2D images. There have been attempts to automatically detect foreshortening using machine learning or other algorithms[16,18,19]. These algorithms need to be run in real-time on the ultrasound machine or trained on other modalities limiting their practicality.

Although adding 3D acquisitions to a study may add value in these cases, it also takes additional time and training. The result is that 3D echo images are much less prevalent. In the Cedars Sinai Medical Center (CSMC), apical 3D echo images are outnumbered by other video acquisitions roughly 11,000 to 1 making 3D echo datasets of reasonable size rare.

There is a large, and quickly growing, body of research dedicated to AI in medicine and specifically cardiology. Several models aim to automate echo measurements or diagnosis[1–3,20]. These models show promise in revolutionizing how echocardiology is performed. Because of the large disparity in prevalence of 2D vs 3D echos and the often-proprietary data format of 3D images, AI models in this field are almost exclusively trained and evaluated on 2D TTE images. 2D datasets curated in this way contain only images acquired by human sonographers in specific views and do not span the full distribution of possible echo images.

It is known that machine learning models can perform unpredictably on out of distribution data[21]. Training methods including data augmentations that translate, rotate and resize images attempt to broaden the coverage of the datasets and mitigate these risks. But these augmentations can only simulate the transformation of an image constrained to the 2D plane. Real ultrasound acquisitions can include rotations and translation in 3D. One of the main goals of AI in medicine is the mitigation of human error. For models that do not perform well with 3D view transformations, the performance of the model could be strongly dependent on the sonographer's acquisition quality.

In this research, we propose methods for evaluating AI model performance on off-axis views by introducing realistic 3D spatial transformations to the acquisition plane in 3D volumes. Although 3D echos remain relatively rare in the CSMC system, searching over 16 years, we curated a dataset of 1,528 apical 3D images. Through reverse engineering, we were able to decode the Phillips 3D DICOM data format these images are stored in. We developed functions for slicing 3D data into 2D images and simulating realistic transformations that could be introduced by sonographer motion. We use a deep learning image view classifier, trained specifically for this task, to find the ideal view to compare performance vs. distance from ideal view.

To test these methods, we chose to evaluate the EchoNet-Dynamic model[1] for measurement of left ventricular ejection fraction (LVEF) as the downstream tasks. LVEF is the ratio of the diastolic LV volume to the systolic LV volume as a percentage of volume ejected. It is an important measurement for assessing cardiac function and heart failure[11,22,23]. Typically, LVEF measurements are made by tracing the LV for systolic and diastolic frames in an A4C view video. EchoNet-Dynamic is a ResNet derived regression model that was trained on 144,184 videos from SHC. These images are primarily of the apical-4-chamber and apical-2-chamber views. It has been well validated on external datasets and even a randomized clinical trial[2]. We evaluate the performance of this model on synthetically produced 2D images with simulated probe rotation, translation, and foreshortening to draw conclusions about the robustness of this model in the real world and dependence on view quality.

2. Methods

To realize the impact of this research, several challenges were to be overcome. One of the largest challenges is simply working with 3D echo. To be able to make use of the 3D echo data, we first needed to pull the DICOM images from the hospital dataset, reverse engineer the proprietary data format, and develop tools for interpreting and slicing 3D volumes. The next crucial step was to align the 3D volumes along standard views so that they could be analyzed together. This was done using a view classifier that we trained just for this project. Finally, we evaluate the performance of the EchoNet EF prediction model.

2.1. *Working With 3D Echo*

The 3D echo dataset used in this research is a subset of all of the echos in CSMC's database between 2012 and 2022, nearly 15 million images. Of these images, 1,349 of these are 3D acquisitions taken in the apical position. The apical 3D echos were used because of their ability to generate A4C and A2C 2D views with relatively benign artifacts from the slicing process. All 3D echos were captured on Philips EPIQ CVx ultrasounds. A breakdown of the relative size relevant factors for the CSMC 2D and 3D datasets can be found in Table 1.

Table 1. Breakdown of CSMC image types from 2006-2022

Dataset	N Studies	N Images	Mean EF	Frame Rate	Acquisition Duration
All Echos	369,306	14,922,383	69.10%	26.42 fps	2.75 seconds
3D Apical Echos	1528	1,349	56.26%	18.32 fps	2.81 seconds

Like standard 2D echos, 3D echos are stored in DICOM format. Unlike 2D echo, the data stored in the "pixel data" tag in the DICOMs is only a snapshot of the volume that the sonographer chose to capture and not the full 3D echo data. The full data is stored in a proprietary compressed format under other tags that we were able to reverse engineer. The decompressed data consists of voxel data and physical bounds for the captured volume. Unlike voxel data captured in MRI and 3D formats, this voxel data is not rectilinear - instead it is defined by a spherical coordinate system, as shown in figure Fig. 1. This coordinate system is parameterized by one linear dimension (ρ), and two rotational dimensions (φ and θ). For each of these axes, the physical bounds given in the DICOM define a section of a sphere containing the scanned region that called the frustum. For convenience, we will also be using a 3D cartesian coordinate system with the origin at the probe on the surface of the skin and the x axis pointing parallel to the probe into the body.

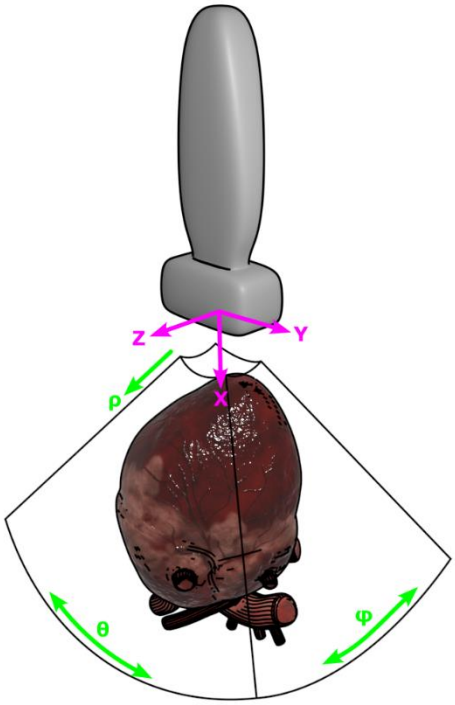

Fig. 1. Diagram showing the 3D world and spherical coordinate systems.

To generate 2D slices of 3D videos, we must first define points on a plane corresponding to the 2D view that we wish to sample. Although there are many degrees of freedom and ways to slice a 3D volume, we decided to constrain our slices to just 4 degrees of freedom to ensure relatively

realistic looking slices and clinical relevance. We first define a square region on the x-y plane centered at the center of the volume and whose width is the max width of the volume to ensure that any slice will be centered and reasonably zoomed. We rotate this plane around the x-axis and then translate it forward or backward through the volume. A translation of 1 corresponds to all the way forward through the volume and -1 corresponds to all the way backward through the volume. Translations of roughly -0.5 to 0.5 result in reasonable slices. Two additional degrees of freedom were added to simulate foreshortening. A horizontal axis is defined on the plane and the slice is rotated forward or backward. We found that an axis location of 30% from the top of the plane to the bottom is reasonable for simulating foreshortening in our dataset.

Once we have defined the plane that we wish to slice, we then define a grid of points on that plane resulting in an array with a shape of (n, m, 3) where the last dimension contains the XYZ location of each point. We then transform these points into spherical coordinates using the following equations resulting in an array with the same shape but whose last dimension contains ρ, φ, θ.

$$\rho = \sqrt{x^2 + y^2 + z^2}$$
$$\varphi = tan^{-1}\left(\frac{z}{x}\right)$$
$$\theta = tan^{-1}\left(\frac{y}{\sqrt{x^2 + z^2}}\right)$$

Eq. 1

Because the spherical coordinates are aligned with the voxel data, we can obtain the voxel indices for each point on the plane by simply renormalizing them using the volume bounds.

$$i = \frac{\rho - \rho_{min}}{\rho_{max} - \rho_{min}}$$
$$j = \frac{\varphi - \varphi_{min}}{\varphi_{max} - \varphi_{min}}$$
$$k = \frac{\theta - \theta_{min}}{\theta_{max} - \theta_{min}}$$

Eq. 2

To generate a 2D image all we need to do is round each index to the nearest integer and lookup its value in the voxel data. Any indices out of bounds of the volume result in an intensity of 0. Although this sampling method works, the relatively low-resolution voxel data results in voxel artifacts due to the relatively low resolution of 3D data. To mitigate this problem, we implemented trilinear interpolation between voxels which results in much smoother images as shown in Fig. 2.

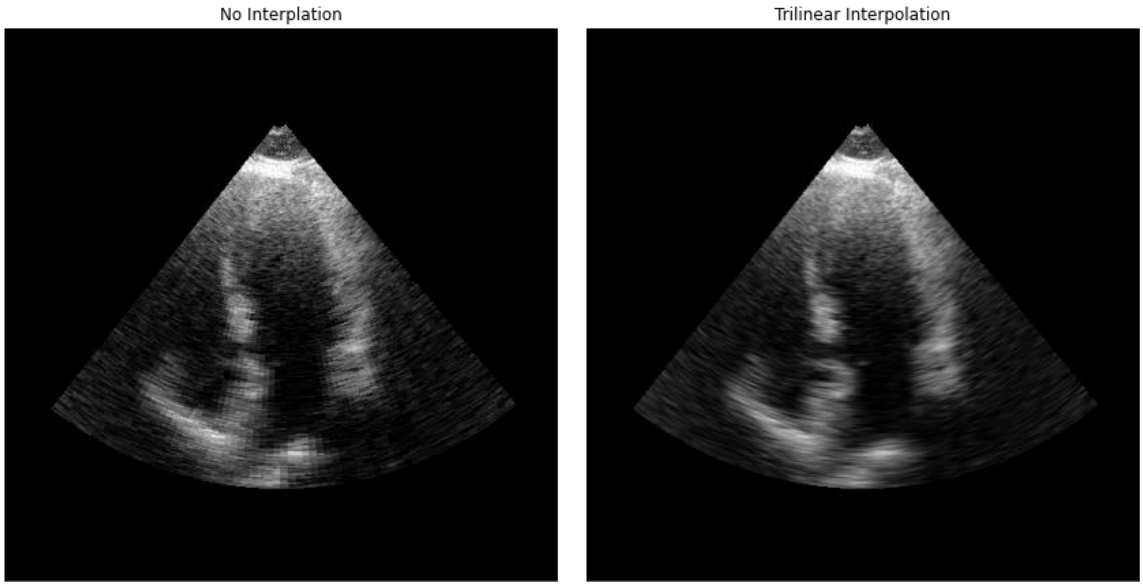

Fig. 2. The impact using trilinear interpolation when generating 2D slices from 3D echo.

2.2. *View Classifier*

With the slicing algorithm that we developed, we are able to accurately simulate the motion of a human moving a probe around a heart, but to characterize a particular view as being a quantifiable rotation and translation away from an optimal view, we need to first define the optimal view. To do this we trained a 2D image view classifier on a standard 2D echo dataset of known standard views. This dataset contains 30,045 echo videos labeled as A4C, A2C, PLAX, Subcostal, or Other views from Stanford Healthcare (SHC). The breakdown of label frequencies can be found in Table 2. During training, random frames are selected from videos in the dataset. Because when running inference on the 3D dataset this model would encounter images unlike anything in the training dataset, we attempted to increase the coverage of the training dataset by adding random mirroring augmentation and additional labels for mirrored A4C, A2C, PLAX and Subcostal. We used a ResNet18[24] image classifier architecture and cross-entropy loss to train the view classifier. The view classifier achieved an AUC of 0.997 for both A4C and A2C views on the SHC test set.

Table 2. Distribution of labels in the view classifier training dataset.

View	N Total	N Train	N Val	N Test
A4C	5,036	4,054	499	483
A2C	3,224	2,577	318	329
PLAX	4,059	3,239	403	417
Subcostal	2,726	2,166	283	276
Other	15,000	12,000	1,500	1,500

2.3. *EF Inference*

The EchoNet model we evaluated has been shown to be accurate on several datasets and even in a randomized clinical trial situation, but it is not known how sensitive it is to small changes in view quality due to poor probe placement and foreshortening.

We addressed this problem by running inference on slices of 3D volumes while varying the rotation, translation and foreshortening from the ideal view. For each 3D volume, we ran both EF and view inference on every combination of translations -0.5 to 0.5 and rotation 0 to 360 degrees. After the best A4C slice, we introduced foreshortening to this view, -40 to 40 degrees, and ran EF inference again. With these results, we were able to draw conclusions about the performance of the EF model as a function of rotation, translation, and foreshortening from the ideal A4C view.

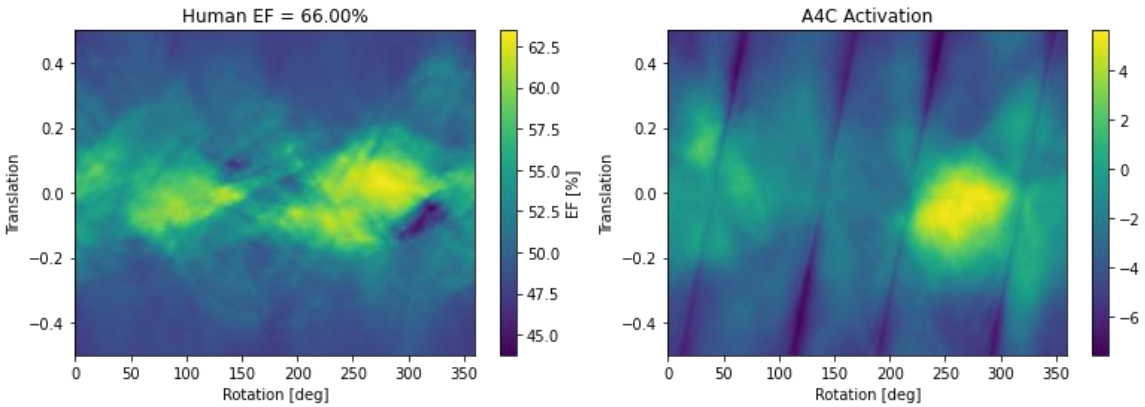

Fig. 3. Phase diagram showing A4C EF prediction and view activation for every combination of rotation and translation. The human measured EF for this patient is 66%.

3. Results

We constrained the slice degrees of freedom to rotation and translation and generated view and EF predictions for every combination of rotation and translation. These predictions were then plotted as a 2D image that summarizes how the model predictions change as the slices are rotated and

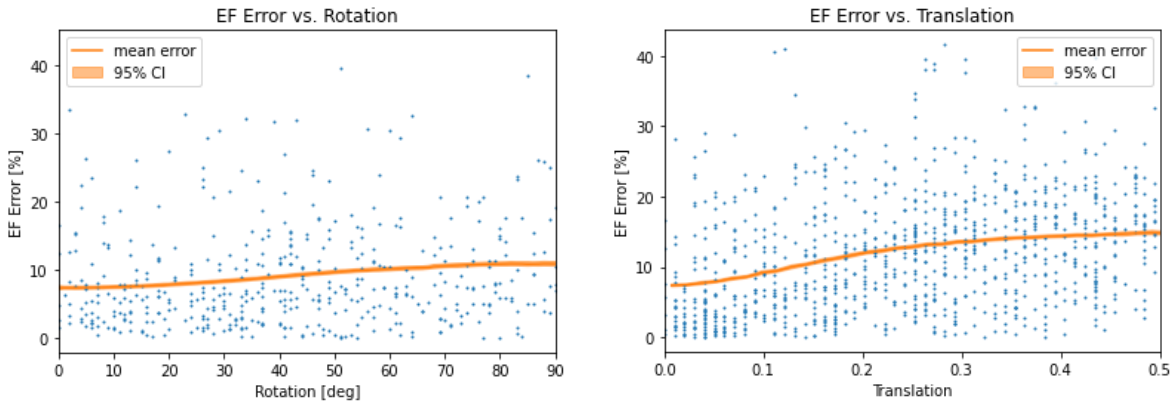

Fig. 4. MAE performance across dataset as view is rotated and translated.

translated shown in Fig. 3. In these plots we can see that regions of high activation for A4C correspond to regions of more accurate EF predictions. The point of maximum A4C activation on this plot for each example is considered to be the optimal view for subsiquent analysis.

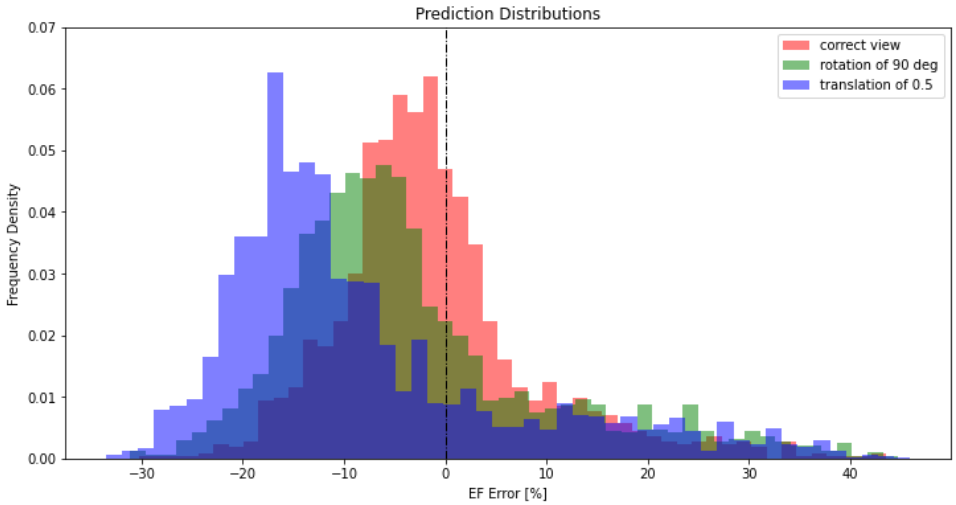

Fig. 5. EF Error distribution for best view, 90 degrees of rotation, and a translation of 0.5.

When EF inference was run on optimal view slices, the mean absolute error (MAE) was 7.3 (7.0-7.7%). Although this is worse than the claimed performance of this model (6.3% comparing model to human or 2.8% comparing model to final value in clinical trial)[2], it is consistent with interobserver variability and the variability between 2D and 3D echo[4,25]. As shown in Fig. 4, when we introduce either rotation or translation to the slice, the error increases. The MAE for rotation increases to 10.9% (10.6-11.1%) while the MAE for translation increases to 14.7% (14.6-14.9%) suggesting that there is more information being used near the center of the volume than near the edges as represented by slices with larger translation error.

One characteristic we noticed was a relatively high frequency of low error, regardless of view quality, especially for patients with near normal EF. This led us to hypothesize that when faced with a poor view, the model makes a guess near the mean of the dataset. We investigate this hypothesis by looking at the prediction trends in various situations. In Fig. 5, we compare the EF prediction distributions of 90-degree rotations and translations of 0.5 to the ideal view slices. We can see that when the view is near ideal, the distribution is relatively tight, and centered around zero. For both introduced rotation and translation, we see that the distributions are shifted to the left, corresponding to underestimates of EF on poorly oriented views. This underestimate cannot be explained by a difference in mean LVEF for the EchoNet training set compared to our 3D dataset. Both datasets have mean LVEF values of roughly 55%[1].

We also analyzed the subset of patients with human measured EF of greater than 70% and the subset of EF less than 30%. In these subsets, the increase in MAE due to introduced rotation and translation is much greater as shown in Fig. 6. This is because for patients with extremely abnormal EF, the model is not able to achieve high accuracy predictions when the view is poor by predicting a value near the mean. For these patients, this effect is stronger than the tendency of the model to underpredict. Therefore, for patients with an LVEF < 30%, the model tends to overpredict EF when the view is poor. An interesting consequence of these two effects is that for low EF patients, there is a threshold where increasing translation decreases error because low EF predictions are nearer to the human measurements for these patients. Fig. 7 illustrates how EF and A4C predictions vary with rotation and translation for a patient with a high human measured EF.

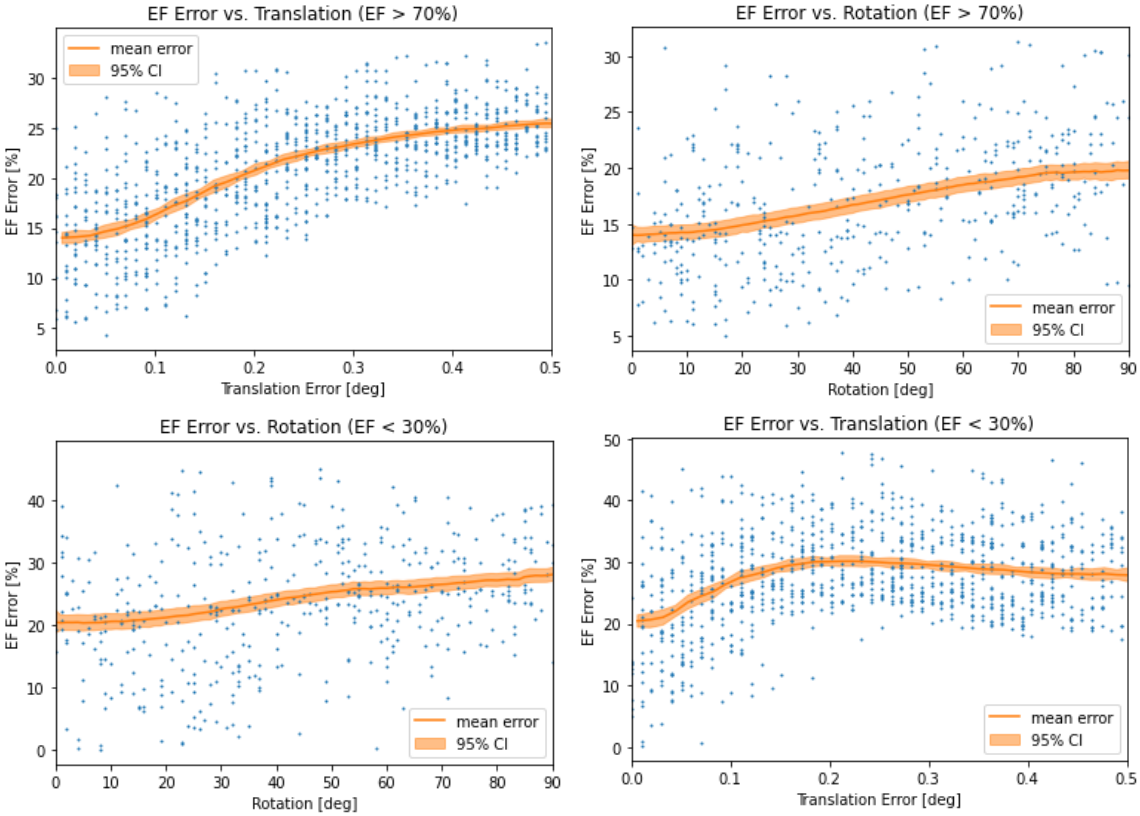

Fig. 6. EF model performance for the >70% and <30% EF subsets.

When looking at foreshortening specifically, we might expect the model to overpredict EF if it calculates EF in the same way as human sonographers, but we see a similar trend as with rotation and translation. This suggests that when predicting EF, the AI model is not segmenting the LV and calculating LV volume to determine EF the way a sonographer would. Fig. shows the results for varying foreshortening from ideal views.

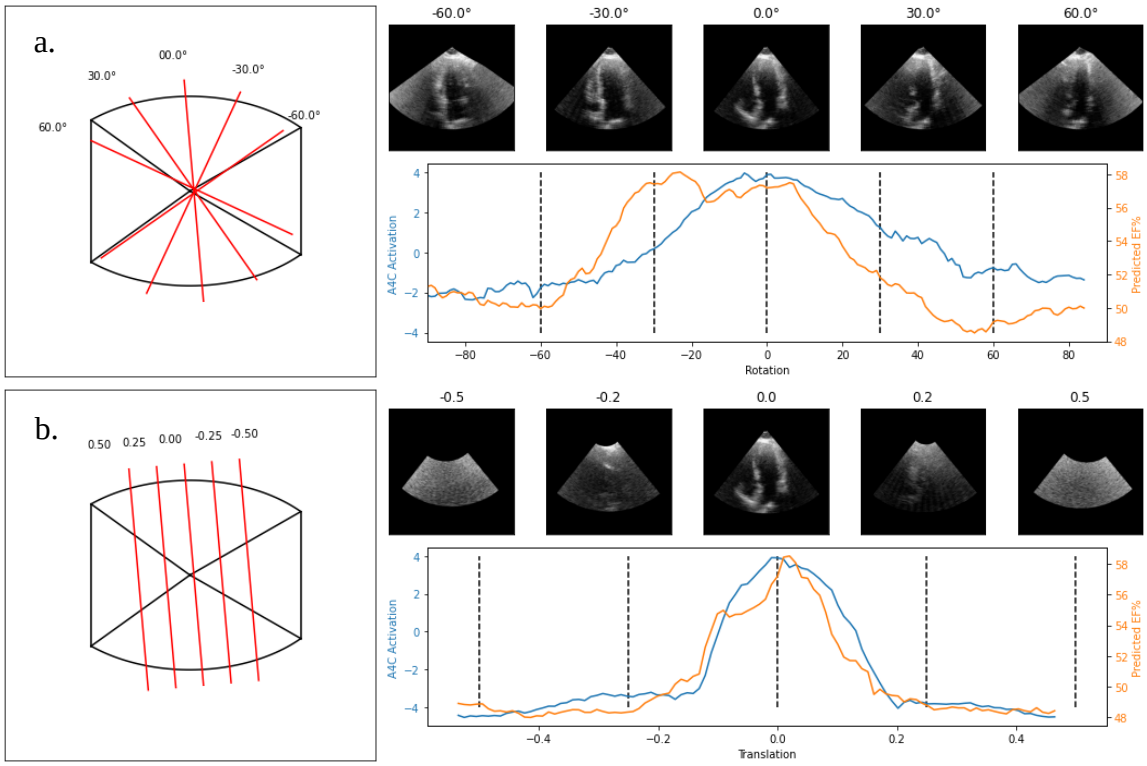

Fig. 7. Example slices and predictions for a range of (a.) rotations and (b.) translations for a selected example volume.

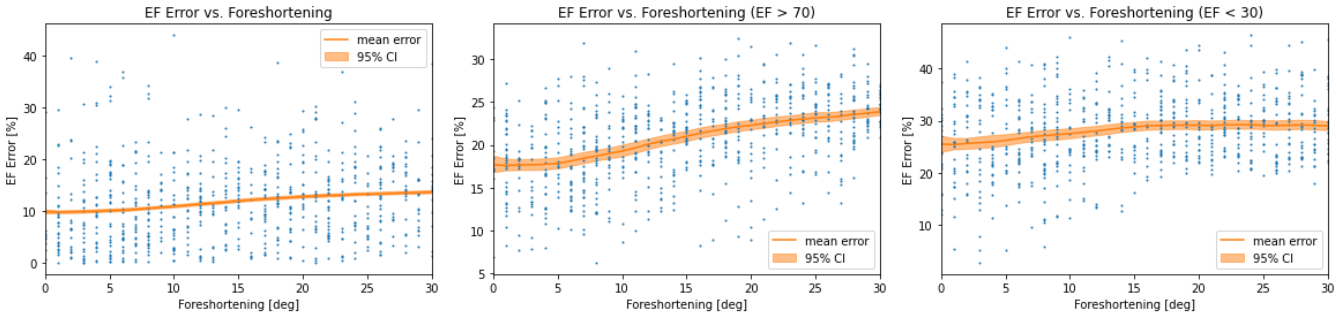

Fig. 8. Performance figures for slices with introduced foreshortening.

4. Discussion

This work demonstrates how 3D echos can be used to evaluate the performance of AI models on realistically out-of-distribution data that these models would likely encounter in real world applications. Understanding distribution shifts and model performance in real world applications may be necessary to understand how AI truly performs in clinical practice, a major barrier in AI research adoption in medicine[26,27]. We presented the methods used for interpreting and utilizing 3D

echo data and evaluated the performance of an established AI model predicting LVEF with these methods.

We found that the EF model we evaluated performed well when the ideal slice is viewed, but error increases as we introduce rotation, translation, or foreshortening. The overall behavior of the model when subjected to OOD data is to guess a value, usually a little below the mean of the training dataset. This overall result of this is a tendency to underestimate EF when the view is poor. The model tends to overpredict EF for patients with very low EF and underestimate EF for patients with very high EF. These trends extent to foreshortening where humans would overestimate EF. Although it makes intuitive sense for the model to guess somewhere near the mean of the dataset when faced with OOD data, the mechanism causing underestimates for OOD data would require further investigation to explain. We hypothesize that the model is gauging the overall amount of motion in the heart to predict EF and for poor views there is a lack of apparent motion, thus the videos look more similar to ones of patients with low EF.

The performance of the EF model even on ideal view slices from 3D echo has lower performance than on 2D videos in prior work. There are several factors that may contribute to this error. First, 3D echo has fundamentally lower spatial and temporal resolution. While the frame rate of standard 2D echos is usually around 30-50 frames per second, 3D echos are much slower, in the range of 13-24 frames per second, with higher framerates associated with lower spatial resolution. Second, the 3D dataset might be comprised of a different distribution of patients than the general population due to selection bias for patients needing additional 3D echos. This is likely, given the average EF of the 3D dataset is 13% lower than the overall CSMC population. Finally, the view classifier we use to find the "ideal" slice is not perfect. It is trained on a dataset of human acquired images that aren't always perfect. Our classifier also only has 4 standard views when in reality there are many more views and several different views may have been grouped together under "A4C". Like the EchoNet model, the view classifier was only trained on 2D images and performance on OOD 3D slices might not be reliable. This would result in the ideal slice for predicting EF not being found.

There is significant opportunity for future research in this field with the use of 3D echo data. An improved view classifier would allow more accurate identification of ideal view orientation. For models trained on clinical 2D datasets, like the EchoNet-Dynamic dataset, it is difficult to quantify the amount of foreshortening and perturbances present. Future work could use 3D echo data to train a model that is able to predict the amount of foreshortening or perturbance in a 2D slice. This would allow us to retrospectively evaluate the view quality and distribution of datasets models are trained on. Additionally, with better tools to simulate and evaluate 3D distribution shifts, there is an opportunity to develop new data augmentations and normalization techniques addressing the spatial nature of echocardiology. Ultimately, as 3D echo data becomes more prevalent, future models could use these techniques to train on 2D slices of 3D data in addition to standard 2D views. These proposed methods would further our understanding and improve the robustness of AI models in echocardiology.

When black box AI models are deployed in healthcare, clinicians may have no sense of whether a model is performing within its operating domain and could lead to either overreliance or mistrust of the AI. In this study, we show how relatively subtle changes to the input data can significantly impact model performance. This has significant impact as with more AI models getting integrated into healthcare systems, it is important to consider how the deployment environment can be different from the environment they were trained and validated in. We show how identifying, simulating, and

evaluating these hypothetical distribution shifts can lead to a better understanding of our AI systems and their performance in the real world.

References

1. Ouyang, D. *et al.* Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).
2. He, B. *et al.* Blinded, randomized trial of sonographer versus AI cardiac function assessment. *Nature* **616**, 520–524 (2023).
3. Johnson, K. W. *et al.* Artificial Intelligence in Cardiology. *J. Am. Coll. Cardiol.* **71**, 2668–2679 (2018).
4. Farsalinos, K. E. *et al.* Head-to-Head Comparison of Global Longitudinal Strain Measurements among Nine Different Vendors: The EACVI/ASE Inter-Vendor Comparison Study. *J. Am. Soc. Echocardiogr.* **28**, 1171–1181, e2 (2015).
5. Hung, J. *et al.* 3D echocardiography: a review of the current status and future directions. *J. Am. Soc. Echocardiogr.* **20**, 213–233 (2007).
6. 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).
7. Feigenbaum, H. Evolution of echocardiography. *Circulation* **93**, 1321–1327 (1996).
8. Ziaeian, B. & Fonarow, G. C. Epidemiology and aetiology of heart failure. *Nat. Rev. Cardiol.* **13**, 368–378 (2016).
9. WRITING COMMITTEE MEMBERS *et al.* 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on practice guidelines. *Circulation* **128**, e240-327 (2013).
10. 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).
11. 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).
12. 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).
13. Foppa, M., Duncan, B. B. & Rohde, L. E. P. Echocardiography-based left ventricular mass estimation. How should we define hypertrophy? *Cardiovasc. Ultrasound* **3**, 17 (2005).
14. Angeli, F. *et al.* Day-to-day variability of electrocardiographic diagnosis of left ventricular hypertrophy in hypertensive patients. Influence of electrode placement. *J. Cardiovasc. Med.* **7**, 812–816 (2006).
15. Ghorbani, A. *et al.* Deep learning interpretation of echocardiograms. *NPJ Digit Med* **3**, 10 (2020).
16. Poon, J., Leung, J. T. & Leung, D. Y. 3D Echo in Routine Clinical Practice - State of the Art in 2019. *Heart Lung Circ.* **28**, 1400–1410 (2019).
17. Ünlü, S. *et al.* Impact of apical foreshortening on deformation measurements: a report from the EACVI-ASE Strain Standardization Task Force. *Eur. Heart J. Cardiovasc. Imaging* **21**, 337–343 (2020).
18. Kim, W.-J. C. *et al.* Automated Detection of Apical Foreshortening in Echocardiography Using Statistical Shape Modelling. *Ultrasound Med. Biol.* **49**, 1996–2005 (2023).
19. Labs, R. B., Zolgharni, M. & Loo, J. P. Echocardiographic image quality assessment using deep neural networks. *arXiv [eess.IV]* (2022).
20. Poplin, R. *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* **2**, 158–164 (2018).

21. Sehwag, V. *et al.* Analyzing the Robustness of Open-World Machine Learning. in *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security* 105–116 (Association for Computing Machinery, 2019).
22. 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).
23. 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).
24. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *arXiv [cs.CV]* (2015).
25. Yuan Neal *et al.* Systematic Quantification of Sources of Variation in Ejection Fraction Calculation Using Deep Learning. *JACC Cardiovasc. Imaging* **0**,.
26. 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).
27. Wu, E. *et al.* How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* **27**, 582–584 (2021).