Self-Supervised Learning of Echocardiogram Videos Enables Data-Efficient Diagnosis of Severe Aortic Stenosis

Gregory Holste [1]   Evangelos Oikonomou [2]   Bobak Mortazavi [3]   Zhangyang Wang [1]   Rohan Khera [2]

Abstract

Given the difficulty of obtaining high-quality labels for medical image recognition tasks, there is a need for deep learning techniques that can be adequately fine-tuned on small labeled data sets. Recent advances in self-supervised learning techniques have shown that such an in-domain representation learning approach can provide a strong initialization for supervised fine-tuning, proving much more data-efficient than standard transfer learning from a supervised pretraining task. However, these applications are not adapted to applications to medical diagnostics captured in a video format. With this progress in mind, we developed a self-supervised learning approach catered to echocardiogram videos with the goal of learning strong representations for downstream fine-tuning on the task of diagnosing aortic stenosis (AS), a common and dangerous disease of the aortic valve. When fine-tuned on 1% of the training data, our best self-supervised learning model achieves 0.818 AUC (95% CI: 0.794, 0.840), while the standard transfer learning approach reaches 0.644 AUC (95% CI: 0.610, 0.677). We also find that our self-supervised model attends more closely to the aortic valve when predicting severe AS as demonstrated by saliency map visualizations.
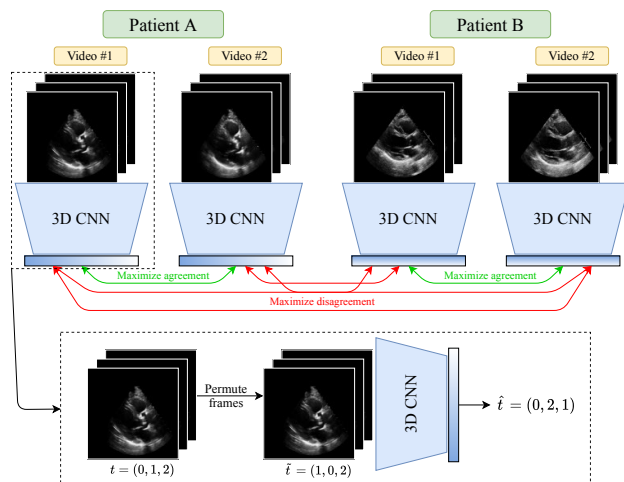
1. Introduction

1.1. Background

Aortic stenosis (AS) is a disease characterized by narrowing of the aortic valve, restricting blood flow and potentially leading to heart failure and even death (Otto & Prendergast, 2014). The standard method for detecting and grading AS

Figure 1. Overview of our proposed self-supervised learning approach for representation learning of echocardiogram videos. The method consists of (1) a contrastive learning objective and (2) a frame re-ordering task. Positive pairs for the contrastive loss are formed between *different* videos from the *same patient*. Frame re-ordering is performed by permuting the frames of an input echo, then training a classifier to predict the original order of frames.

is echocardiography, the practice of capturing ultrasound videos of the heart from a set of canonical views designed to highlight different areas of the heart to assess cardiac function and detect pathology (Baumgartner et al., 2017). However, AS screening is typically performed by Doppler imaging, a specialized type of echocardiography that can additionally quantify the velocity of blood flow (Eugène et al., 2021). While this represents the current standard, this requires specialized equipment that is not widely available outside of large, well-endowed facilities. Particularly, many handheld ultrasound machines now provide broad access to echocardiography without Doppler imaging to clinical practice in remote areas, but without the ability to detect aortic stenosis. Thus, this work aims to automatically diagnose severe AS from single-view two-dimensional (2D) echocardiograms, without any Doppler imaging.

Recently, deep learning techniques have been developed and adapted for a variety of cardiac diseases from 2D

---

[1]The University of Texas at Austin [2]Yale School of Medicine [3]Texas A&M University. Correspondence to: Rohan Khera <rohan.khera@yale.edu>.

echo videos (Ouyang et al., 2020; Ghorbani et al., 2020; Goto et al., 2021). Some studies have successfully applied learning algorithms to detect AS specifically from 2D echo (Duffy et al., 2022; Huang et al., 2021). However, one study (Huang et al., 2021) for example is trained on individual frames, ignoring the temporal aspect of a beating heart that is captured in echocardiography. Broadly, many of these successful applications to echocardiography are trained and evaluated on large labeled data sets that are extremely difficult and expensive to acquire. For this reason, we aim to develop a data-efficient self-supervised learning approach to learn representations of echocardiograms that can serve as a powerful initialization for downstream fine-tuning on very small labeled data sets.

1.2. Related Work

Self-supervised learning (SSL) has proven very effective for label-efficient fine-tuning in natural image classification (Chen et al., 2020; He et al., 2020), video classification (Diba et al., 2021; Kuang et al., 2021), and now even medical image classification and segmentation tasks (Azizi et al., 2021; Taleb et al., 2020; Tang et al., 2021). However, most successful medical applications of SSL operate on 2D data such as histopathological images and radiographs (Chen & Krishnan, 2022; Wang et al., 2021). Some recent studies have developed SSL methods for 3D medical image data, though this has been applied to CT and MRI, where this third dimension is *spatial*, not temporal (Tang et al., 2021; Taleb et al., 2020). That is, medical videos should require special treatment to handle the structure of the temporal content – specifically for echo, representing the spatial translocation of the same cardiac structure over time. It must also be noted that all spatial 3D applications of SSL (in CT and MRI) are in high-resolution modalities with standardized acquisition, making them amenable to the data augmentation required to power most SSL approaches. In contrast, ultrasound is a noisy modality, producing low-resolution images that are extremely sensitive to transformations that require pixel interpolation. Moreover, echos are sensitive to the skill of the individual acquiring the image, with frequent off-axis images containing distortions and artifacts (Le et al., 2016).

1.3. Our Contribution

There are two main challenges to applying self-supervised contrastive learning to echocardiograms:

1. ultrasound is a noisy imaging modality that is brittle to heavy data augmentation
2. echocardiograms contain rich temporal content that most contrastive learning approaches ignore.

To tackle the first issue, we employ "multi-instance" contrastive learning. In our setting, a patient often will have multiple videos captured from a single view type during routine acquisition; we can then leverage these different instances of the same video type for a given patient as positive pairs. Assuming that different videos of the same patient

	Training	Validation	Test
Videos	17,601	2,387	3,460
Studies	5,311	708	1,063
Severe AS	1,184 (22.3%)	160 (22.6%)	246 (23.1%)

Table 1. Summary of training, validation, and test sets. Values in parentheses represent the percentage of studies in each split with a diagnosis of severe AS.

should capture the same underlying characteristics, this (a) forms challenging positive pairs for contrastive representation learning and (b) removes the need for aggressive augmentation to artificially generate two distinct "views" of a patient. To address the second issue of incorporating the temporal nature of echocardiography, we utilize a pretext task of frame re-ordering. For this, we randomly shuffle the frames of an input echo and train a classifier to predict the original order of frames. This imbues the model with a sense of temporal coherence that we observe to aid downstream fine-tuning for severe AS classification and generate more interpretable visual explanations of these predictions.

2. Materials & Methods

2.1. Dataset & Preprocessing

This study includes patients who underwent transthoracic echocardiograms (TTE) – cardiac ultrasounds obtained from the chest wall surface – between the years 2016 and 2020 at the Yale New Haven Hospital. Of all TTE studies from 2016-2020, a sample of 10,000 studies was extracted with severe AS oversampled by a factor of 50. To obtain all echos from the parasternal long axis (PLAX) view, the first and most common view obtained in TTE, an automatic view classifier was employed (Zhang et al., 2018), verified by a clinical expert reader. If a video was most confidently classified as PLAX, it was retained for later preprocessing. Following view classification, the resulting PLAX videos underwent more thorough deidentification by finding the convex hull of the largest contour, masking pixels outside the central image content. After keeping all study with AS labels, the resulting 23,448 videos from 7,082 studies were spatially downsampled to 112×112, then 75% of studies were randomly assigned for training, 10% for validation, and 15% for testing (Table 1). We frame severe AS detection as a binary classification task, where the positive class is severe AS and the negative class consists of all other AS designations including healthy patients.
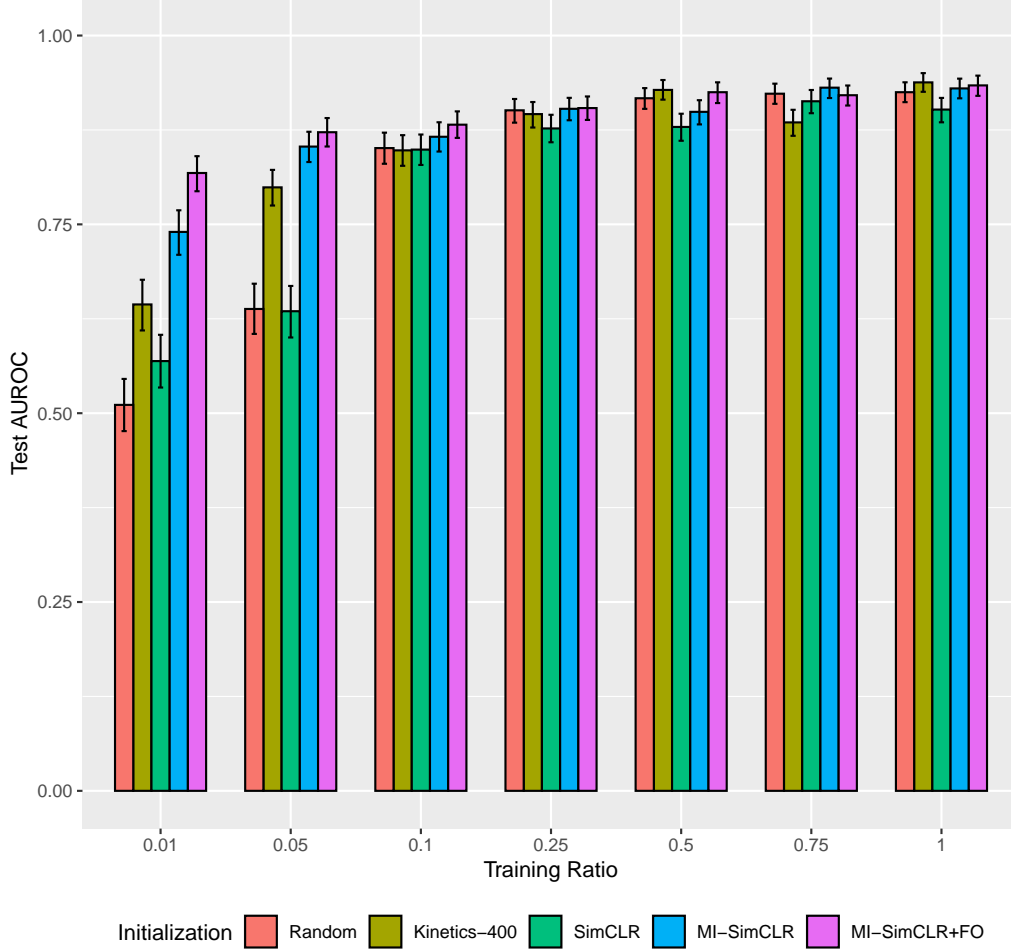
Figure 2. Severe AS classification performance for different weight initializations when fine-tuning on various percentages of the full training set. Performance is measured by AUROC on the test set, and error bars represent 95% confidence intervals obtained by bootstrapping.

2.2. Self-Supervised Learning Approaches

SimCLR. Our first approach to self-supervised learning of echocardiograms is to directly adopt the SimCLR framework (Chen et al., 2020). Specifically, we randomly sample a video clip from an echo, then perform a temporally-consistent set of spatial augmentations to each frame, generating two "views" of the echo. Since ultrasound is noisy relative to natural images and other medical image modalities such as tomography and MRI, we are limited in the level of augmentation we can apply without destroying valuable signal. For this reason, augmentations are limited to random zero-padding on all sides by up to 8 pixels with probability 1, a random horizontal flip with probability 0.5, and a random rotation within -10 to 10 degrees with probability 0.5.

MI-SimCLR. Since the images included in ultrasound videos are noisy, the typical aggressive augmentation

scheme for SimCLR will destroy valuable signal leading to poor representation learning. Borrowing from the findings of one study (Azizi et al., 2021) and using the fact that a patient will often have several PLAX videos, we instead form positive pairs between all unique combinations of *distinct* PLAX videos for the same patient. This "**m**ulti-**i**nstance" **SimCLR** (MI-SimCLR) operates under the hypothesis that different PLAX videos of the same patient should capture the same underlying qualities of the patient. Additionally, this approach should provide more robust features for contrastive learning since it is more challenging to learn similar representations of truly distinct videos rather than augmented views of the same video.

MI-SimCLR+FO. While MI-SimCLR is better equipped for echocardiography than SimCLR in that it removes the need for aggressive augmentation, this method completely ignores the temporal content in echo videos. In this final
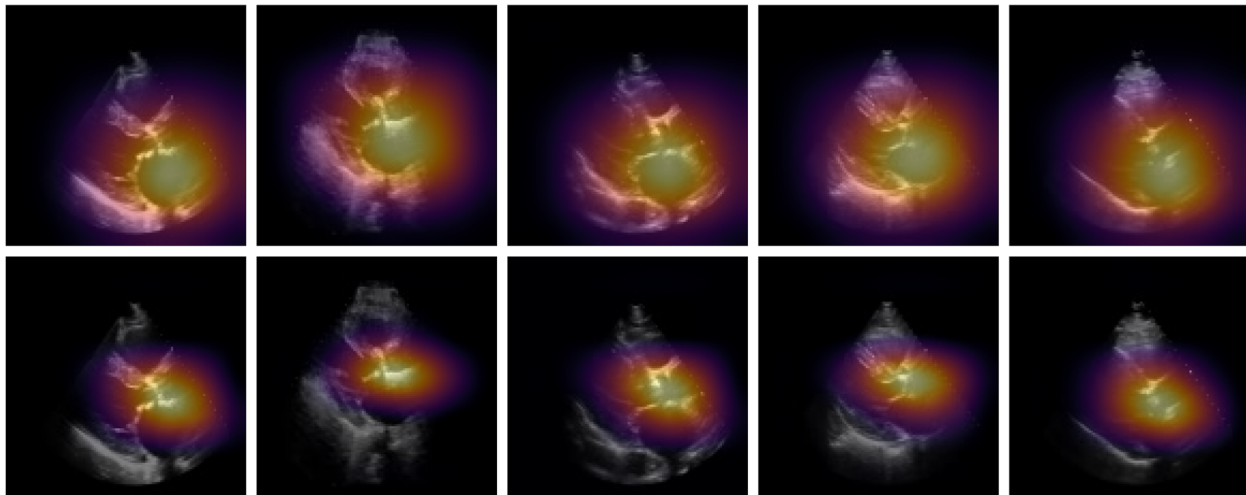
Figure 3. Saliency map visualizations for the five most confident severe AS predictions as determined by the Kinetics-400-pretrained model fine-tuned on all training data. The top row shows saliency maps from the Kinetics-400-pretrained model, and the bottom row comes from the proposed MI-SimCLR+FO model. While saliency maps were computed for entire video clips, only the first frame of each video is displayed above in each column. Saliency maps were obtained by applying Grad-CAM (Selvaraju et al., 2017) to the first 32 frames of each video, then computing the pixelwise maximum along the temporal axis to obtain a single 2D heatmap.

model, **m**ulti-**i**nstance **SimCLR** with **f**rame re-**o**rdering (MI-SimCLR+FO), we incorporate a pretext task that involves randomly shuffling each echo video clip and training the model to predict the correct frame order. Similar to another ultrasound SSL study (Jiao et al., 2020), if a video clip consists of K frames, this is treated as a $K!$-way classification problem and implemented with a simple fully-connected classification head that takes in the learned representation of each echo. This pretraining method allows the fine-tuned model to inherit a sense of temporal coherence more appropriate for the medical video domain.

2.3. Experimental Setup

For downstream supervised fine-tuning on the task of severe AS diagnosis, we evaluate five different weight initializations: **Random**, **Kinetics-400**, **SimCLR**, **MI-SimCLR**, and **MI-SimCLR+FO**. These approaches are used to initialize a 3D ResNet18 (Tran et al., 2018), which is then fine-tuned to classify severe AS with a standard binary cross-entropy loss. Fine-tuning is performed on varying percentages of the full set of training studies (1%, 5%, 10%, 25%, 50%, 75%, and 100%) to understand how the different initializations perform in data-scarce settings. Performance is measured by area under the receiver operating characteristic curve (AUROC). For full implementation details, please see our code repository: https://github.com/CarDS-Yale/echo-ssl-aortic-stenosis.

3. Results

3.1. Fine-tuning Experiments

When fine-tuning on a small subset of labeled data (10% and below), multi-instance self-supervised pretraining initializations consistently outperform all other initialization methods (Figure 2). For example, after fine-tuning on 1% of training studies (53 studies), the proposed MI-SimCLR+FO achieves 0.818 AUC (95% CI: 0.794, 0.840), while the standard transfer learning approach of using a Kinetics-400 initialization reaches 0.644 AUC (95% CI: 0.610, 0.677). Similarly, on 5% of training studies, MI-SimCLR+FO reaches 0.872 AUC (95% CI: 0.853, 0.891), while MI-SimCLR achieves 0.853 AUC (95% CI: 0.833, 0.873) and Kinetics-400 0.799 AUC (95% CI: 0.775, 0.822).

Overall, we find that naively adopting SimCLR to this setting is inadequate: SimCLR alone performs worst on all training ratios except one. However, both MI-SimCLR and MI-SimCLR+FO are the top two methods when fine-tuning on less than or equal to 10% of studies. In other words, *using different videos from the same patient to construct positive pairs is critical for effective representation learning of echocardiograms*.

We additionally find that the frame reordering task provides downstream benefits for severe AS diagnosis; namely, MI-SimCLR+FO outperforms MI-SimCLR for all training ratios except one. However, there in general appears to be a saturation point at 25% of the full training cohort (1,327 studies), at which point there is no clear advantage to any specific initialization approach.

3.2. Interpretability Analysis

We observe that both the standard transfer learning model and our self-supervised model generally focus on the aortic valve and left atrium – clinically relevant abnormalities – when predicting severe AS (Figure 3). However, our final SSL model, MI-SimCLR+FO, consistently produces more spatially localized visual explanations, more closely attending to the aortic valve than the transfer learning model.

4. Conclusion & Discussion

We have presented a novel self-supervised learning method for echocardiogram videos, demonstrating its effectiveness for downstream fine-tuning on severe AS diagnosis with very little labeled data. Our proposed method MI-SimCLR+FO is specifically designed to tackle the challenges and opportunities presented by echocardiography: 1) it contrasts multiple videos from the same PLAX view of a given patient to avoid destroying signal via augmentation, and 2) it enforces temporal coherence via a frame re-ordering pretext objective. This work is among the first to apply self-supervised learning techniques to medical video data, and provides a promising avenue toward fine-tuning echo video-based AI models on small labeled datasets.

References

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3478–3488, 2021.

Baumgartner, H., Hung, J., Bermejo, J., Chambers, J. B., Edvardsen, T., Goldstein, S., Lancellotti, P., LeFevre, M., Miller Jr, F., Otto, C. M., et al. Recommendations on the echocardiographic assessment of aortic valve stenosis: a focused update from the european association of cardiovascular imaging and the american society of echocardiography. *European Heart Journal-Cardiovascular Imaging*, 18(3):254–275, 2017.

Chen, R. J. and Krishnan, R. G. Self-supervised vision transformers learn visual concepts in histopathology. *arXiv preprint arXiv:2203.00585*, 2022.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Diba, A., Sharma, V., Safdari, R., Lotfi, D., Sarfraz, S., Stiefelhagen, R., and Van Gool, L. Vi2clr: Video and image for visual contrastive learning of representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1502–1512, 2021.

Duffy, G., Cheng, P. P., Yuan, N., He, B., Kwan, A. C., Shun-Shin, M. J., Alexander, K. M., Ebinger, J., Lungren, M. P., Rader, F., et al. High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. *JAMA cardiology*, 7(4):386–395, 2022.

Eugène, M., Duchnowski, P., Prendergast, B., Wendler, O., Laroche, C., Monin, J.-L., Jobic, Y., Popescu, B. A., Bax, J. J., Vahanian, A., et al. Contemporary management of severe symptomatic aortic stenosis. *Journal of the American College of Cardiology*, 78(22):2131–2143, 2021.

Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J. H., Harrington, R. A., Liang, D. H., Ashley, E. A., and Zou, J. Y. Deep learning interpretation of echocardiograms. *NPJ digital medicine*, 3(1):1–10, 2020.

Goto, S., Mahara, K., Beussink-Nelson, L., Ikura, H., Katsumata, Y., Endo, J., Gaggin, H. K., Shah, S. J., Itabashi, Y., MacRae, C. A., et al. Artificial intelligence-enabled fully automated detection of cardiac amyloidosis using electrocardiograms and echocardiograms. *Nature communications*, 12(1):1–12, 2021.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Huang, Z., Long, G., Wessler, B., and Hughes, M. C. A new semi-supervised learning benchmark for classifying view and diagnosing aortic stenosis from echocardiograms. In *Machine Learning for Healthcare Conference*, pp. 614–647. PMLR, 2021.

Jiao, J., Droste, R., Drukker, L., Papageorghiou, A. T., and Noble, J. A. Self-supervised representation learning for ultrasound video. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1847–1850. IEEE, 2020.

Kuang, H., Zhu, Y., Zhang, Z., Li, X., Tighe, J., Schwertfeger, S., Stachniss, C., and Li, M. Video contrastive learning with global context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3195–3204, 2021.

Le, H. T., Hangiandreou, N., Timmerman, R., Rice, M. J., Smith, W. B., Deitte, L., and Janelle, G. M. Imaging artifacts in echocardiography. *Anesthesia & Analgesia*, 122(3):633–646, 2016.

Otto, C. M. and Prendergast, B. Aortic-valve stenosis—from patients at risk to severe valve obstruction. *New England Journal of Medicine*, 371(8):744–756, 2014.

Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C. P., Heidenreich, P. A., Harrington, R. A., Liang, D. H., Ashley, E. A., et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580 (7802):252–256, 2020.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., and Lippert, C. 3d self-supervised methods for medical imaging. *Advances in Neural Information Processing Systems*, 33:18158–18172, 2020.

Tang, Y., Yang, D., Li, W., Roth, H., Landman, B., Xu, D., Nath, V., and Hatamizadeh, A. Self-supervised pre-training of swin transformers for 3d medical image analysis. *arXiv preprint arXiv:2111.14791*, 2021.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Huang, J., Yang, W., and Han, X. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 186–195. Springer, 2021.

Zhang, J., Gajjala, S., Agrawal, P., Tison, G. H., Hallock, L. A., Beussink-Nelson, L., Lassen, M. H., Fan, E., Aras, M. A., Jordan, C., et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation*, 138(16):1623–1635, 2018.