1 ViViEchoformer: Deep Video Regressor Predicting Ejection Fraction

Taymaz Akan¹, Sait Alp², Md. Shenuarin Bhuiyan³, Tarek Helmy¹, A. Wayne Orr^{3,4}, Md. Mostafizur
 Rahman Bhuiyan⁵, Steven A. Conrad¹, John A. Vanchiere^{1,6}, Christopher G. Kevil^{3,4}, Mohammad A. N
 Bhuiyan^{1*}

 ¹Department of Medicine, Louisiana State University Health Sciences Center at Shreveport, Shreveport, LA 71103, USA;
 ²Department of Computer Engineering, Erzurum Technical University, Erzurum, Turkey;
 ³Department of Pathology and Translational Pathobiology, Louisiana State University Health Sciences Center at Shreveport, LA 71103, USA;
 ⁴Department of Molecular and Cellular Physiology, Louisiana State University Health Sciences Center at Shreveport, Shreveport, LA 71103, USA;
 ⁵Department of Pediatric Cardiology, Bangabandhu Sheikh Mujib Medical University, Bangladesh, ⁶Department of Pediatrics, Louisiana State University Health Sciences Center at Shreveport, LA 71103, USA;

11 Correspondence: Mohammad Alfrad Nobel Bhuiyan, Ph.D., Department of Medicine, Louisiana State University Health Sciences 12 Center at Shreveport, PO Box 33932, Shreveport, LA 71130- 3932. Email: <u>Nobel.Bhuiyan@lsuhs.edu</u>

Heart disease is the leading cause of death worldwide, and cardiac function as measured by ejection fraction (EF) is an important determinant of outcomes, making accurate measurement a critical parameter in PT evaluation. Echocardiograms are commonly used for measuring EF, but human interpretation has limitations in terms of intra- and inter-observer (or reader) variance. Deep learning (DL) has driven a resurgence in machine learning, leading to advancements in medical applications. We introduce the ViViEchoformer DL approach, which uses a video vision transformer to directly regress the left ventricular function (LVEF) from echocardiogram videos. The study used a dataset of 10,030 apical-4-chamber echocardiography videos from patients at Stanford University Hospital. The model accurately captures spatial information and preserves inter-frame relationships by extracting spatiotemporal tokens from video input, allowing for accurate, fully automatic EF predictions that aid human assessment and analysis. The ViViEchoformer's prediction of ejection fraction has a mean absolute error of 6.14%, a root mean squared error of 8.4%, a mean squared log error of 0.04, and an R^2 of 0.55. ViViEchoformer predicted heart failure with reduced ejection fraction (HFrEF) with an area under the curve of 0.83 and a classification accuracy of 87 using a standard threshold of less than 50% ejection fraction. Our video-based method provides precise left ventricular function quantification, offering a reliable alternative to human evaluation and establishing a fundamental basis for echocardiogram interpretation.

13 **INTRODUCTION**

14 Cardiovascular diseases (CVDs) encompass a range of conditions that can negatively impact the health of the cardiovascular system, which consists of the heart and blood vessels. CVDs are consistently 15 ranked as one of the top causes of death worldwide ¹. Heart failure (HF) is a rapidly growing 16 cardiovascular condition, with an estimated prevalence of 37.7 million individuals worldwide. HF is a 17 18 chronic phase of cardiac functional impairment, causing symptoms such as dyspnea, fatigue, poor exercise tolerance, and fluid retention, which impact patients' quality of life and contribute to the global 19 health crisis². It also carries a high mortality rate. Diagnosing heart failure requires an accurate 20 21 assessment of cardiac function, which can be done using various methodologies to guantify and 22 characterize. Left ventricular EF is one of the most important metrics for assessing cardiac function, which 23 measures how well the left ventricle can eject blood 3,4 .

Standard methods for estimating left ventricular ejection fraction include echocardiograms, cardiac MRI, cardiac computed tomography (CT), and Equilibrium Radionuclide Angiocardiography (ERNA). Echocardiography uses ultrasound to create real-time images of the heart's chambers, valves, and blood flow, assessing the volume of blood pumped out of the left ventricle with each contraction. MRI provides detailed images of the heart's structure and function but has limitations such as cost, availability, and potential contraindications. CT uses X-rays to produce detailed heart images but has limitations such as radiation exposure, allergic reactions to contrast media, and limited dynamic heart function assessment.

Equilibrium radionuclide angiography is a method used in nuclear medicine studies. Still, it has some drawbacks, like taking a long time to process, injecting radiopharmaceutical agents, and yielding low resolution for regional ventricular function in heart disease patients ^{3,5}. Clinically echocardiography is the preferred most common method for estimating LVEF because it is widely available, provides real-time imaging, is non-invasive, and is more cost-effective than other options. This makes it particularly useful for quick and detailed assessments in various clinical situations.

37 Traditional echocardiography typically includes a visual interpretation to estimate LVEF, providing a 38 qualitative assessment without precise numerical values. This approach is well-suited for managing acute 39 patients but falls short when it comes to serial evaluations, particularly in patients with valvular lesions 40 causing regurgitation. There are also quantitative capabilities for echocardiography using the Simpson's 41 method and fractional shortening to calculate EF. The human calculation of ejection fraction is subject to 42 variability due to irregular heart rate and the nature of the calculation, which necessitates manual ventricle size tracing for every beat⁴. The variability in estimating LVEF among different observers can often result 43 44 in requests for additional testing, review of the study, and reinterpretation, which can impact the timing of 45 therapeutic interventions 5-7.

46 Conventional Machine learning (ML) has recently led to substantial advancements in diverse fields, 47 including medical applications. Conventional ML has been utilized in echocardiography to determine the 48 ejection fraction, with significant interest in their potential to provide improvement in disease diagnoses, aid decision-making, and serve as a confirmatory assessment ^{8,9}. However, conventional ML has a 49 potential disadvantage in its reliance on feature engineering, which is a manual and time-consuming 50 process. Moreover, despite obtaining images in various positions and orientations, these conventional 51 52 echocardiographic systems lack 3D localization and spatial relation measurements for volume 53 computation.

Deep learning has driven a significant resurgence in machine learning due to availability of large data 54 sets, and advances in computing power ¹⁰⁻¹⁵. This field has revolutionized machine learning by 55 understanding and manipulating data, including images ^{16,17}, and incorporation of natural language 56 processing (NLP)¹⁸. Moreover, deep learning differs from conventional methods as it avoids manual 57 feature engineering. Also, using deep learning techniques in medical diagnostics improves the accuracy 58 of diagnoses ^{19,20}. It plays a crucial role in predictive analytics, allowing for detecting possible health risks 59 or outcomes ²¹. These techniques provide healthcare professionals with valuable predictive insights 60 through the assimilation and analysis of various datasets, including patient information, genetic profiles. 61 imaging studies, and clinical records, and enables early detection of diseases or health deterioration ²². 62

63 Deep learning techniques can be used to determine ejection fraction, estimate end-diastole and endsystole volumes, and calculate the percentage difference between them, rather than relying on actual 64 echocardiogram videos ²³⁻²⁷. A recently proposed method, EchoNet-Dynamic ⁴, directly regresses LVEF 65 66 from video inputs using spatiotemporal models, which avoids the need to estimate EDV and ESV 67 separately. EchoNet-Dynamic, a video-based deep learning model, has been proposed for echocardiograms, demonstrating its ability to assess ejection fraction accurately across the entire video. It 68 is a CNN model that uses atrous convolution ²⁸ for semantic segmentation of the left ventricle, a CNN 69 model²⁹ with residual connections and spatiotemporal convolutions for predicting the ejection fraction, 70 71 and video-level predictions for beat-to-beat estimations of cardiac function. Moreover, another video-72 based method performs LV segmentation using echocardiogram sequences and then converts the predicted context into an end-to-end video regression model ³⁰. However, segmentation, a sensitive 73 process involving categorizing entire regions, may increase computational requirements and processing 74 75 times. Inaccuracies in segmentation can impact subsequent classification or regression tasks, making the 76 overall process more sensitive to segmentation quality. Recent advances in deep learning have shown 77 that it can accurately and reproducibly identify human-identifiable phenotypes and characteristics not recognized by human experts, overcoming limitations in human interpretation ^{31–33}. 78

Parameter name	Values
Hyperparameters of ViViT	
Optimizer	SGD
Batch size	128
Epoch	100
Input Shape	(52, 52, 32, 1)
Layer norm	1e-6
Learning rate	1e-4
Number of heads	12
Number of Layers	10
Patch size	(32, 8, 8)
Projection dim	512

Table 1. Details of model variants

79 Herein, we propose an end-to-end deep learning approach, ViViEchoformer, which leverages a video

80 vision transformer (ViViT) ³⁴ to regress LVEF from echocardiogram videos directly. We converted ViViT

81 from classification to regression to predict LVEF. The model captures spatial information and preserves

82 inter-frame relationships by extracting spatiotemporal tokens from the input video. While utilizing the

83 video vision transformer to capture spatiotemporal patterns in the video accurately, this method performs

84 precise, fully automatic EF predictions that facilitate human assessment and subsequent analysis.

85 RESULTS

86 Our neural network architecture was implemented in Python using the TensorFlow and Keras libraries. A

87 workstation equipped with 62 GB of RAM and an NVIDIA GeForce GTX 4080 GPU was used for all

experiments. We trained our transformer model (**Fig 1**) on a data set with 10,030 echocardiogram videos

89 provided by Stanford University Hospital ³⁵. We converted the classification model ViViT, into a regression

90 model and trained it to estimate the left ventricular ejection fraction from echocardiogram videos using a

training and validation set of over 30700 and 1200 videos, respectively, and a test set of over 1200

92 videos. The analysis focused on the 32 frames of videos that were resized to 52x52 dimensions.



Fig 1. The model pipeline for video regression. The Tubelet embedding technique extracts and linearly embeds nonoverlapping tubelets across the spatio-temporal input volume. Using spatial-temporal attention, the transformer encoder forwards all spatio-temporal tokens extracted from the video.



Fig 2. The graph illustrates the model's loss over epochs for training (blue) and validation (orange) datasets.

We employed the SGD optimizer for training, and the training process was conducted over 100 epochs with a batch size of 128 and a learning rate of 1e-4. **Table 1** provides a summary of the configuration of the training parameters. The model checkpoint is configured to save only the optimal solution discovered during training based on the loss function evaluation during validation. The model checkpoint is saved when a metric improves on a validation set during training. As depicted in **Fig 2**, the model demonstrates a significant reduction in loss in the initial epochs, which indicates the model's capacity to learn quickly from the training data.

100 We have employed the evaluation metrics for evaluating the performance of ViViEchoformer on the 101 EchoNet test dataset, which were not previously used during model training. The estimation of EF has 102 been associated with interobserver variability of up to 14% ³⁶. The ViViEchoformer's prediction of ejection 103 fraction had a mean absolute error of 6.14%, root mean squared error of 8.4%, mean squared log error of 104 0.04, and an R^2 of 0.55.

105 The visual assessment has been carried out using six plots (Fig 3). These plots are used to evaluate the 106 performance of a predictive model, providing information about the accuracy, distribution of errors, and 107 independence of errors, which are crucial for validating the robustness of the model. The scatter plot Fig 108 3a shows the model's predictions against the actual values, with points scattered around the line of 109 perfect agreement. This indicates that the model captures the trend in the data, but the spread of points 110 away from the line indicates variances in prediction accuracy. The violin plot and histogram of error 111 distribution Fig 3b, c provide insight into the distribution of prediction errors, with a long tail of errors 112 indicating a right-skewed distribution. The line plot of errors in Fig 3d shows variability, with most falling 113 within two standard deviations of the mean. However, occasional spikes beyond this range suggest more 114 significant errors, possibly due to outliers or less valid assumptions. The autocorrelation and partial 115 autocorrelation plots in Fig 3e, f show that the errors are mostly independent, indicating a positive 116 predictive model performance.

117 Table 2 reports the model's classification performance distinguishing between Heart Failure with Reduced 118 Ejection Fraction (HFrEF) and Non-HFrEF cases. Precision, recall, f1-score, and support numbers are 119 reported for both categories. The classification report shows ViViEchoformer's prediction of HFrEF with an 120 area under the curve of 0.83 (Fig 4a), using a common threshold of an EF of less than 50%. The model 121 achieves a precision of 0.77 for HFrEF cases, indicating 77% correctness, and a recall of 0.83, indicating 122 83% correct identification. The f1-score balances these metrics, indicating the model's effectiveness in 123 HFrEF cases. However, the model performs better for non-HFrEF cases, with a precision of 0.91 and 124 recall of 0.92, resulting in a higher f1-score of 0.89. The overall accuracy across both classes is 0.87, 125 indicating 87% correct classifications. The macro average f1-score is 0.83, considering the balance between classes without weighting for their representation in the dataset. The weighted average f1-score 126 127 is also 0.87, indicating consistent high performance across classes when accounting for the number of 128 samples in each. Fig 4b illustrates the confusion matrix for our model's classification performance where



HFrEF is labeled 0, and Non-HFrEF is labeled 1. The matrix visually represents the model's predictionscompared to the actual labels.



Fig 3. Model Evaluation. a Comparison of ViViEchoformer predicted, and EchoNet dataset reported ejection fractions (n = 1288). **b** the violin plot showcasing the model error distribution. **c** errors distribution histogram. **d** error values across samples. **e** Autocorrelation plot of residuals. **f** partial autocorrelation of the residuals.

131 DISCUSSION

132 The most prominent architecture of choice in sequence modeling is the transformer, which uses a multi-133 headed self-attention mechanism instead of convolution. ViViEchoformer is a video transformer-based 134 deep learning model for echocardiogram video understanding tasks, allowing for accurate, fully automatic 135 EF predictions that aid human assessment and analysis. To our knowledge, ViViEchoformer is the first 136 deep-learning model that uses transformers to estimate the ejection fraction from echocardiogram videos. 137 Previous attempts to use deep learning techniques are primarily used to determine EFs, end-diastole and end-systole volumes, and percentage differences in echocardiogram videos rather than actual data. 138 139 These methods typically do not account for inter-frame relationships or temporal dependencies within the 140 video sequences during their analysis. To process video sequences, ViViEchoformer splits them up into 141 smaller temporal and spatial units known as tokens. The model can then comprehend temporal dependencies throughout the sequence and spatial relationships within individual frames thanks to 142 143 extracting and processing information from these tokens across frames.

144 Some video-based methods perform LV segmentation using echocardiogram sequences, but 145 segmentation may increase computational requirements and processing times due to its sensitive nature. However, when analyzing massive datasets, DL techniques can reveal hidden patterns that were 146 147 previously not apparent. Recent advancements in DL techniques have demonstrated their ability to "see 148 the unseen" in images and videos. Consequentially, determining EFs without end-diastole and end-149 systole volumes could be possible for DL techniques. Without infusing knowledge awareness and using 150 any pre-processing, such as segmentation, our method directly regresses EF among the video frames. ViViEchoformer's predictions have a variance comparable to or less than human experts' measurements 151 of cardiac function ³⁷. ViViEchoformer achieved high prediction accuracy for estimating ejection fraction 152 153 performed by human interpreters. Its prediction of ejection fraction had a mean absolute error of 6.14%, which is within the typical inter-observer variation of 14%. 154

	precision	recall	f1-score	support	
HFrEF	0.77	0.74	0.75	285	
Non-HFrEF	0.91	0.92	0.71	794	
accuracy			0.87	1079	
macro avg	0.84	0.83	0.83	1079	
weighted avg	0.87	0.87	0.87	1079	

Table 2. Classification performance for HFrEF and Non-HFrEF cases



Fig 4. Classification accuracy of 85 and AUC of 0.83 using a standard threshold of less than 50% EF. a ROC curve, b Confusion matrix where HFrEF, labeled as 0 and non-HFrEF labeled 1.

In the study by Ouyang et al.⁴, five expert sonographers and cardiologists conducted a blinded review of 155 156 echocardiogram videos that exhibited the largest absolute differences between the initial human labels and the predictions made by EchoNet-Dynamic. These experts independently assessed the relevant 157 videos and two blinded measurements of ejection fraction. The findings revealed that 38% (15 out of 40) 158 of the videos had significant issues related to video quality or the acquisition process. In comparison, 13% 159 (5 out of 40) were characterized by marked arrhythmias, which constrained the experts' capacity to 160 161 assess ejection fraction accurately. A critical limitation of the EchoNet-Dynamic dataset stems from the 162 inaccuracy in the initial human labeling of echocardiogram videos, compounded by issues related to poor 163 image quality, arrhythmias, and variations in heart rate. These factors significantly impact on the training and evaluation performance of our model. 164

165 In developing a model to regress the left ventricular ejection fraction (LVEF) from echocardiogram videos, 166 we encountered a nuanced issue at the intersection of statistical significance and clinical utility, 167 particularly when classifying LVEF based on the 50% cutoff. Our model is capable of closely 168 approximating actual LVEF values. Yet, we observe instances where minor discrepancies-such as a predicted LVEF of 49.9% versus an actual measurement of 50.01%-raise important considerations. 169 While these small differences may be statistically significant, they highlight the clinical uncertainty of near-170 171 threshold predictions in model evaluation. This distinction is important because, in clinical practice, the 172 marginal difference may not change treatment or patient outcome, calling statistically significant but 173 clinically marginal model predictions into question. This is a limitation for most methodologies and should 174 be acknowledged.

175 Fig 5 presents a scatter plot evaluating the performance of a regression model that predicts left 176 ventricular ejection fraction (LVEF). The true LVEF values are on the X-axis, while the Y-axis displays the 177 model's predicted LVEF values. The overlay of a green zone and an orange area indicates the boundary 178 of correct and incorrect classifications by the model relative to the critical threshold of 50%. The green 179 zone indicates regions where the model's predictions align correctly with the true classifications-180 predictions of LVEF less than 50% that are indeed below 50% (lower left) and predictions above 50% that are actually above 50% (upper right). Conversely, the orange zone indicates regions of misclassification-181 182 predictions above 50% for true values below 50% (lower right) and vice versa (upper left). Central to the 183 plot is a highlighted square around the 50% line, visually representing the area of uncertainty where the 184 model's predictions are close to the threshold, encapsulating the challenge of near-threshold predictions. This zone of uncertainty underscores the difficulty in achieving precise classifications around the 50% 185 186 cutoff point, which is critical for clinical decision-making based on LVEF values.



Fig 5. Scatter plot of true vs. predicted LVEF values by the regression model, illustrating classification accuracy relative to the 50% threshold. The green area represents correct classifications, while the orange area signifies incorrect classifications. The highlighted square around the 50% line delineates the zone of uncertainty, emphasizing the model's challenge in making near-threshold predictions.

The study suggests that future research could focus on developing advanced classification models to identify videos with poor image quality, arrhythmias, and heart rate variations. This would improve the reliability of automated assessments by reducing the impact of the issues mentioned earlier on model predictions, thereby enhancing the accuracy of ejection fraction prediction.

191 **METHOD**

192 Data management

The study used a dataset of 10,030 apical-4-chamber echocardiography videos from patients at Stanford 193 University Hospital between 2016 and 2018 ³⁸. The data was meticulously preprocessed to ensure 194 195 integrity and uniformity, including cropping and masking operations. The videos were then down-sampled 196 to a uniform resolution of 112x112 pixels using cubic interpolation, ensuring the quality of the visual data and compatibility with the analytical framework. This dataset is crucial for understanding cardiac function 197 198 representations in full resting echocardiogram studies. The dataset was divided into test, validation, and 199 training sets, with 1,277, 1,288, and 7,462 videos in each set. The histogram in Fig 6a visually represents 200 the EF values in the training set, showcasing the range from 6.90 to 96.96. The histogram shows a 201 dataset's imbalanced distribution of ejection fraction values, with a significant concentration in the 55% to 202 70% range. Consequently, the pattern and spread of EFs around the line indicate how the points in the 203 55% and 70% ranges are closely scattered around a diagonal line (Fig 3a). This imbalance can affect the 204 performance of predictive models trained on this data, potentially leading to bias toward predicting values 205 in the most common range. Additionally, a scatter plot was included to illustrate the spread of ejection 206 fraction values within the training dataset (Fig 6c, d).



Fig 6. Comparative distribution of EF values in the training dataset before and after data augmentation and down-sampling techniques. The initial dataset (**a**) consisted of 7462 samples, while the augmented dataset (**b**) expanded to 30787 samples, illustrating the effect of augmentation and balancing strategies on the EF value distribution. **c** and **d** represent the spread of EF before and after augmentation in the training dataset.

207 In the initial examination of our training dataset, we identified a skewed distribution of EF values, which 208 threatened to bias our predictive model towards the more common EF ranges, thereby impairing its 209 generalizability. We first addressed the variability in frame counts to ensure uniformity in video clip length. 210 Videos with fewer than 32 frames were lengthened by padding the last frame, whereas for videos with 211 fewer than 64 frames, we employed 32 random samples to standardize their length. For videos containing 212 64 frames or more, we generated 32-frame echocardiogram clips by sampling every second frame. This preprocessing protocol was applied to all videos to create a consistent structure for subsequent steps. 213 214 Following this standardization, we specifically targeted the underrepresented EF values for augmentation. 215 For videos with an excess of 64 frames, we generated two distinct clips with variable starting points by sampling every other frame, effectively doubling the representation of these EF ranges. This 216 217 augmentation, performed prior to any down-sampling, was crucial in addressing the initial data imbalance. 218 In the subsequent phase, we down-sampled the overrepresented EF values to balance the dataset. Later, 219 the underrepresented values are applied to each frame through a series of image transformations, 220 including rotation, zoom, shift, and shear. A random factor between 0.99 and 1.01 also changed the EF 221 value for each augmented video. This was done to maintain physiological plausibility and add a realistic 222 range. The histogram in **Fig 6b** visually represents the EF values in the training set after augmentation.

223 Preprocessing

Accurately assessing cardiac function using echocardiograms is crucial to minimize noise and ensure high-quality data for accurate interpretation. To address this, we developed a comprehensive preprocessing pipeline that enhances the interpretability of echocardiogram frames.

The preprocessing method starts with 32 echocardiogram frames with a 52x52 pixel resolution. The median frame is calculated by determining the median value of each pixel location across all 32 frames (temporal dimension), resulting in a singular 52x52 matrix. Then, a frame-wise multiplication operation is performed on each original video frame, resulting in a transformed video with identical dimensions but

modified pixel values by multiplying the corresponding median pixel values. This meticulous operation is performed for all 32 frames in the sequence. Subsequently, histogram equalization was applied to each frame to adjust contrast and improve the visibility of cardiac structures, followed by a median blur filter with a 3x3 pixel mask to reduce noise while preserving essential anatomical details. **Fig 7** compares the first 10 frames of the original video and their preprocessed counterparts, showing the significant improvements.



Fig 7. Sequential visualization of echocardiogram frame preprocessing. The top row (a-k) displays the first 11 original frames from the echocardiogram video, demonstrating the raw imaging data. After applying our preprocessing steps, the bottom row (a-k) illustrates the corresponding frames: median frame calculation, frame-wise multiplication, histogram equalization, and median blur filtering. The processed frames reveal a marked enhancement in the definition and contrast of cardiac structures, providing a clear visual distinction from the original frames and underscoring the efficacy of the preprocessing technique.

237 Video-based transformer model and training

The Vision Transformer (ViT) is a pure-transformer architecture that has outperformed convolutional neural networks in image classification, offering a competitive alternative to the widely used convolutional neural networks in computer vision ^{34,39}. The ViViT architecture, inspired by the ViT, provides a new approach to video classification. It uses transformer-based models, leveraging attention-based mechanisms to model long-range contextual relationships in video content. This innovative approach offers a strategic alternative to conventional 3D CNNs ⁴⁰ and RNNs⁴¹, allowing for more accurate and efficient video classification.

Even though ViViT is an efficient video classification model, we trained the ViViT from scratch to directly regress the LVEF from echocardiogram videos. The model performed self-attention, computed on a sequence of spatio-temporal patches we extracted from the echocardiogram videos. We initially replaced the final layer of the classifier head, intended to output various classes, with a new layer designed to produce a single, continuous output. There is only one output unit in this new layer and no activation function. The tubelet embedding of the echocardiogram frames feeds the model with nonoverlapping spatiotemporal information.

The model is structured around a sequence of <u>ten</u> transformer layers. Each layer consists of <u>twelve</u> heads. The token size (model dimension) was set to d = 512. The hidden size of multi-layer perceptron



Fig 8. The layered structure of a neural network model. The diagram showcases the arrangement of various layers, including multihead attention, encoder, and dense layers, illustrating the flow of information within the model.

(MLP) was <u>768</u>. The output of the tokens is then transformed into a regression prediction via an MLP as
 non-linearity in the three hidden layers of 512, 128, and 64. Fig 8 illustrates the layered structure of our
 model.

257 Data availability

The EchoNet-Dynamic dataset was used in this project. It is a public dataset of de-identified echocardiogram videos found at <u>https://echonet.github.io/dynamic/</u>.

260 **ACKNOWLEDGEMENTS**

Three National Institutes of Health grants supported this work: R01HL145753, R01HL145753-01S1, and R01HL145753-03S1; in addition, the work was supported by LSUHSC-S CCDS Finish Line Award, COVID-19 Research Award, and LARC Research Award to MSB; Jane Cheever Powell Foundation for Cardiovascular Research Related to Gender and Isolation, LSUHS to SRB; and Institutional Development Award (IDeA) from the National Institutes of General Medical Sciences of the NIH under grant number P20GM121307 and R01HL149264 to CGK.

267 Author Contributions Statement

Initial study concept and design: MANB, TA. Acquisition of data: TA. Model training: TA, SA. Analysis and
 interpretation of data: TA, SA. Drafting of the paper: TA. Critical revision of the manuscript for important
 intellectual content: MANB. Statistical analysis: TA, SA, MANB. Reading, editing, and approving the
 paper: All the authors.

272 Competing interests

273 The authors declare no competing interests.

274 Ethical Approval and Consent to participate

- 275 Not applicable.
- 276 Replication of results
- The codes and data used are available on request to enable the method proposed in the manuscript to be replicated by readers.

279 **REFERENCES**

- 2801.Robinson S. Cardiovascular disease. Priorities for Health Promotion and Public Health [Internet]2812021[cited2023Nov14];355–93.Availablefrom:282https://www.taylorfrancis.com/chapters/edit/10.4324/9780367823689-16/cardiovascular-disease-sally-robinson
- 284 2. Ziaeian B, Fonarow GC. Epidemiology and aetiology of heart failure. Nature Reviews Cardiology 285 2016 13:6 [Internet] 2016 [cited 2023 Nov 23];13(6):368-78. Available from: 286 https://www.nature.com/articles/nrcardio.2016.25
- Savarese G, Stolfo D, Sinagra G, Lund LH. Heart failure with mid-range or mildly reduced ejection
 fraction. Nature Reviews Cardiology 2021 19:2 [Internet] 2021 [cited 2023 Nov 18];19(2):100–16.
 Available from: https://www.nature.com/articles/s41569-021-00605-5
- Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature 2020 580:7802 [Internet] 2020 [cited 2023 Nov 18];580(7802):252–6. Available from: https://www.nature.com/articles/s41586-020-2145-8
- 5. Gopal AS, Shen Z, Sapin PM, et al. Assessment of Cardiac Function by Three-dimensional Echocardiography Compared With Conventional Noninvasive Methods. Circulation [Internet] 1995

- 295
 [cited
 2023
 Nov
 18];92(4):842–53.
 Available
 from:

 296
 https://www.ahajournals.org/doi/abs/10.1161/01.CIR.92.4.842
 from:
 from:
 from:
- 2976.Lang RM, Badano LP, Mor-Avi V, et al. Recommendations for Cardiac Chamber Quantification by298Echocardiography in Adults: An Update from the American Society of Echocardiography and the299European Association of Cardiovascular Imaging. Eur Heart J Cardiovasc Imaging [Internet] 2015300[cited 2023 Nov 24];16(3):233–71. Available from: https://dx.doi.org/10.1093/ehjci/jev014
- 3017.Farsalinos KE, Daraban AM, Ünlü S, Thomas JD, Badano LP, Voigt JU. Head-to-Head302Comparison of Global Longitudinal Strain Measurements among Nine Different Vendors: The303EACVI/ASE Inter-Vendor Comparison Study. Journal of the American Society of Echocardiography3042015;28(10):1171-1181.e2.
- Cannesson M, Tanabe M, Suffoletto MS, et al. A Novel Two-Dimensional Echocardiographic Image
 Analysis System Using Artificial Intelligence-Learned Pattern Recognition for Rapid Automated
 Ejection Fraction. J Am Coll Cardiol 2007;49(2):217–26.
- 3089.Kim Y, Garvin JH, Goldstein MK, et al. Extraction of left ventricular ejection fraction information309from various types of clinical reports. J Biomed Inform 2017;67:42–8.
- 31010.Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nature311Medicine 2019 25:1 [Internet] 2019 [cited 2023 Nov 24];25(1):24–9. Available from:312https://www.nature.com/articles/s41591-018-0316-z
- van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. Nature
 Medicine 2021 27:5 [Internet] 2021 [cited 2023 Nov 24];27(5):775–84. Available from:
 https://www.nature.com/articles/s41591-021-01343-4
- 316 Wainberg M, Merico D, Delong A, Frey BJ. Deep learning in biomedicine. Nature Biotechnology 12. 317 2018 36:9 [Internet] 2018 [cited 2023 Nov 24];36(9):829-38. Available from: 318 https://www.nature.com/articles/nbt.4233
- 31913.Moen E, Bannon D, Kudo T, Graf W, Covert M, Van Valen D. Deep learning for cellular image320analysis. Nature Methods 2019 16:12 [Internet] 2019 [cited 2023 Nov 24];16(12):1233–46.321Available from: https://www.nature.com/articles/s41592-019-0403-1
- 32214.Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision. npj Digital323Medicine 2021 4:1 [Internet] 2021 [cited 2023 Nov 24];4(1):1–9. Available from:324https://www.nature.com/articles/s41746-020-00376-2
- 32515.Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical
imaging: a systematic review and meta-analysis. npj Digital Medicine 2021 4:1 [Internet] 2021
[cited 2023 Nov 24];4(1):1–23. Available from: https://www.nature.com/articles/s41746-021-00438-
z328z
- 32916.Lin TY, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context. Lecture Notes in
Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in
Bioinformatics) [Internet] 2014 [cited 2023 Nov 24];8693 LNCS(PART 5):740–55. Available from:
https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48
- 33317.Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge. Int J334Comput Vis [Internet] 2015 [cited 2023 Nov 24];115(3):211–52. Available from:335https://link.springer.com/article/10.1007/s11263-015-0816-y
- 33618.Hirschberg J, Manning CD. Advances in natural language processing. Science (1979) [Internet]3372015[cited 2023 Nov 24];349(6245):261–6.Available from:338https://www.science.org/doi/10.1126/science.aaa8685
- Litjens G, Sánchez CI, Timofeeva N, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific Reports 2016 6:1 [Internet] 2016 [cited 2024 Mar 18];6(1):1–11. Available from: https://www.nature.com/articles/srep26286
- 34220.Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical343imaging: a systematic review and meta-analysis. npj Digital Medicine 2021 4:1 [Internet] 2021

344 [cited 2024 Mar 18];4(1):1–23. Available from: https://www.nature.com/articles/s41746-021-00438 345 z

- 346 21. Odigwe BE, Rajeoni AB, Odigwe CI, Spinale FG, Valafar H. Application of machine learning for 347 patient response prediction to cardiac resynchronization therapy. Proceedings of the 13th ACM 348 International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB 349 2022 [Internet] 2022 [cited 2024 Feb 11];Available from: 350 https://dl.acm.org/doi/10.1145/3535508.3545513
- Rajeoni B, Pederson A;, Clair B;, et al. Automated Measurement of Vascular Calcification in
 Femoral Endarterectomy Patients Using Deep Learning. Diagnostics 2023, Vol 13, Page 3363
 [Internet] 2023 [cited 2024 Feb 11];13(21):3363. Available from: https://www.mdpi.com/20754418/13/21/3363/htm
- 35523.Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of
echocardiograms using deep learning. npj Digital Medicine 2018 1:1 [Internet] 2018 [cited 2023
Nov 25];1(1):1–8. Available from: https://www.nature.com/articles/s41746-017-0013-1
- 35824.Ghorbani A, Ouyang D, Abid A, et al. Deep learning interpretation of echocardiograms. npj Digital359Medicine 2020 3:1 [Internet] 2020 [cited 2023 Nov 25];3(1):1–10. Available from:360https://www.nature.com/articles/s41746-019-0216-8
- Wei H, Cao H, Cao Y, et al. Temporal-Consistent Segmentation of Echocardiography with Colearning from Appearance and Shape. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) [Internet] 2020 [cited 2023 Nov 25];12262 LNCS:623–32. Available from: https://link.springer.com/chapter/10.1007/978-3-030-59713-9_60
- 366 Reynaud H, Vlontzos A, Hou B, Beqiri A, Leeson P, Kainz B. Ultrasound Video Transformers 26. 367 for Cardiac Ejection Fraction Estimation. Lecture Notes in Computer Science (including subseries 368 Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) [Internet] 2021 [cited 369 2023 Nov 25];12906 LNCS:495-505. Available from: 370 https://link.springer.com/chapter/10.1007/978-3-030-87231-1 48
- 371 27. Jafari MH, Woudenberg N Van, Luong C, Abolmaesumi P, Tsang T. Deep bayesian image
 372 segmentation for a more robust ejection fraction estimation. Proceedings International
 373 Symposium on Biomedical Imaging 2021;2021-April:1264–8.
- 28. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking Atrous Convolution for Semantic Image
 375 Segmentation. 2017 [cited 2023 Nov 25];Available from: https://arxiv.org/abs/1706.05587v3
- Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M. A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018. p. 6450–9.
- 37930.Dai W, Li X, Ding X, Cheng KT. Cyclical Self-Supervision for Semi-Supervised Ejection Fraction380Prediction From Echocardiogram Videos. IEEE Trans Med Imaging 2023;42(5):1446–61.
- 38131.Lang RM, Badano LP, Mor-Avi V, et al. Recommendations for Cardiac Chamber Quantification by382Echocardiography in Adults: An Update from the American Society of Echocardiography and the383European Association of Cardiovascular Imaging. Eur Heart J Cardiovasc Imaging [Internet] 2015384[cited 2023 Nov 25];16(3):233–71. Available from: https://dx.doi.org/10.1093/ehjci/jev014
- 385 32. Poplin R, Varadarajan A V., Blumer K, et al. Prediction of cardiovascular risk factors from retinal
 fundus photographs via deep learning. Nature Biomedical Engineering 2018 2:3 [Internet] 2018
 [cited 2023 Nov 25];2(3):158–64. Available from: https://www.nature.com/articles/s41551-018 0195-0
- Hughes JW, Tooley J, Torres Soto J, et al. A deep learning-based electrocardiogram risk score for
 long term cardiovascular death and disease. npj Digital Medicine 2023 6:1 [Internet] 2023 [cited
 2023 Nov 25];6(1):1–9. Available from: https://www.nature.com/articles/s41746-023-00916-6

- 392 34. Arnab A, Dehghani M, Heigold G, Sun C, Luči c ML, Schmid C. ViViT: A Video Vision Transformer.
 393 2021;6836–46.
- 394 35. Ouyang D, He B, Ghorbani A, et al. Echonet-dynamic: a large new cardiac motion video data
 resource for medical machine learning. In: NeurIPS ML4H Workshop: Vancouver, BC, Canada.
 2019.
- 397 36. Thavendiranathan P, Popović ZB, Flamm SD, Dahiya A, Grimm RA, Marwick TH. Improved
 398 Interobserver Variability and Accuracy of Echocardiographic Visual Left Ventricular Ejection
 399 Fraction Assessment through a Self-Directed Learning Program Using Cardiac Magnetic
 400 Resonance Images. Journal of the American Society of Echocardiography 2013;26(11):1267–73.
- 40137.Lang RM, Badano LP, Mor-Avi V, et al. Recommendations for cardiac chamber quantification by
echocardiography in adults: an update from the American Society of Echocardiography and the
European Association of Cardiovascular Imaging. European Heart Journal-Cardiovascular Imaging
2015;16(3):233–71.
- 40538.Ouyang D, He B, Ghorbani A, et al. Echonet-dynamic: a large new cardiac motion video data406resource for medical machine learning. In: NeurIPS ML4H Workshop: Vancouver, BC, Canada.4072019.
- 408 39. Ouyang D, He B, Ghorbani A, et al. Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. In: NeurIPS ML4H Workshop: Vancouver, BC, Canada.
 410 2019.
- 41140.Feichtenhofer C. X3d: Expanding architectures for efficient video recognition. In: Proceedings of412the IEEE/CVF conference on computer vision and pattern recognition. 2020. p. 203–13.
- 41. Bhardwaj S, Srinivasan M, Khapra MM. Efficient video classification using fewer frames. In:
 414 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. p.
 415 354–63.

416