EchoNet-Dynamic: a Large New Cardiac Motion Video Data Resource for Medical Machine Learning

David Ouyang[1] **Bryan He**[2]

Amirata Ghorbani[3] **Matt P. Lungren**[4] **Euan A. Ashley**[1]

David H. Liang[1] **James Y. Zou**[2,5]

1. Department of Medicine, Stanford University
2. Department of Computer Science, Stanford University
3. Department of Electrical Engineering, Stanford University
4. Department of Radiology, Stanford University
5. Department of Biomedical Data Science, Stanford University

{ouyangd, bryanhe, amiratag, mlungren, euan, dliang, jamesz} @stanford.edu

Abstract

Machine learning analysis of biomedical images has seen significant recent advances. In contrast, there has been much less work on medical videos, despite the fact that videos are routinely used in many clinical settings. A major bottleneck for this is the the lack of openly available and well annotated medical video data. Computer vision has benefited greatly from many open databases which allow for collaboration, comparison, and creation of medical task specific architectures. We present the EchoNet-Dynamic Dataset of 10,036 echocardiography videos, spanning the range of typical echocardiography lab imaging conditions, with corresponding labeled measurements including ejection fraction, left ventricular volume at end-systole and end-diastole, and human expert tracings of the left ventricle as an aid in studying automated approaches to evaluate cardiac function. We additionally present the performance of three 3D convolutional architectures for video classification used to assess ejection fraction to near-expert human performance and as a benchmark for further collaboration, comparison, and creation of task-specific architectures. To the best of our knowledge, this is the largest labeled medical video dataset made available publicly to researchers and medical professionals and first public report of video-based 3D convolutional architectures to assess cardiac function.

1 Introduction

Echocardiography is the most widely used and readily available imaging technique to assess cardiac function and structure. Combining rapid image acquisition, high temporal resolution, and without the risks of ionizing radiation, echocardiography serves as the backbone of cardiovascular imaging [22, 9] and is one of the most frequently used imaging studies in the United States [5]. Information from echocardiography is used by cardiologists, surgeons, emergency physicians, anesthesiologists, and oncologists among other physicians as echocardiography is used for perioperative risk stratification, manage cardiovascular risk of patients with oncologic disease undergoing chemotherapy, and aid

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

in the diagnosis and surgical planning of cardiovascular disease[24, 1, 17]. For indications ranging from cardiomyopathies to valvular heart diseases, echocardiography is both necessary and sufficient to diagnose many cardiovascular diseases.

Despite its importance in clinical phenotyping, there is variance in the human interpretation of echocardiogram images that could impact clinical care [29, 14, 18] and there is significant interest in improving measurement precision and reproducibility [26, 11]. Formalized training guidelines for cardiologists recognize the value of experience in interpreting echocardiogram images and basic cardiology training might be insufficient to interpret echocardiograms at the highest level [33] even as other physicans are beginning to use point-of-care ultrasound for bedside diagnosis [7]. Previous work using deep learning to estimate ejection fraction from echocardiography has been limited to still-image based methods and have significant variance from human measurements [35, 21, 23, 15]. The limited published work on the direct comparison of performance of different architectures to predict ejection fraction [3, 34].

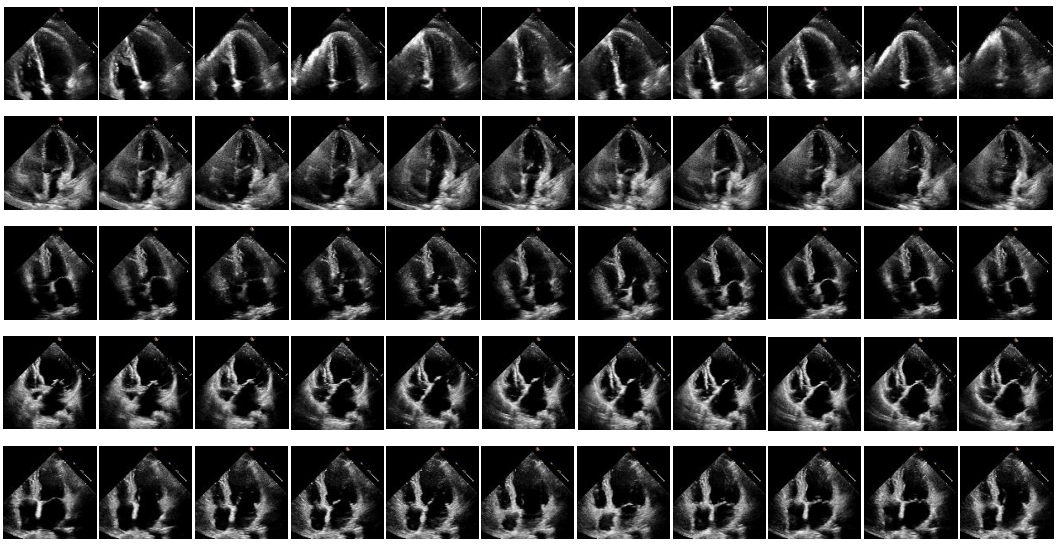

Figure 1: Representative frames from the EchoNet-Dynamic dataset. Eleven frames of five independent videos are shown after pre-processing removing ECG data, text labels, and ultrasound acquisition information. Each video is matched with calculations and measurements from the clinical report.

Echocardiography is a uniquely well-suited imaging modality for the application of deep learning in cardiology. In addition to standardized imaging windows and views, echocardiography reporting systems often uses structured reporting, readily making clinical databases available for model training on diverse phenotypes. Many previous studies of deep learning on medical imaging focused on resource-intensive imaging modalities common in resource-rich settings [6, 2] or subspeciality imaging with focused indication [25, 10, 28]. These modalities often need retrospective annotation by experts, while the standard clinical workflow of echocardiography includes detailed measurements, labels, and region specific interpretations.

In this paper, we introduce a new large video dataset of echocardiograms for computer vision research. The EchoNet-Dynamic database was created to provide images to study cardiac motion and chamber volumes using echocardiography, or cardiac ultrasound, videos obtained in real clinical practice for diagnosis and medical decision making. Clinically important metrics, such as ejection fraction, are linked with representative echocardiography videos for supervised training tasks. We developed this dataset because there is a current lack of such a dataset of echocardiogram videos, and we believe a dataset large enough to train deep neural networks and also large enough to act as a performance benchmark can be used to assess different model architectures can advance the field of deep learning on echocardiography.

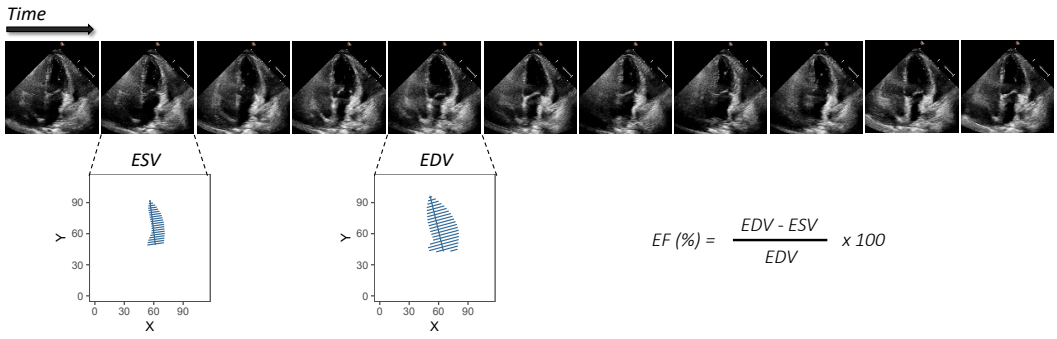

Figure 2: Diagram representing standard clinical workflow in assessing cardiac functions. Based on cardiac timing, human experts trace the chamber size of the left ventricle to obtain end diastolic volume (EDV) and end systolic volume (ESV). The ratio of ESV and EDV is used to calculate the ejection fraction (EF).

2 Overview of the EchoNet-Dynamic Dataset

2.1 Composition

A standard full resting echocardiogram study consists of a series of 50-100 videos and still images visualizing the heart from different angles, locations, and image acquisition techniques (2D images, tissue Doppler images, color Doppler images, and others). In this dataset, one apical-4-chamber 2D gray-scale video is extracted from each study. Each video represents an unique individual as the dataset contains 10,036 echocardiography videos from 10,036 random individuals who underwent echocardiography between 2006 and 2018 as part of clinical care at a University Hospital. The apical-4-chamber view video was identified by extracting the Digital Imaging and Communications In Medicine (DICOM) file linked to measurements of ventricular volume used to calculate the ejection fraction in the apical-4-chamber view.

Table 1: Dataset Label Variables

Variable	Description
FileName	Hashed file name used to link videos, labels, and annotations
Age	Age in years, rounded to nearest year
Sex	Sex reported in medical record
EF	Ejection fraction calculated by ratio of ESV and EDV
ESV	End systolic volume calculated by method of discs
EDV	End diastolic volume calculated by method of discs
Height	Video Height
Width	Video Width
FPS	Frames Per Second
NumFrames	Number of Frames in whole video
Split	Classification of train/validation/test sets used for benchmarking

2.2 Clinical Measurements and Calculations

In addition to the video itself, each study is linked to clinical measurements and calculations obtained by a registered sonographer and verified by a level 3 echocardiographer in the standard clinical workflow. A central metric of cardiac function is the left ventricular ejection fraction [20, 13, 29], used to diagnose cardiomyopathy, assess eligibility for certain chemotherapies, and determine indication for medical devices. Left ventricular ejection fraction has a significant relationship with mortality in many disease states, with lower ejection fraction correlating with worse prognosis [31].

The ejection fraction is expressed as a percentage and is the ratio of left ventricular end systolic volume (ESV) and left ventricular end diastolic volume (EDV) determined by (EDV - ESV) / EDV. In our dataset, and in standard echocardiography practice, the left ventricle is traced at the endocardial border at two separate time points representing end-systole and end-diastole for each video (Fig. 3). Each tracing is used to estimate ventricular volume by integration of ventricular area over the length of the major axis of the ventricle [20]. For each video file, the corresponding labels of end systolic volume, end diastolic volume, and ejection fraction are provided as CSV files (Table 1). The tracing file structure is described in Appendix (Table 4).

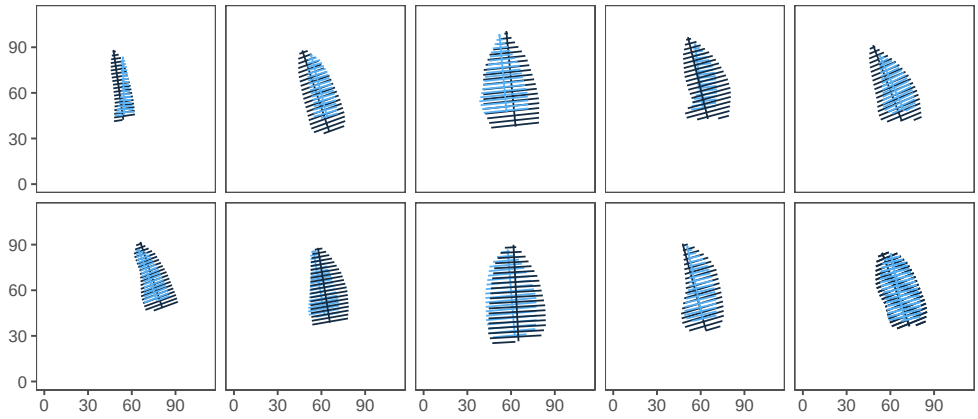

Figure 3: Twenty human expert tracings from 10 videos of the left ventricle endocardium used to calculate end systolic volume and end diastolic volume. Tracings used to estimate left ventricle volume by integration of area over the length of the ventricle. The ratio of end diastolic volume (EDV, black) and end systolic volume (ESV, blue) is used to calculate ejection fraction.

2.3 Statistics

The videos are canonical apical-4-chamber or zoomed-in apical-4-chamber echocardiographic views of sufficient quality for a sonographer to trace left ventricular volumes as part of the standard clinical workflow.The videos consistent of a series of gray-scale images of 112 by 112 pixels. Each video has between 24-1002 frames at a mean of 51 frames per second. There is no known relationship between videos and each video is of a unique individual. Subpopulation data is not available at the initial release of the dataset, however there is minimal known variation by gender or race within the range of measurement error for the released corresponding metrics [27, 4, 11]. The current videos are divided into three splits with roughly 75% for training, 12.5% for validation, and 12.5% for testing.

Table 2: Dataset Summary Statistics

Metric	Total Dataset	Training	Validation	Test
Number of Videos	10,036	7,465	1,289	1,282
Female (%)	4885 (48%)	3662 (49%)	611 (44%)	612 (44%)
Age (Years)	68 (21)	70 (22)	62 (18)	62 (17)
Frames Per Second	50.9 (6.8)	50.8 (6.7)	51.0 (6.5)	51.3 (7.3)
Number of Frames	175 (57)	175 (57)	176 (52)	176 (60)
Ejection Fraction (%)	55.7 (12.5)	55.7 (12.5)	55.8 (12.3)	55.3 (12.4)
End Systolic Volume (mL)	43.3 (34.5)	43.2 (36.1)	43.3 (34.5)	43.9 (36.0)
End Diastolic Volume (mL)	91.0 (45.7)	91.0 (46.0)	91.0 (43.8)	91.4 (46.0)

2.4 Deidentification

This research was approved by the University Institutional Review Board and data privacy review through a standardized workflow by the Center for Artificial Intelligence in Medicine and Imaging (AIMI) and the University Privacy Office. In addition to masking of text, ECG information, and extra data outside of the scanning sector in the video files as described below, each DICOM file's pixel data was parsed out and saved as an AVI file to prevent any leakage of identifying information through public or private DICOM tags. Each video was subsequently manually reviewed by an employee of the Hospital with familiarity with imaging data to confirm the absence of any identifying information.

2.5 Video Processing

The raw images and measurements were obtained from the clinical database of the University Hospital Echocardiography Lab. Images were acquired by skilled sonographers using iE33, Sonos, Acuson SC2000, Epiq 5G, or Epiq 7C ultrasound machines and processed images were stored in Philips Xcelera picture achiving and communication system. Video views were identified through implicit knowledge of view classification in the clinical database by identifying images and videos labeled with measurements done in the corresponding view.The apical-4-chamber view video was identified by extracting the Digital Imaging and Communications In Medicine (DICOM) file linked to measurements of ventricular volume used to calculate the ejection fraction. Videos were spot checked for quality control, confirm view classification, and exclude videos with color Doppler. Each subsequent video was cropped and masked to remove text, ECG and respirometer information, and other information outside of the scanning sector. The resulting square images were either 600x600 or 768x768 pixels depending on the ultrasound machine and downsampled by cubic interpolation using OpenCV into standardized 112x112 pixel videos. The pixel data from the DICOM files was saved into AVI files with hashed file names to future prevent release of potentially identifying metadata in public or private DICOM tags.

2.6 Non-exhaustive annotation

The ejection fraction, end systolic volume, and end diastolic volumes were measured in the clinical setting and included in the clinical report, however there can be variation in the setting of atrial fibrillation, premature atrial contractions, and other sources of ectopy. The convention is to identify at least one representative cardiac cycle, and use this representative cardiac cycle to perform measurements. For this reason, test time augmentation is reasonable if the input clip is significantly smaller in length than the total video length. Similar non-exhaustive annotation is used in classification datasets such as Kinetics and ImageNet[30, 19].

2.7 Distribution and Maintenance

The dataset is available electronically at https://echonet.github.io/echoNet/. The authors will maintain the EchoNet-Dynamic dataset with changes and updates to be described in the corresponding Github page. Users of this dataset must agree to a Research Use Agreement with the University attesting to ethical use of the dataset and limitations of its use, such as excluding commercial or clinical use. Each user must individually register and sign the Research Use Agreement and cannot independently share the dataset to others.

3 Benchmark Performance

In this section, we briefly describe three convolutional neural network architectures for action classification [32] repurposed to predict ejection fraction, end systolic volume, and end diastolic volume from the previously described dataset. We use these architectures as baselines and compare their performance to human clinical measurements.

We chose three architectures with various forms of spatiotemporal convolutions which has been previously benchmarked on action classification tasks on Sports-1M, Kinetics, UCF101, and HMDB51 [32]. These architectures have shown superior performance to image-based 2D CNN architectures on individual frames of the video for image classification. Image-based 2D CNN architectures have

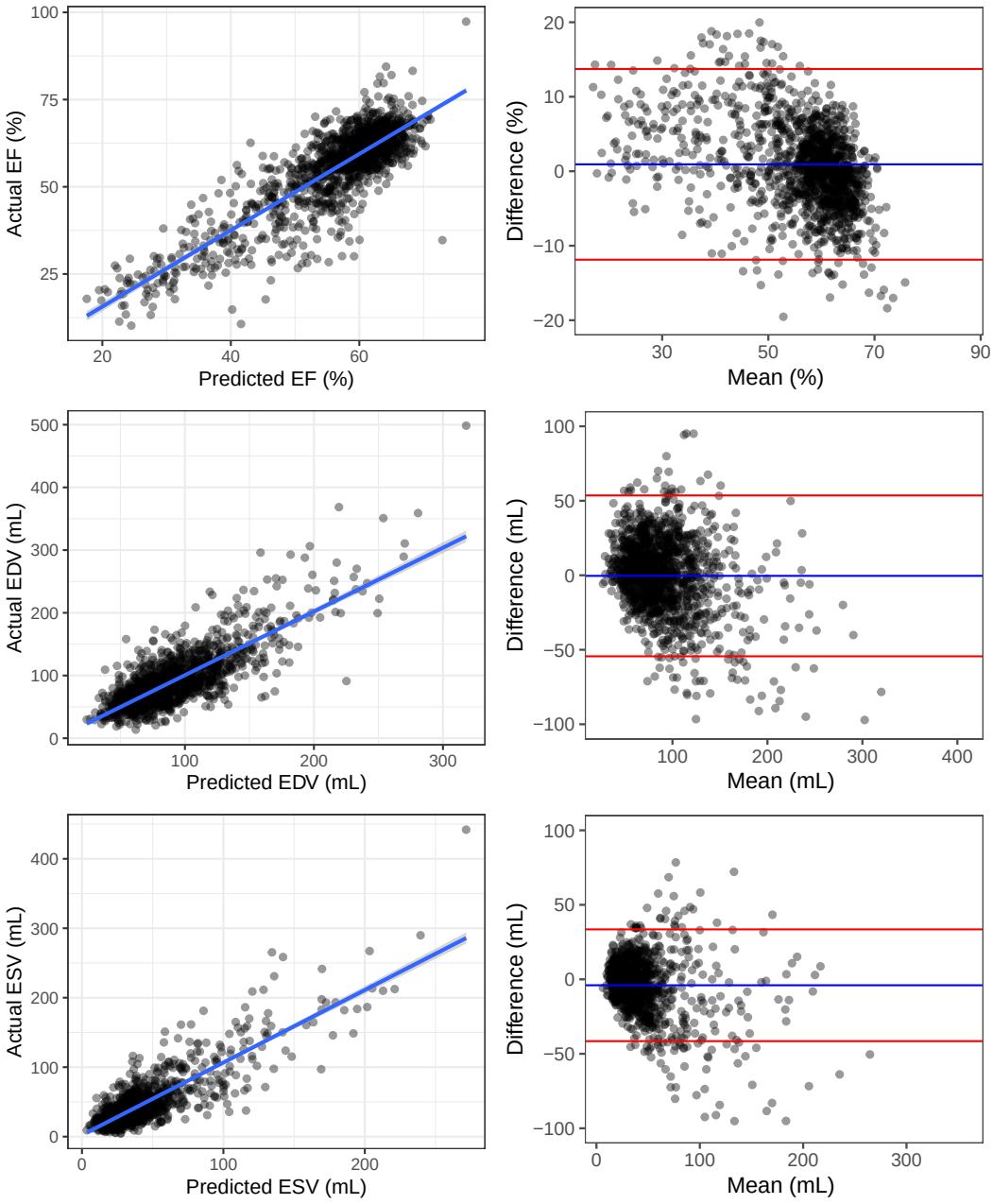

Figure 4: Performance on test dataset in predicting ejection fraction, end diastolic volume, and end systolic volume by R3D model, clip length of 16, sampling rate of 1 in every 4 frames. Represented as scatter plot and Bland-Altman plot.

been attempted to predict ejection fraction [35, 15] however their performance has been significantly inferior compared to human performance.

3.1 Model Architectures

Prior literature has described the performance of convolutional networks coupled with LSTM, two-stream, and 3D convolutional networks for human action classification in videos[8, 12, 32]. In the interest of creating high performance baselines on standard videos with gray-scale images without proprocessing or creation of optical flow frames, we evaluated the performance of three architectures

Table 3: Benchmark Model Performance on Test Set (Hyperparameters chosen from Validation Set)

Task	Model	Clip Length	Sampling Rate	MAE	RMSE	R^2
Ejection Fraction	Human Experts	Entire Video	Every Frame	3.12	4.57	0.88
Ejection Fraction	R3D	16	1 in 4	5.44	6.16	0.71
Ejection Fraction	MC3	16	1 in 4	5.91	6.80	0.69
Ejection Fraction	R2+1D	16	1 in 4	6.87	7.55	0.66
End Systolic Volume	R3D	16	1 in 4	12.7	19.3	0.72
End Systolic Volume	MC3	16	1 in 4	12.4	18.3	0.71
End Systolic Volume	R2+1D	16	1 in 4	12.4	19.7	0.74
End Diastolic Volume	R3D	16	1 in 4	20.0	30.3	0.64
End Diastolic Volume	MC3	16	1 in 4	51.8	35.2	0.61
End Diastolic Volume	R2+1D	16	1 in 4	21.1	28.8	0.60

which combine 3D convolutions over spatiotemporal video volume with residual connections between layers[32]. Each model uses ResNet-18 as the base architecture and consists of 18 convolutional layers with residual connections connecting odd numbered layers [16].

The main difference between the three models are the various filters used in each layer and described in previously [32]. The R3D model uses convolutional filters of equal size in three dimensions in width, height, and time. The mixed convolution 3 (MC3) model combines 3D convolutional filters in the first nine layers with subsequent layers using 2D filters. The R2+1D models factorizes 3D convolutional filters into separate spatial and temporal components with different spatial and temporal sizes such that the total number of parameters is similar to the R3D model. The R3D model with a clip length of 16 and frame sampling rate of 4 performed the best, with a mean absolute error of 5.44%. For context, human accuracy has been described to be about a mean absolute error of 4-5% for skilled echocardiographers in controlled settings [26, 3, 11].

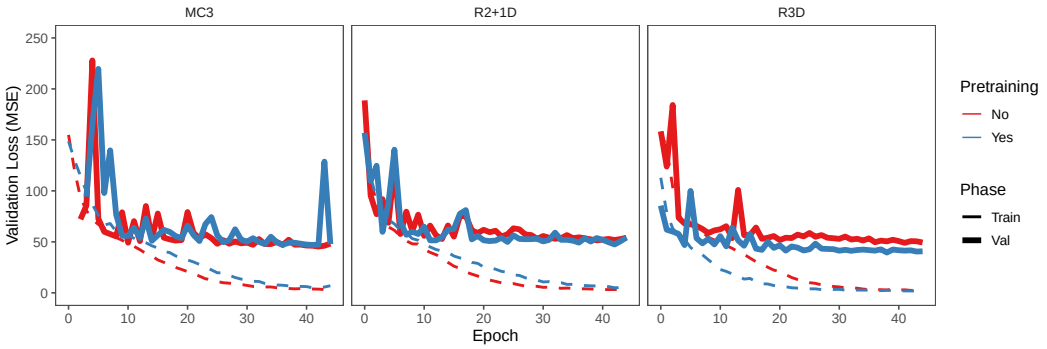

Figure 5: Training and Validation Errors for MC3, R2+1D, R3D. There was no significant difference in performance by transfer learning.

3.2 Implementation Details

The input images scaled to 112×112, with no further cropping applied at either training or test time. For videos whose length was greater than the input clip length, uniform random sampling was performed both at training and test time. Hyperparameter search of the input clip length (total number of frames) and sampling rate was performed using the validation set (Fig. 6). Training time increased linearly with increased input clip length and increased input clip length was associated with better model performance. Increasing sampling rate (which decreased the input clip length for an video of equal length) did not significantly affect model performance. Balancing model performance and training time, we show model performance for an input clip length of 16 frames and a sampling rate of 1 in 4 frames. This would sample a representative video of 64 frames (approximately 1.2 seconds of the video).

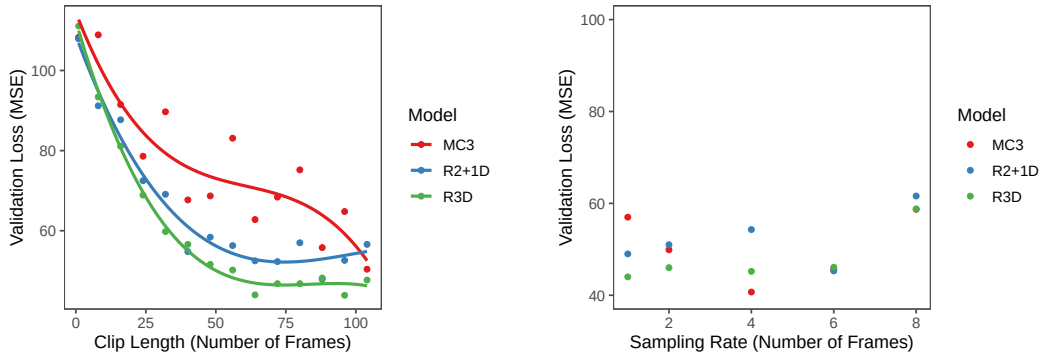

Figure 6: Model performance on validation dataset by variable clip length and frame sampling rate. Performance improved with increasing clip length but plateaued at around 64 frames. Performance was similar at various sampling rates for an equal total sampled clip length of 64 frames.

All models were initialized with pretrained weights from the the Kinetics-400 dataset. The models were then trained to minimize the squared loss between the prediction and true ejection fraction using the SGD optimizer with a learning rate of 0.00005, momentum of 0.9, and batch size of 20 for 45 epochs. Our results were similar to prior experience with the model architectures [32], however we did not see significant different in model performance with or without pre-training with Kinetics-400 weights (Fig. 5).

3.3 Relative Performance by Metric

In each prediction task attempted, 3D convolutions with residual connections (R3D model) outperformed decomposing spatial and temporal convolutions (R2+1D model) and mixing 3D and 2D convolutions (MC3 model). All architectures had the highest performance in predicting ejection fraction. This relative better performance in predicting ejection fraction compared to left ventricular chamber volumes is consistent with knowledge that the actual ultrasound images are scaled by the sonographer during acquisition such that the entire left ventricle as well as surrounding structures are visualized in the apical-4-chamber view. This makes volumetric assessment difficult from the videos alone, however the ejection fraction depends on the relative change in volume, which is scale-invariant, and all the necessary information is captured in the video.

4 Discussion

In this paper, we describe the EchoNet-Dynamic Cardiac Motion Video dataset, a set of 10,271 apical-4-chamber echocardiography videos with tracings and labels of cardiac function. This is the largest medical video dataset to be made publically available and first large release of echocardiogram data with labels and tracings. We describe our process for deidentifying and preprocessing medical videos for public release and identifying relevant labels and tracings. We also present the performance of three 3D convolutional architectures to assess ejection fraction to close to expert human performance and as a benchmark for further collaboration, comparison, and creation of task-specific architectures.

Cardiac function is a medical example of a task dependent on motion that would be difficult to assess based on still motion images alone. While ejection fraction is imperfect, we show that training on videos of cardiac motion using three different convolutional architectures with residual connections that have previously only been used for classification tasks also predicts ejection fraction well, close to human expert performance. Additionally, we show that transfer learning has marginal/limited benefit on ultrasound images as pre-training with Kinetics-400 did not have significant improvement in performance.

Important open questions for the community that can be answered with this dataset include how to improve the reliability and accuracy of cardiac function assessment, evaluation of beat-to-beat ejection fraction for assessment of cardiac function with an irregular heart rates, supervised or self-supervised video alignment of the cardiac cycle, and other important cardiology questions.

While this is a comprehensive dataset for the use of video images to evaluate cardiac function, echocardiography has much more additional phenotypic information. Future work could include the inclusion of additional echocardiographic views to visualize all segments of the left ventricle, linking ECG data with echocardiogram data, inclusion of cardiovascular history of patients, and the addition of temporally separated studies of the sample patients to assess the effect of aging and prognosticate disease progression.

Acknowledgments

The collection of this dataset was funded by the Stanford Echocardiography Lab and the Stanford Center for Artificial Intelligence in Medicine and Imaging (AIMI). We are very grateful for help by Curt Langlotz, Sachin Dutt, Srikanth Eluru, Fatima Rodriguez, Robert Harrington, Ingela Schnittger, and Paul Heidenreich.

Appendix

Table 4: Tracings File Variables

Variable	Description
FileName	Hashed file name used to link videos, labels, and annotations
X1	X coordinate of left most point of line segment
Y1	Y coordinate of left most point of line segment
X1	X coordinate of right most point of line segment
Y1	Y coordinate of right most point of line segment
Length	Native length in centimeters of line segment
Frame	Frame number of video on which tracing was performed

References

[1] Husam Abdel-Qadir, Peter C. Austin, Douglas S. Lee, Eitan Amir, Jack V. Tu, Paaladinesh Thavendiranathan, Kinwah Fung, and Geoffrey M. Anderson. A population-based study of cardiovascular mortality following early-stage breast cancer. *JAMA Cardiology*, 2(1):88–93, 2017.

[2] Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj amd Bokyung Choi, Joshua J. Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, David P. Naidich, and Shravya Shetty. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25:954—-961, 2019.

[3] Federico M. Asch, Theodore Abraham, Madeline Jankowski, Jayne Cleve, Mike Adams, Nathanael Romano, Nicolas Poilvert, Ha Hong, and Roberto Lang. Accuracy and reproducibility of a novel artificial intelligence deep learning-based algorithm for automated calculation of ejection fraction in echocardiography, ACC, 2019.

[4] Federico M. Asch, Jose Banchs, Rhonda Price, Vera Rigolin, James D. Thomas, Neil J. Weissman, and Roberto M. Lang. Need for a global definition of normative echo values-rationale and design of the world alliance of societies of echocardiography normal values study (wase). *Journal of the American Society of Echocardiography*, 13(1):157–162, 2019.

[5] Virnig BA, Shippee ND, O'Donnell B, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Agency for Healthcare Research and Quality (US); 2011-.*, 2014.

[6] Ghalib A Bello, Timothy JW Dawes, Jinming Duan, Carlo Biffi, Antonio de Marvao, Luke SGE Howard, J Simon R Gibbs, Martin R Wilkins, Stuart A Cook, Daniel Rueckert, et al. Deep-learning cardiac motion analysis for human survival prediction. *Nature Machine Intelligence*, 1(2):95, 2019.

[7] Bennett CE, Samavedam S, Jayaprakash N, Kogan A, Gajic O, and Sekiguchi H. When to incorporate point-of-care ultrasound (pocus) into the initial assessment of acutely ill patients: a pilot crossover study to compare 2 pocus-assisted simulation protocols. *Cardiovasc Ultrasound.*, 11(16), 2018.

[8] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[9] PS Douglas, MJ Garcia, DE Haines, WW Lai, et al. Accf/ase/aha/asnc/hfsa/hrs/scai/sccm/scct/scmr 2011 appropriate use criteria for echocardiography. 24(3):229–267, 2011.

[10] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[11] Konstantinos E. Farsalinos, Ana M. Daraban, Serkan Ünlü, James D. Thomas, Luigi P. Badano, and Jens-Uwe Voigt. Head-to-head comparison of global longitudinal strain measurements among nine different vendors. *Journal of the American Society of Echocardiography*, 28(10):1171–1182, 2015.

[12] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[13] Thomas A Foley, Sunil V Mankad, Nandan S Anavekar, Crystal R Bonnichsen, Michael F Morris, Todd D Miller, and Philip A Araoz. Measuring left ventricular ejection fraction - techniques and potential pitfalls, 2012.

[14] D De Geer, A Oscarsson, and J Engvall. Variability in echocardiographic measurements of left ventricular function in septic shock patients. *J. Cardiovasc Ultrasound.*, 13(19), 2015.

[15] Amirata Ghorbani1, David Ouyang, Abubakar Abid, Bryan He, Jonathan H. Chen, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Deep learning interpretation of echocardiograms. bioarxiv: 10.1101/681676v1, 2019.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. arxiv:1512.03385v1, 2015.

[17] PA Heidenreich, JG Trogdon, OA Kavjou, J Butler, K Dracup, et al. Forecasting the future of cardiovascular disease in the united states: a policy statement from the american heart association. *Circulation*, 123(8):933–944, 2011.

[18] Abbott JA and Gentile-Solomon JM. Echocardiographic variables used to estimate pulmonary artery pressure in dogs. *J Vet Intern Med.* ., 31, 2017.

[19] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

[20] Ateet Kosaraju and Amgad N. Makaryus. Left ventricular ejection fraction, 2019.

[21] Ali Madani, Ramy Arnaout, Mohammad Mofrad, and Rima Arnaout. Fast and accurate view classification of echocardiograms using deep learning. *npj Digital Medicine*, 1(1):6, 2018.

[22] Ali Madani, Jia Rui Ong, Ashul Tiberwal, and Mohammad RK Mofrad. U.s. hospital use of echocardiography: Insights from the nationwide inpatient sample. *J Am Coll Cardiol*, 67(5):502–511, 2016.

[23] Ali Madani, Jia Rui Ong, Ashul Tiberwal, and Mohammad RK Mofrad. Deep echocardiography: data-efficient supervised and semisupervised deep learning towards automated diagnosis of cardiac disease. *npj Digital Medicine*, 1(59), 2018.

[24] Ford MK, Beattie WS, and Wijeysundera DN. Systematic review: prediction of perioperative cardiac complications and mortality by the revised cardiac risk index. *Annals of Internal Medicine*, 152(2):26–35, 2010.

[25] Kunal Nagpal, Davis Foote, Yun Liu, Ellery Wulczyn, Fraser Tan, Niels Olson, Jenny L Smith, Arash Mohtashamian, James H Wren, Greg S Corrado, et al. Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *arXiv preprint arXiv:1811.06497*, 2018.

[26] David Ouyang, Amirata Ghorbani, Cindy Wang, Alberta Yen, Francois Haddad, James Zou, Euan Ashley, and David Liang. Defining 'no significant change': Standard error of standard measurements, ASE, 2019.

[27] Stefan Pfaffenberger, Phillipp Bartko, Alexandra Graf, Elisabeh Pernicka, Jamil Babayev, Emina Lolic, Diana Bonderman, Helmut Baumgartner, Gerald Maurer, and Julia Mascherbauer. Size matters! impact of age, sex, height, and weight on the normal heart size. *Circ Cardiovascular Imaging*, 6(1):1073–1079, 2013.

[28] Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158, 2018.

[29] Wood PW, Choy JB, Nanda NC, and Becher H. Left ventricular ejection fraction and volumes: it depends on the imaging method. *Echocardiography*, 31, 2014.

[30] O Russakovsky, J Deng, H Su, J Krause, S Satheesh, S Ma, Z Huang, A Karpathy, A Khosla, M Bernstein, et al. Imagenet large scale visual recognition challenge. arxiv: 1409.0575, 2015.

[31] Kevin S. Shah, Haolin Xu, Roland A. Matsouaka, Deepak L. Bhatt, Paul A. Heidenreich, Adrian F. Hernandez, Adam D. Devore, Clyde W. Yancy, and Gregg C. Fonarow. Heart failure with preserved, borderline, and reduced ejection fraction 5-year outcomes. *Journal of the American College of Cardiology*, 2017.

[32] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. arxiv: 11711.11248v3, 2018.

[33] Susan E. Wiegers, Thomas Ryan, James A. Arrighi, Samuel M. Brown, et al. 2019 acc/aha/ase advanced training statement on echocardiography (revision of the 2003 acc/aha clinical competence statement on echocardiography). 19, 2019.

[34] Alberta Yen, Ali Khalid Chaudhry, Cindy Wang, Xiu Tang, Ha Hong, Nicolas Poilvert, and David Liang. Echogps and autoef help novices perform efficient and accurate echocardiographic monitoring in cancer patients, ACC, 2019.

[35] Jeffrey Zhang, Sravani Gajjala, Pulkit Agrawal, Geoffrey H Tison, Laura A Hallock, Lauren Beussink-Nelson, Mats H Lassen, Eugene Fan, Mandar A Aras, ChaRandle Jordan, et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation*, 138(16):1623–1635, 2018.