npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Deep learning interpretation of echocardiograms

Amirata Ghorbani[1,6], David Ouyang[2,6*], Abubakar Abid[1], Bryan He[3], Jonathan H. Chen[2], Robert A. Harrington[2], David H. Liang[2], Euan A. Ashley[2] and James Y. Zou[1,3,4,5*]

Echocardiography uses ultrasound technology to capture high temporal and spatial resolution images of the heart and surrounding structures, and is the most common imaging modality in cardiovascular medicine. Using convolutional neural networks on a large new dataset, we show that deep learning applied to echocardiography can identify local cardiac structures, estimate cardiac function, and predict systemic phenotypes that modify cardiovascular risk but not readily identifiable to human interpretation. Our deep learning model, EchoNet, accurately identified the presence of pacemaker leads (AUC = 0.89), enlarged left atrium (AUC = 0.86), left ventricular hypertrophy (AUC = 0.75), left ventricular end systolic and diastolic volumes ($R^2 = 0.74$ and $R^2 = 0.70$), and ejection fraction ($R^2 = 0.50$), as well as predicted systemic phenotypes of age ($R^2 = 0.46$), sex (AUC = 0.88), weight ($R^2 = 0.56$), and height ($R^2 = 0.33$). Interpretation analysis validates that EchoNet shows appropriate attention to key cardiac structures when performing human-explainable tasks and highlights hypothesis-generating regions of interest when predicting systemic phenotypes difficult for human interpretation. Machine learning on echocardiography images can streamline repetitive tasks in the clinical workflow, provide preliminary interpretation in areas with insufficient qualified cardiologists, and predict phenotypes challenging for human evaluation.

npj Digital Medicine (2020)3:10 ; https://doi.org/10.1038/s41746-019-0216-8

INTRODUCTION

Cardiovascular disease has a substantial impact on overall health, well-being, and life-expectancy. In addition to being the leading cause of mortality for both men and women, cardiovascular disease is responsible for 17% of the United States' national health expenditures.[1] Even as the burden of cardiovascular disease is expected to rise with an aging population,[1] there continues to be significant racial, socioeconomic, and geographic disparities in both access to care and disease outcomes.[2,3] Variation in access to and quality of cardiovascular imaging has been linked to disparities in outcomes.[3,4] It has been hypothesized that automated image interpretation can enable more available and accurate cardiovascular care and begin to alleviate some of the disparities in cardiovascular care.[5,6] The application of machine learning in cardiology is still in its infancy, however there is significant interest in bringing neural network based approaches to cardiovascular imaging.

Machine learning has transformed many fields, ranging from image processing and voice recognition systems to super-human performance in complex strategy games.[7] Many of the biggest recent advances in machine learning come from computer vision algorithms and processing image data with deep learning.[8–11] Recent advances in machine learning suggest deep learning can identify human-identifiable characteristics as well as phenotypes unrecognized by human experts.[12,13] Efforts to apply machine learning to other modalities of medical imaging have shown promise in computer-assisted diagnosis.[12–16] Seemingly unrelated imaging of individual organ systems, such as fundoscopic retina images, can predict systemic phenotypes and predict cardiovascular risk factors.[12] Additionally, deep learning algorithms perform well in risk stratification and classification of disease.[14,16] Multiple recent medical examples outside of cardiology show convolutional neural network (CNN) algorithms can match or even exceed human experts in identifying and classifying diseases.[13,14]

Echocardiography is a uniquely well-suited approach for the application of deep learning in cardiology. The most readily available and widely used imaging technique to assess cardiac function and structure, echocardiography combines rapid image acquisition with the lack of ionizing radiation to serve as the backbone of cardiovascular imaging.[4,17] Echocardiography is both frequently used as a screening modality for healthy, asymptomatic patients as well as in order to diagnose and manage patients with complex cardiovascular disease.[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.[18–20] 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.[21]

Given the importance of imaging to cardiovascular care, an automated pipeline for interpreting cardiovascular imaging can improve peri-operative risk stratification, manage the cardiovascular risk of patients with oncologic disease undergoing chemotherapy, and aid in the diagnosis of cardiovascular disease.[1,22,23] While other works applying machine learning to medical imaging required re-annotation of images by human experts, the clinical workflow for echocardiography inherently includes many measurements and calculations and often is reported through structured reporting systems. The ability to use previous annotations and interpretations from clinical reports can greatly accelerate adoption of machine learning in medical imaging. Given the availability of previously annotated clinical

[1]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [2]Department of Medicine, Stanford University, Stanford, CA, USA. [3]Department of Computer Science, Stanford University, Stanford, CA, USA. [4]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [5]Chan-Zuckerberg Biohub, San Francisco, CA, USA. [7]These authors contributed equally: Amirata Ghorbani, David Ouyang. *email: ouyangd@stanford.edu; jamesz@stanford.edu

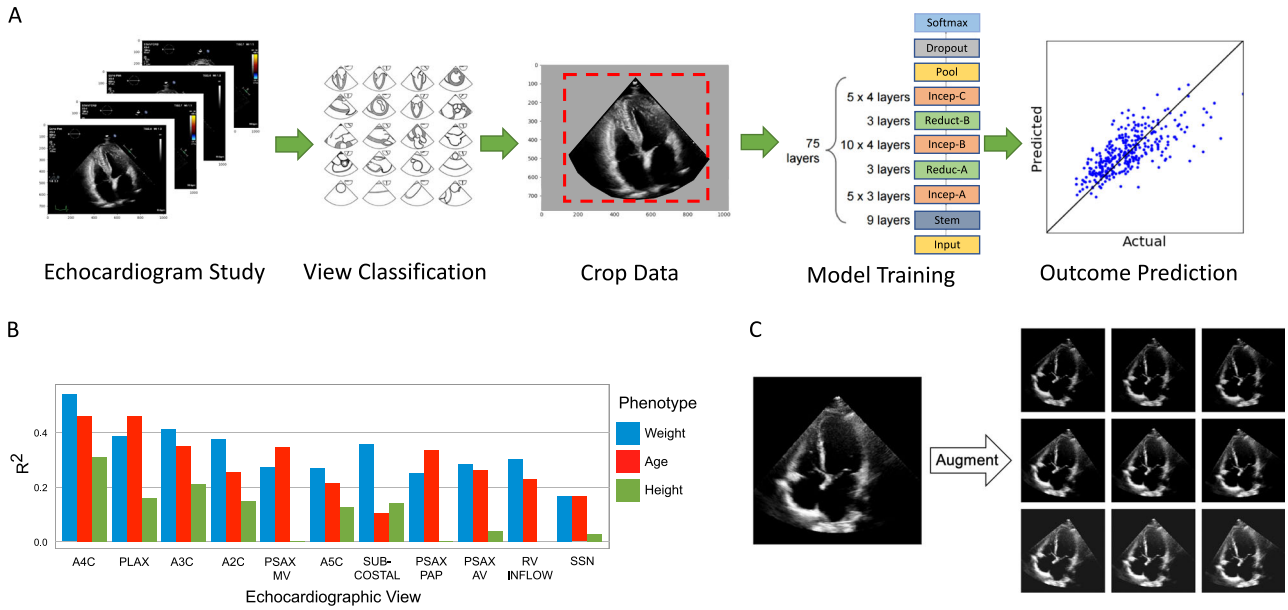

Fig. 1 EchoNet machine learning pipeline for outcome prediction. a EchoNet workflow for image selection, cleaning, and model training. **b** Comparison of model performance with different cardiac views as input. **c** Examples of data augmentation. The original frame is rotated (left to right) and its intensity is increase (top to bottom) as augmentations.

reports, the density of information in image and video datasets, and many available machine learning architectures already applied to image datasets, echocardiography is a high impact and highly tractable application of machine learning in medical imaging.

Current literature have already shown that it is possible to identify standard echocardiogram views from unlabeled datasets.[5,6,24] Previous works have used CNNs trained on images and videos from echocardiography to perform segmentation to identify cardiac structures and derive cardiac function. In this study, we extend previous analyses to show that EchoNet, our deep learning model using echocardiography images, can reliably identify local cardiac structures and anatomy, estimate volumetric measurements and metrics of cardiac function, and predict systemic human phenotypes that modify cardiovascular risk. Additionally, we show the first application of interpretation frameworks to understand deep learning models from echocardiogram images. Human-identifiable features, such as the presence of pacemaker and defibrillator leads, left ventricular hypertrophy, and abnormal left atrial chamber size identified by our CNN were validated using interpretation frameworks to highlight the most relevant regions of interest. To the best of our knowledge, we develop the first deep learning model that can directly predict age, sex, weight, and height from echocardiogram images and use interpretation methods to understand how the model predicts these systemic phenotypes difficult for human interpreters.

RESULTS

We trained a CNN model on a data set of more than 2.6 million echocardiogram images from 2850 patients to identify local cardiac structures, estimate cardiac function, and predict systemic risk factors (Fig. 1). Echocardiogram images, reports, and measurements were obtained from an accredited echocardiography lab of a large academic medical center (Table 1). Echocardiography visualizes cardiac structures from various different orientations and geometries, so images were classified by cardiac view to homogenize the input data set. Echocardiogram images were sampled from echocardiogram videos, pre-processed by de-identifying the images, and cropped to eliminate information outside of the scanning sector. These processed images were used to train EchoNet on the relevant medical classification or prediction task.

Predicting anatomic structures and local features

A standard part of the clinical workflow of echocardiography interpretation is the identification of local cardiac structures and characterization of its location, size, and shape. Local cardiac structures can have significant variation in image characteristics, ranging from bright echos of metallic intracardiac structures to dark regions denoting blood pools in cardiac chambers. As our first task, we trained EchoNet on three classification tasks frequently evaluated by cardiologists that rely on recognition of local features (Fig. 2). Labels of the presence of intracardiac devices (such as catheters, pacemaker, and defibrillator leads), severe left atrial dilation, and left ventricular hypertrophy were extracted from the physician-interpreted report and used to train EchoNet on unlabeled apical-4-chamber input images. The presence of a pacemaker lead was predicted with high accuracy (AUC of 0.89, F1 score of 0.73), followed by the identification of a severely dilated left atrium (AUC of 0.85, F1 score of 0.68), and left ventricular hypertrophy (AUC of 0.75, F1 score of 0.57). Similarly high performance was achieved in predicting right atrium major axis length and left atrial volume estimate. Scatter plots are shown in the Supplemental Materials. To understand the model's predictions, we used gradient-based sensitivity map methods[25] to identify the regions of interest for the interpretation and show that EchoNet highlights relevant areas that correspond to intracardiac devices, the left atrium, and the left ventricle respectively. Models' prediction robustness was additionally examined with direct input image manipulations, including occlusion of human recognizable features, to validate that EchoNet arrives at its predictions by focusing on biologically plausible regions of interest.[26] For example, in the frames in Fig. 2 with pacemaker lead, when we manually mask out the lead in the frame, EchoNet changes its prediction to no pacemaker.

Table 1. Baseline characteristics of patients in the training and test datasets.

Characteristics	Complete data		A4C view data	
	Train data	Test data	Train data	Test data
Number of patients	2850	373	2546	337
Number of images	1,624,780	169,880	172,080	21,540
Sex (% Male)	52.4%	52.8%	52.2%	53.7%
Age: mean, years (std)	61.3 (17.2)	62.8 (16.8)	61.1 (17.1)	63.2 (16.9)
Weight: mean, Kg (std)	78.8 (22.7)	78.9 (20.8)	78.0 (21.7)	78.5 (20.2)
Height: mean, m (std)	1.69 (0.11)	1.69 (0.11)	1.69 (0.12)	1.69 (0.11)
BMI: mean (std)	27.3 (6.7)	27.5 (6.5)	27.1 (6.5)	27.3 (6.1)
Pacemaker or defibrillator lead (% Present)	13.2	14.7	13.1	15.1
Severe left atrial enlargement (% Present)	17.2	20.3	18.0	21.9
Left ventricular hypertrophy (% Present)	33.3	38.0	32.7	37.9
End diastolic volume, mL: mean (std)	94.3 (47.2)	94.6 (13.0)	95.1 (48.2)	96.9 (48.0)
End systolic volume, mL: mean (std)	45.6 (38.3)	46.2 (36.1)	46.0 (39.3)	47.0 (36.6)
Ejection fraction: mean (std)	55.2 (12.3)	54.7 (13.0)	55.1 (12.2)	54.8 (13.1)

Predicting cardiac function

Quantification of cardiac function is a crucial assessment addressed by echocardiography. However, it has significant variation in human interpretation.[18,19] The ejection fraction, a measure of the volume change in the left ventricle with each heart beat, is a key metric of cardiac function, but its measurement relies on the time-consuming manual tracing of left ventricular areas and volumes at different times during the cardiac cycle. We trained EchoNet to predict left ventricular end systolic volume (ESV), end diastolic volume (EDV), and ejection fraction from sampled apical-four-chamber view images (Fig. 3). Left ventricular ESV and EDV were accurately predicted. For the prediction of ESV, an R^2 score of 0.74 and mean absolute error (MAE) of 13.3 mL was achieved versus MAE of 25.4 mL if we use mean prediction which is to predict every patient's ESV as the average ESCV value of patients. The result for the EDV prediction was an R^2 score of 0.70 and MAE of 20.5 mL (mean prediction MAE = 35.4 mL). Conventionally, ejection fraction is calculated from a ratio of these two volumetric measurements, however, calculated ejection fraction from the predicted volumes were less accurate (Fig. 3c) than EchoNet trained directly on the ejection fraction (Fig. 3d). We show the relative performance of a deep learning model undergoing a standard human workflow of evaluating ESV and EDV then subsequently calculating ejection fraction from the two volumetric measurements vs. direct "end-to-end" deep learning prediction of ejection fraction and show that the "end-to-end" deep learning prediction model had improved performance. Using the trained EchoNet, an R^2 score of 0.50 and MAE of 7.0% is achieved (MAE of mean prediction = 9.9%). For each model, interpretation methods show appropriate attention over left ventricle as the region of interest to generate the predictions. A comparison of model performance based on number of sampled video frames did not show gain in model performance after 11 frames per prediction task.

Predicting systemic cardiovascular risk factors

With good performance in identifying local structures and estimating volumetric measurements of the heart, we sought to determine if EchoNet can also identify systemic phenotypes that modify cardiovascular risk. Previous work has shown that deep CNNs have powerful capacity to aggregate the information on visual correlations between medical imaging data and systemic phenotypes.[12] EchoNet predicted systemic phenotypes of age ($R^2 = 0.46$, MAE = 9.8 year, mean prediction MAE = 13.4 year), sex (AUC = 0.88), weight ($R^2 = 0.56$, MAE = 10.7 Kg, mean prediction MAE = 15.4 Kg), and height ($R^2 = 0.33$, MAE = 0.07 m, mean prediction MAE = 0.09 m) with similar performance to previous predictions of cardiac specific features (Fig. 4a). It is recognized that characteristics such as heart chamber size and geometry vary by age, sex, weight, and height,[27,28] however, human interpreters cannot predict these systemic phenotypes from echocardiogram images alone. We also investigated multi-task learning—sharing some of the model parameters while predicting across the different phenotypes—and this did not improve the model performance. Bland-Altman plots of the model accuracy in relationship to the predictions are shown in Fig. 5 and in the Supplemental Materials.

Lastly, we used the same gradient-based sensitivity map methods to identify regions of interest for models predicting systemic phenotypes difficult for human experts to predict. These regions of interest for these models tend to be more diffuse, highlighting the models for systemic phenotypes do not rely as much on individual features or local regions (Fig. 4b). The interpretations for models predicting weight and height had particular attention on the apex of the scanning sector, suggesting information related to the thickness and characteristics of the chest wall and extra-cardiac tissue was predictive of weight and height.

DISCUSSION

In this study, we show that deep CNNs trained on standard echocardiograms images can identify local features, human-interpretable metrics of cardiac function, and systemic phenotypes, such as patient age, sex, weight, and height. Our models achieved high prediction accuracy for tasks readily performed by human interpreters, such as estimating ejection fraction and chamber volumes and identifying of pacemaker leads, as well as for tasks that would be challenging for human interpreters, such as predicting systemic phenotypes from images of the heart alone. Unique from prior work in the field, instead of using hand-labeled outcomes, we describe and exemplify an approach of using previously obtained phenotypes and interpretations from clinical records for model training, which can allow for more external validity and more rapid generalization with larger training data sets.

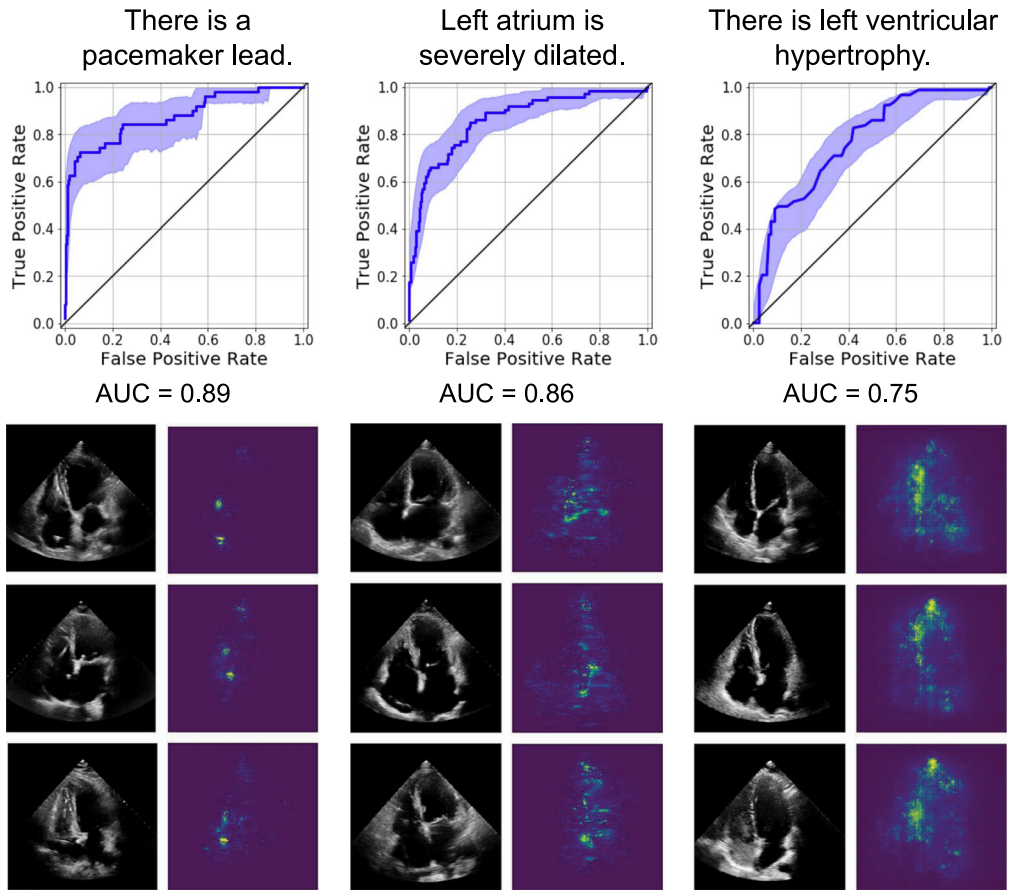

Fig. 2 EchoNet performance and interpretation for three clinical interpretations of local structures and features. For each task, representative positive examples are shown side-by-side with regions of interest from the respective model. Shaded areas indicate 95% confidence intervals.

One common critique of deep learning models on medical imaging datasets is the "black-box" nature of the predictions and the inability to understand the models ability to identity relevant features. In addition to showing the predictive performance of our methods, we validate the model's predictions by highlighting important biologically plausible regions of interest that correspond to each interpretation. These results represent the first presentation of interpretation techniques for deep learning models on echocardiographic images and can build confidence in simple models as the relevant pixels are highlighted when identifying local structures such as pacemaker leads. In addition, this approach of using interpretability frameworks to identify regions of interest may lay additional groundwork toward understanding human physiology when interpreting outputs of deep learning models for challenging, human-unexplainable phenotypes in medical imaging. These results represent a step towards automated image evaluation of echocardiograms through deep learning. We believe this research could supplement future approaches to screen for subclinical cardiovascular disease and understand the biological basis of cardiovascular aging.

While age, sex, weight, and height are relatively obvious visual phenotypes, our paper presents models predicting these systemic phenotypes in the roadmap of progressing from simple local feature based predictions, to more complex dynamic measurement predictions, and finally to human-difficult classifications of systemic phenotypes without obvious local features. Previous studies have shown that medical imaging of other organ systems can predict cardiovascular risk factors including age, gender, and blood pressure by identifying local features of systemic phenotypes.[12] Recently, 12-lead ECG based deep learning models have been shown to accurately predict age and sex, further validating a cardiac phenotype for aging and gender dysmorphism.[29] Our results identify another avenue of detecting systemic phenotypes through organ-system specific imaging. These results are supported by previous studies that showed population level normative values for the chamber sizes of cardiac structures as participants vary by age, sex, height, and weight.[27,28] Age-related changes in the heart, in particular changing chamber sizes and diastolic filling parameters, have been well characterized,[30,31] and our study builds upon this body of work to demonstrate that these signals are present to allow for prediction of these phenotypes to a degree of precision not previously reported. As systemic phenotypes of age, sex, and body mass index are highly correlated with cardiovascular outcomes and overall life expectancy, the ability of deep learning models to identify predictive latent features suggest that future work on image-based deep learning models can identify features hidden from human observers and predict outcomes and mortality.[32–34]

In addition to chamber size, extracardiac characteristics as well as additional unlabeled features, are incorporated in our models to predict patient systemic phenotypes. The area closest to the transducer, representing subcutaneous tissue, chest wall, lung parenchyma, and other extracardiac structures are highlighted in the weight and height prediction models. These interpretation maps are consistent with prior knowledge that obese patients often have challenging image acquisition,[35,36] however, it is surprising the degree of precision it brings to predicting height

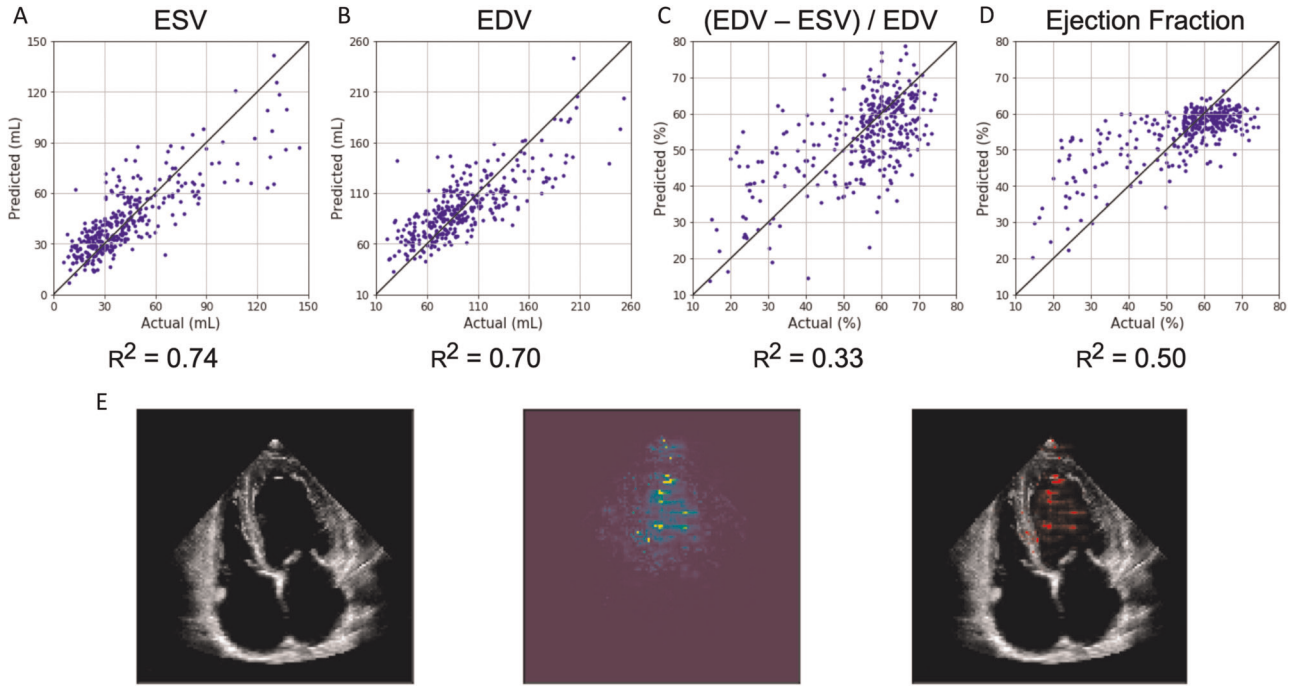

Fig. 3 EchoNet performance and interpretation for ventricular size and function. EchoNet performance for **a** predicted left ventricular end systolic volume, **b** predicted end diastolic volume, **c** calculated ejection fraction from predicted ESV and EDV, and **d** predicted ejection fraction. **e** Input image, interpretation, and overlap for ejection fraction model.

and weight. Retrospective review of predictions by our model suggest human-interpretable features that show biologic plausibility. In the saliency maps for the age prediction model, significant attention was paid to the crux of the heart, involving the intra-atrial septum, where the aortic annulus as the view becomes closer to an apical-five-chamber view, septal insertion of the mitral and tricuspid leaflets, and the mitral apparatus. This is an area of where differential calcification can be seen, particularly of the aortic valve and mitral annulus, and is known to be highly correlated with age-related changes.[37,38] Images predicted to be of younger patients also show preference for small atria and is consistent with prior studies showing age-related changes to the left atrium.[31,39] The feedback loop between physician and machine learning models with clinician review of appropriate and inappropriately predicted images can assist in greater understanding of normal variation in human echocardiograms as well as identify features previously neglected by human interpreters. Understanding misclassifications, such as patients with young biological age but high predicted age, and further investigation of extreme individuals can potentially help identify subclinical cardiovascular disease and better understand the aging process.

Prior foundational work on deep learning interpretation of echocardiogram images have focused on the mechanics of obtaining the correct echocardiographic view and hand-crafted scenarios with closely curated patient populations and multi-step processing and post-processing feature selection and calculation.[5,24] The work described here focuses on using more modern deep learning architectures and techniques in the framework of using previously adjudicated phenotypes with the potential of rapid scaling of algorithms to clinical practice. With the continued rapid expansion of computational resources, we were able to input higher resolution images (299×299 instead of 60×80 in prior studies)[24] and present an 'end-to-end' approach to predicting complex phenotypes like ejection fraction that has decreased variance over multi-step techniques which require identification of end-systole, end-diastole, and separate segmentation steps.[5]

While our model performance improves upon the results of prior work, EchoNet's evaluation of clinical measurements of ESV, EDV, and EF have non-negligible variance and does not surpass human assessment of these metrics. For these tasks, clinical context and understanding of contextual information and other measurements likely has significant relevance to the training task. For example, evaluation of EF as a ratio of ESV and EDV magnifies errors and performs worse than estimation of ESV or EDV individually. Future work requires greater integration of temporal information between frames to better assess cardiac motion and interdependencies in cardiac structures. In addition to quantitative measurements, human evaluation of cardiac structures, such as tracings of the left ventricle, are potentially high value training datasets.

Recent novel machine learning techniques for interpreting network activations are also presented for the first time to understand regions of interest in the interpretation of echocardiogram images.[25] While prior work used hand-labeled outcomes and patient cohorts for the majority of their outcome labels, we describe and showcase an approach of using previously obtained phenotypes and interpretations from clinical records for model training, which can allow for more external validity and more rapid generalization with larger training data sets. Additionally, given the significant difference between images in ImageNet vs. echocardiogram images, pretraining with ImageNet weights did not significantly help model performance, however, our models trained on systemic phenotypes can be good starting weights for future work on training on echocardiogram images of more complex phenotypes.

Previous studies of deep learning on medical imaging focused on resource-intensive imaging modalities common in resource-rich settings[40,41] or sub-speciality imaging with focused indication.[12,13,16] These modalities often need retrospective annotation by experts as the clinical workflow often does not require

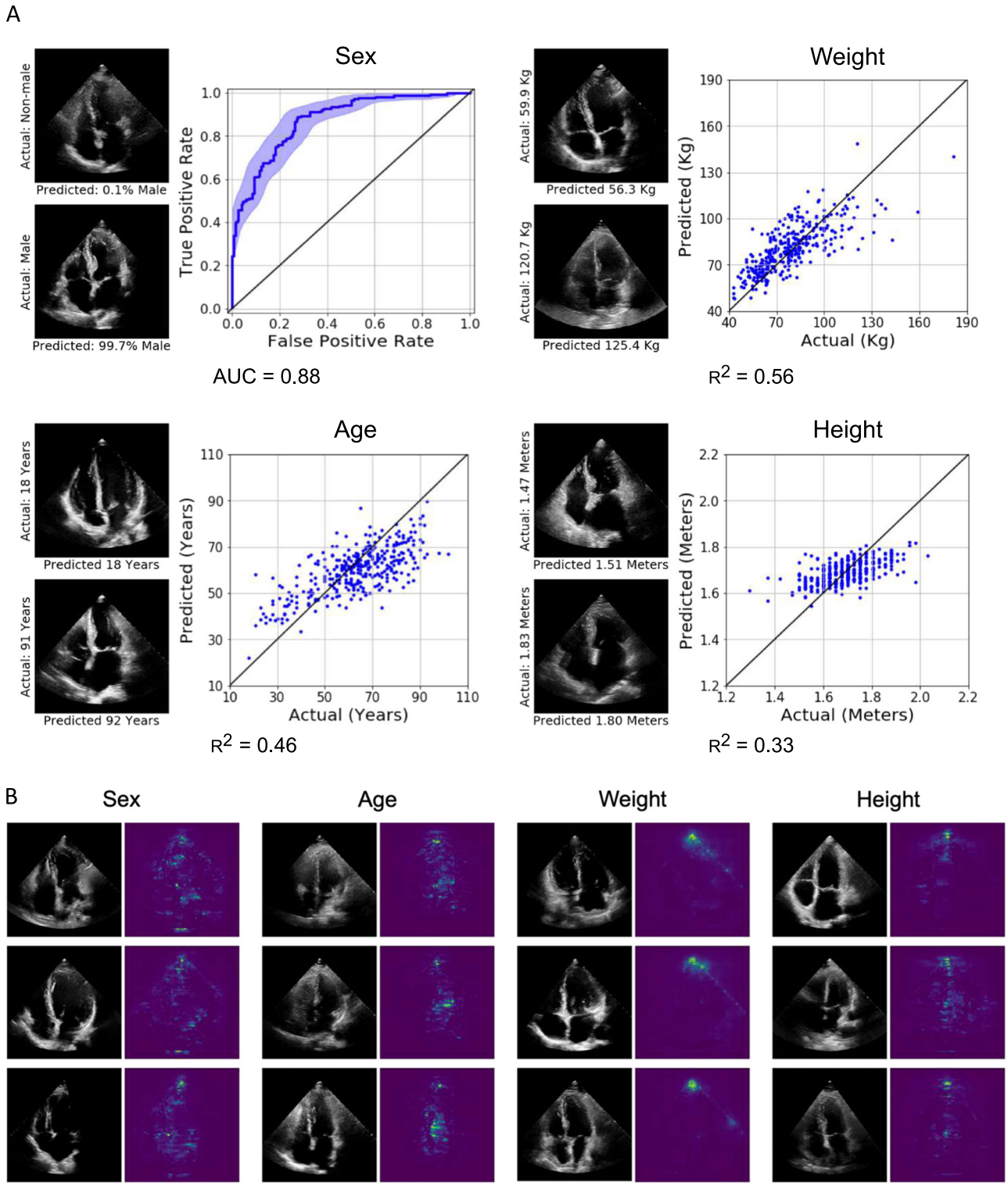

Fig. 4 EchoNet performance and interpretation for systemic phenotypes. a EchoNet performance for prediction of four systemic phenotypes (sex, weight, height and age) using apical-4-chamber view images. Shaded areas indicate 95% confidence intervals. **b** Interpretation of systemic phenotype models with representative positive examples shown side-by-side with regions of interest.

detailed measurements or localizations. In the development of any machine learning models to healthcare questions, external validity of first-order importance. An important caveat of our work is that the images obtained were from one type of ultrasound machine and our test dataset was of different patients but also scanned using the sample machine and at the same institution. Our approach trains deep learning models on previous studies and associated annotations from the EMR to leverage past data for rapid deployment of machine learning models. This approach leverages two advantages of echocardiography, first that echocardiography is one of the most frequently using imaging studies in the United States[42] and second,

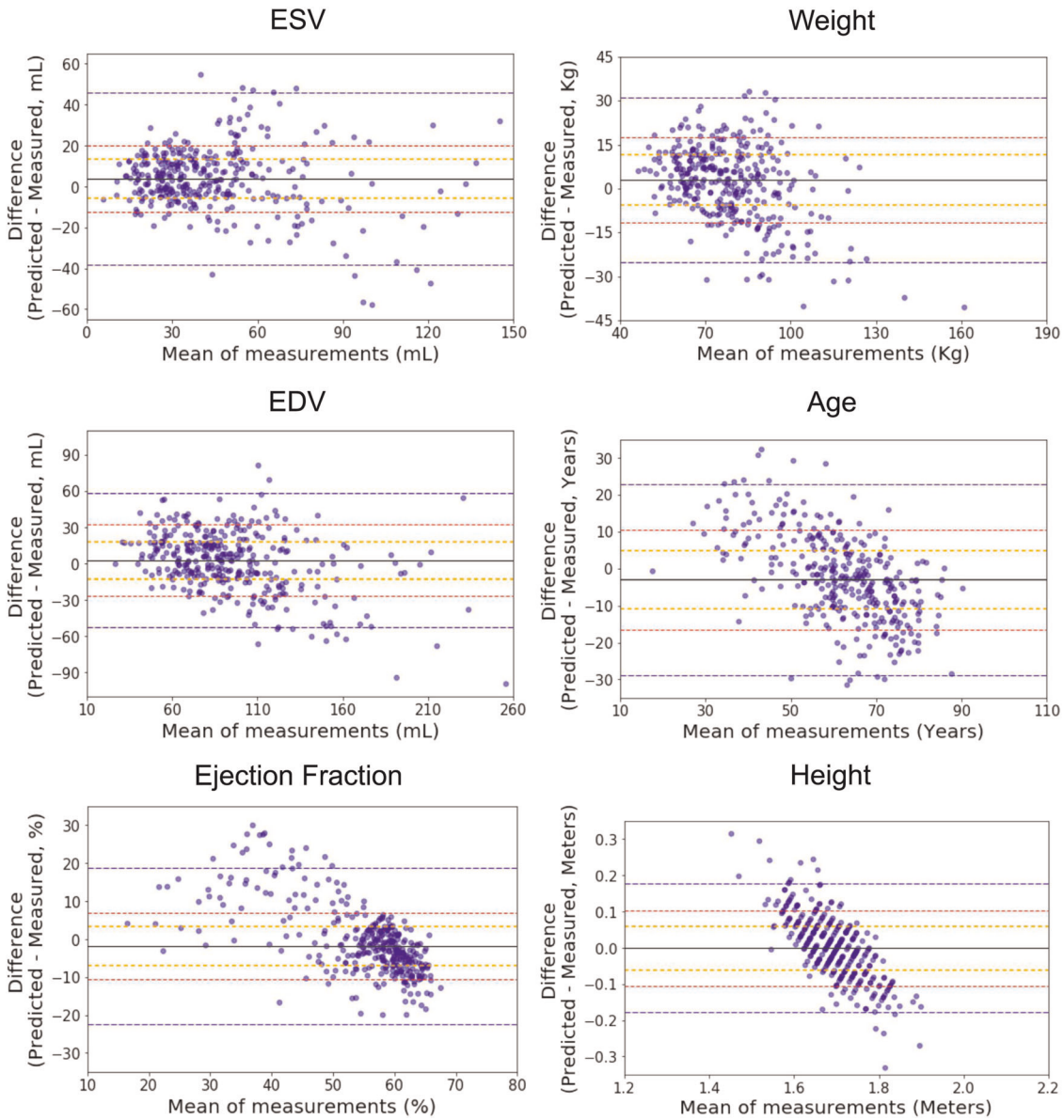

Fig. 5 Bland-Altman plotsBland-Altman plots of EchoNet performance for regression predictiontasks. The solid black line indicates the median. Orange, red, and blue dashed lines delineate the central 50%, 75%, and 95% of cases based on differences between automated and measured values.

echocardiography often uses structured reporting, making advances in deep learning particularly applicable and generalizable. However, such a method depends on the clinical standard, as there is known variability between MRI and echocardiography derived methods and training on clinical reports require rigorous quality control from the institution's echocardiography lab. Future work on deep learning of echocardiography would need to confirm the performance in broader populations and settings. Automation of echocardiography interpretation through deep learning can make cardiovascular care more readily available. With point-of-care ultrasound is being more frequently used by an increasing number of physicians, ranging from emergency room physicians, internists, to anesthesiologists, and deep learning on cardiac ultrasound images can provide accurate predictions and diagnoses to an even wider range of patients.

In summary, we provide evidence that deep learning can reproduce common human interpretation tasks and leverage additional information to predict systemic phenotypes that could allow for better cardiovascular risk stratification. We used interpretation methods that could feedback relevant regions of interest for further investigation by cardiologists to better understand aging and prevent cardiovascular disease. Our work could enable assessment of cardiac physiology, anatomy, and risk stratification at the population level by automating common workflows in clinical echocardiography and democratize expert interpretation to general patient populations.

METHODS

Dataset

The Stanford Echocardiography Database contains images, physician reports, and clinical data from patients at Stanford Hospital who underwent echocardiography in the course of routine care. The accredited echocardiography laboratory provides cardiac imaging to a range of patients with a variety of cardiac conditions including atrial fibrillation, coronary artery disease, cardiomyopathy, aortic stenosis, and amyloidosis.

For this study, we used 3312 consecutive comprehensive non-stress echocardiography studies obtained between June 2018 and December 2018, and randomly split the patients into independent training, validation, and test cohorts. Videos of standard cardiac views, color Doppler videos, and still images comprise each study and is stored in Digital Imaging and Communications in Medicine (DICOM) format. The videos were sampled to obtain 1,624,780 scaled 299×299 pixel images. The sampling rate was chosen to optimize model size and training time while maintaining model performance and additional preprocessing details are described in the Supplementary Materials. For each image, information pertained to image acquisition, identifying information, and other information outside the imaging sector was removed through masking. Human interpretations from the physician-interpreted report and clinical features from the electronic medical record were matched to each echocardiography study for model training. This study was approved by the Stanford University IRB. Written informed consent was waived for retrospective review of imaging obtained in the course of standard care.

Model

We chose a CNN architecture that balances network width and depth in order to manage the computational cost of training. We used the architecture based on Inception-Resnet-v1[10] to predict all of our phenotypes. This architecture has strong performance on benchmark datasets like ILSVR2012 image recognition challenge (Imagenet)[9] and is computationally efficient compared to other networks.[43] Pretraining Inception-ResNet with ImageNet did not significantly increase model performance, and our ultimate model used randomly initiated weights.

For each prediction task, one CNN architecture was trained on individual frames from each echocardiogram video with output labels that were extracted either from the electronic medical record or from the physician report. From each video, we sampled 20 frames (one frame per 100 milliseconds) starting from the first frame of the video. The final prediction was performed by averaging all the predictions from individual frames. Several alternative methods were explored in order to aggregate frame-level predictions into one patient-level prediction and did not yield better results compared to simple averaging.

Model training was performed using the TensorFlow library[44] which is capable of utilizing parallel-processing capabilites of Graphical Processing Units (GPUs) for fast training of deep learning models. We chose Adam optimizer as our optimization algorithm which is computationally efficient, has little memory usage, and has shown superior performance in many deep learning tasks.[45] As our prediction loss, we used cross-entropy loss for classification tasks and squared error loss for regressions tasks along with using weight-decay regularization loss to prevent over-fitting.[46] We investigated other variants of prediction loss (absolute loss, Huber loss[47] for regression and Focal loss[48] for classification), and they did not improve performance. For each prediction task, we chose the best performing hyper-parameters using grid search (24 models trained for each task) to optimize learning rate and weight decay regularization factor. In order to perform model selection, for each tasks, we split the training data into training and validation set by using 10% of train data as a held-out validation set in; the model with the best performance on the validation set is then examined on the test set to report the final performance results. After the models were trained, they were evaluated on a separate set of test frames gathered from echocardiogram studies of 337 other patients with similar demographics (Table 1). These patients were randomly chosen for a 10% held-out test set and were not seen by the model during training.

Data augmentation

Model performance improved with increasing input data sample size. Our experiments suggested additional relative improvement with increase in the number of patients represented in the training cohort compared to oversampling of frames per patient. Data augmentation using previously validated methods,[49,50] also greatly improving generalization of model predictions by reducing over-fitting on the training set. Through the training process, at each optimization step each training image is transformed through geometric transformations (such as flipping, reflection, and translation) and changes in contrast and saturation. As a result, the training data set is augmented into a larger effective data set. In this work, mimicking variation in echocardiography image acquisition, we used random rotation and random saturation augmentation for data augmentation (Fig. 1c). During each step of stochastic gradient descent in the training process, we randomly sample 24 training frames, and we perturb each training frame with a random rotation between -20 to 20 degrees and with adding a number sampled uniformly between -0.1 to 0.1 to image pixels (pixels values are normalized) to increase or decrease brightness of the image. Data augmentation results in improvement for all of the tasks; between 1–4% improvement in AUC metric for classification tasks and 2–10% improvement in R^2 score for regression tasks.

Cardiac view selection

We first tried using all echocardiogram images for prediction tasks but given the size of echocardiogram studies, initial efforts struggled with long training times, poor model convergence, and difficulty with model saturation. With the knowledge that, in a single comprehensive echocardiography study, the same cardiac structures are often visualized from multiple views to confirm and corroborate assessments from other views, we experimented with model training using subsets of images by cardiac view. As described in Fig. 1b, a selection of the most common standard echocardiogram views were evaluated for model performance. Images from each study were classified using a previously described supervised training method.[5] We sought to identify the most information-rich views by training separate models on the subsets of dataset images of only one cardiac view. Training a model using only one cardiac view results in one order of magnitude reduction of training time and computational cost with the benefit of maintaining similar predictive performance when information-rich views were used. For each of the prediction tasks and specific choice of hyper-parameters, training a model on the A4C-View data set converges in ~30 h using one Titan XP GPU. The training process of the same model and prediction task converges in ~240 h using all the views in the dataset. Given the favorable balance of performance to computational cost as well as prior knowledge on which views most cardiologists frequently prioritize, we chose the apical-four-chamber view as the input training set for subsequent experiments on training local features, volumetric estimates and systemic phenotypes.

Interpretability

Interpretability methods for deep learning models have been developed to explain the predictions of the black-box deep neural network. One family of interpretations methods are the sensitivity map methods that seek to explain a trained model's prediction on a given input by assigning a scalar importance score to each of the input features or pixels. If the model's input is an image, the resulting sensitivity map could be depicted as a two-dimensional heat-map with the same size as the image where more important pixels of the image are brighter than other pixels. The sensitivity map methods compute the importance of each input feature as the effect of its perturbation on model's prediction. If the pixel is not important, the change should be small and vice versa.

Introduced by Baehrens et al.[51] and applied to deep neural networks by Simonyan et al.,[52] the simplest way to compute such score is to have a first-order linear approximation of the model by taking the gradient of the output with respect to the input; the weights of the resulting linear model are the sensitivity of the output to perturbation of their corresponding features (pixels). More formally, given the d-dimensional input $\mathbf{x}_t \in \mathbb{R}^d$ and the model's prediction function $f(.)$, the importance score of the j'th feature is $|\nabla_{\mathbf{x}} f(\mathbf{x}_t)_j|$. Further extensions to this gradient method were introduced to achieve better interpretations of the model and to output sensitivity maps that are perceptually easier to understand by human users: LRP,[53] DeepLIFT,[54] Integrated Gradients,[55] and so forth. These sensitivity map methods, however, suffer from visual noise[25] and sensitivity to input perturbations.[56] SmoothGrad[25] method alleviates both problems[57] by adding white noise to the image and then take the average of the resulting sensitivity maps. In this work, we use SmoothGrad with the simple gradient method due to its computational efficiency. Other interpretation methods including Integrated Gradients were tested but did not result in better visualizations.

Lessons from model training and experiments

EchoNet performance greatly improved with efforts to augment data size, homogenize input data, and with optimize model training with hyperparameter search. Our experience shows that increasing number of unique patients in the training set can significantly improve the model, more so than increasing the sampling rate of frames from the same patients. Homogenizing the input images by selection of cardiac view prior to model training greatly improved training speed and decreased computational time without significant loss in model performance. Finally,

we found that results can be significantly improved with careful hyperparameter choice; between 7–9% in AUC metric for classification tasks and 3–10% in R^2 score for regression tasks.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The data comes from medical records and imaging from Stanford Healthcare and is not publicly available. The de-identified data is available from the authors upon reasonable request and with permission of the institutional review board.

CODE AVAILABILITY

The code is freely available at https://github.com/amiratag/EchoNet.

Received: 25 June 2019; Accepted: 19 December 2019;
Published online: 24 January 2020

REFERENCES

1. Heidenreich, P. et al. Forecasting the future of cardiovascular disease in the united states: a policy statement from the american heart association. *Circulation* **123**, 933–944 (2011).
2. Cohen, M. et al. Racial and ethnic differences in the treatment of acute myocardial infarction: findings from the get with the guidelines-coronary artery disease program. *Circulation* **121**, 2294–2301 (2010).
3. Havranek, E. et al. Social determinants of risk and outcomes of cardiovascular disease a scientific statement from the american heart association. *Circulation* **132**, 873–898 (2015).
4. Madani, A., Ong, J. R., Tiberwal, A. & Mofrad, M. R. US hospital use of echocardiography: Insights from the nationwide inpatient sample. *J. Am. Coll. Cardiol.* **67**, 502–511 (2016).
5. Zhang, J. et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation* **138**, 1623–1635 (2018).
6. Madani, A., Ong, J. R., Tiberwal, A. & Mofrad, M. R. Deep echocardiography: data-efficient supervised and semisupervised deep learning towards automated diagnosis of cardiac disease. *npj Digital Med.* **1**, 59 (2018).
7. Chen, J. H. & Asch, S. M. Machine learning and prediction in medicine-beyond the peak of inflated expectations. *N. Engl. J. Med.* **376**, 2507 (2017).
8. Dong, C., Loy, C.C., He, K. & Tang, X. Learning a deep convolutional network for image super-resolution. in *European conference on computer vision*, 184–199 (Springer, 2014).
9. Russakovsky, O. et al. Imagenet large scale visual recognition challenge. *Int. j. comp. vis.* **115**, 211–252 (2015).
10. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A.A., Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence* (AAAI.org, 2017).
11. Karpathy, A. et al. Large-scale video classification with convolutional neural networks. In *Proc. of the IEEE conference on Computer Vision and Pattern Recognition*, 1725–1732 (IEEE, 2014).
12. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158 (2018).
13. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115 (2017).
14. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559 (2018).
15. Ounkomol, C., Seshamani, S., Maleckar, M. M., Collman, F. & Johnson, G. R. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nat. Methods* **15**, 917 (2018).
16. Nagpal, K. et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *npj Digit. Med.* **2**, 48 (2019).
17. Douglas, P. et al. Accf/ase/aha/asnc/hfsa/hrs/scai/sccm/scct/scmr 2011 appropriate use criteria for echocardiography. *J. Am. Soc. Echocardiogr.* **24**, 229–267 (2011).
18. Wood, P.W., Choy, J.B., Nanda, N.C. & Becher, H. Left ventricular ejection fraction and volumes: it depends on the imaging method. *Echocardiography* **31**, 87–100 (2014).
19. Geer, D. D., Oscarsson, A. & Engvall, J. Variability in echocardiographic measurements of left ventricular function in septic shock patients. *J. Cardiovasc Ultrasound.* **13**, 19 (2015).
20. JA, A. & JM, G.-S. Echocardiographic variables used to estimate pulmonary artery pressure in dogs. *J. Vet. Intern. Med.* **31**, 1622–1628 (2017).
21. 2019 ACC/AHA/ASE advanced training statement on echocardiography (Revision of the 2003 ACC/AHA Clinical Competence Statement on Echocardiography): a report of the ACC competency management committee. *J. Am. Coll. Cardiol.* **19**, S0735–S1097 (2019)
22. MK, F., WS, B. & DN, W. Systematic review: prediction of perioperative cardiac complications and mortality by the revised cardiac risk index. *Ann. Intern. Med.* **152**, 26–35 (2010).
23. Abdel-Qadir, H. et al. A population-based study of cardiovascular mortality following early-stage breast cancer. *JAMA Cardiol.* **2**, 88–93 (2017).
24. Madani, A., Arnaout, R., Mofrad, M. & Arnaout, R. Fast and accurate view classification of echocardiograms using deep learning. *npj Digital Med.* **1**, 6 (2018).
25. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).
26. Abid, A. et al. Gradio: Hassle-free sharing and testing of ml models in the wild. in *Proc. 36th International Conference on Machine Learning,* Vol. 72 (JMLR.org, 2019).
27. Kou, S. et al. Echocardiographic reference ranges for normal cardiac chamber size: results from the norre study. *Eur. Heart J. Cardiovasc. Imaging* **15**, 680–690 (2014).
28. Pfaffenberger, S. et al. Size matters! Impact of age, sex, height, and weight on the normal heart size. *Circ. Cardiovasc. Imaging* **6**, 1073–1079 (2013).
29. Attia, Z. et al. Age and sex estimation using artificial intelligence from standard 12-lead ecgs. *Circ.: Arrhythm. Electrophysiol.* **12**, e007284 (2019).
30. Munagala, V. et al. Association of newer diastolic function parameters with age in healthy subjects: a population-based study. *J. Am. Soc. Echocardiogr.* **16**, 1049–1056 (2003).
31. D'Andrea, A. et al. Left atrial volume index in healthy subjects: clinical and echocardiographic correlates. *Echocardiography* **30**, 1001–1007 (2013).
32. Bhaskaran, K., dos Santos Silva, I., Leon, D. A., Douglas, I. J. & Smeeth, L. Body-mass index and mortality among 1.46 million white adults. *N. Engl. J. Med.* **363**, 2211–2219 (2010).
33. de Gonzalez A, B., P, H. & JR, C. Association of BMI with overall and cause-specific mortality: a population-based cohort study of 3.6 million adults in the UK. *Lancet Diabetes Endocrinol.* **6**, 944–953 (2018).
34. Xu, H., Cupples, L. A. & Stokes, A., & Liu, C.T. et al. Association of obesity with mortality over 24 years of weight history findings from the framingham heart study. *JAMA Netw. Open* **1**, e184587 (2018).
35. Madu, E. C. Transesophageal dobutamine stress echocardiography in the evaluation of myocardial ischemia in morbidly obese subjects. *Chest.* **117**, 657–661 (2000).
36. Medical Advisory Secretariat. Use of contrast agents with echocardiography in patients with suboptimal echocardiography. *Ont. Health Technol. Assess. Ser.* **10**, 1–17 (2010).
37. Kälsch, H. et al. Aortic calcification onset and progression: Association with the development of coronary atherosclerosis. *J Am Heart Assoc.* **6**, e005093 (2017).
38. Eleid, M.F., Foley, T.A., Said, S.M., Pislaru, S.V. & Rihal, C.S. Severe mitral annular calcification: multimodality imaging for therapeutic strategies and interventions. *JACC: Cardiovas. Imaging* **9**, 1318–1337 (2016).
39. Aurigemma, G. et al. Left atrial volume and geometry in healthy aging: the cardiovascular health study. *Circ. Cardiovasc. Imaging* **2**, 282–289 (2009).
40. Bello, G. A. et al. Deep-learning cardiac motion analysis for human survival prediction. *Nat. Mach. Intell.* **1**, 95 (2019).
41. Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
42. Virnig, B.A. et al. Trends in the Use of Echocardiography. Echocardiography Trends. Data Points #20 (prepared by the University of Minnesota DEcIDE Center, under Contract No. HHSA29020100013I). Rockville, MD: Agency for Healthcare Research and Quality; May 2014. AHRQ Publication No. 14-EHC034-EF (2007–2011).
43. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proc. of the IEEE conference on computer vision and pattern recognition*, 2818–2826 (IEEE, 2016).
44. Abadi, M. et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* 265–283 (2016).
45. Kingma, D.P. & Ba, J. Adam: a method for stochastic optimization. 3rd International Conference on Learning Representations, {ICLR} 2015, (San Diego, CA, USA, 2015) Conference Track Proceedings.
46. Krogh, A. & Hertz, J.A. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, 950–957 (1992).
47. Huber, P.J. Robust estimation of a location parameter. in *Breakthroughs in statistics*, 492–518 (Springer, 1992).
48. Lin, T.Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proc. of the IEEE international conference on computer vision,* 2980–2988 (IEEE, 2017).

49. Perez, L. & Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621* (2017).
50. Lim, S., Kim, I., Kim, T., Kim, C. & Kim, S. Fast AutoAugment In Advances in Neural Information Processing Systems, 6662–6672 (2019).
51. Baehrens, D. et al. How to explain individual classification decisions. *Journal of Machine Learning Research* **11**, 1803–1831 (2010).
52. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).
53. Bach, S. et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS ONE* **10**, e0130140 (2015).
54. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *Proc. of the 34th International Conference on Machine Learning*, vol. 70, 3145–3153 (JMLR, 2017)
55. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. in *Proc. 34th International Conference on Machine Learning,* Vol. 70, 3319–3328 (JMLR. org, 2017).
56. Ghorbani, A., Abid, A. & Zou, J. Interpretation of neural networks is fragile. In *Proc. of the AAAI Conference on Artificial Intelligence*, Vol. 33, 3681–3688 (AAAI.org, 2019).
57. Levine, A., Singla, S. & Feizi, S. Certifiably robust interpretation in deep learning. *arXiv preprint arXiv:1905.12105* (2019).

ACKNOWLEDGEMENTS

This work is supported by the Stanford Translational Research and Applied Medicine pilot grant and an Stanford Artificial Intelligence in Imaging and Medicine Center seed grant. D.O. is supported by the American College of Cardiology Foundation/ Merck Research Fellowship. A.G. is supported by the Stanford-Robert Bosch Graduate Fellowship in Science and Engineering. J.Y.Z. is supported by NSF CCF 1763191, NIH R21 MD012867-01, NIH P30AG059307, and grants from the Silicon Valley Foundation and the Chan-Zuckerberg Initiative.

AUTHOR CONTRIBUTIONS

Initial study concept and design: D.O. Acquisition of data: D.O., D.H.L. Model training: A.G., J.Y.Z. Analysis and interpretation of data: A.G., D.O., E.A.A., and J.Y.Z. Drafting of the paper: A.G., D.O. Critical revision of the manuscript for important intellectual content: A.G., D.O., A.A., B.H., J.H.C., R.A.H., D.H.L., E.A.A. and J.Y.Z. Statistical analysis: A. G., D.O., E.A.A., J.Y.Z.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41746-019-0216-8.

Correspondence and requests for materials should be addressed to D.O. or J.Y.Z.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.