npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Fast and accurate view classification of echocardiograms using deep learning

Ali Madani[1], Ramy Arnaout[2], Mohammad Mofrad[1] and Rima Arnaout[3]

Echocardiography is essential to cardiology. However, the need for human interpretation has limited echocardiography's full potential for precision medicine. Deep learning is an emerging tool for analyzing images but has not yet been widely applied to echocardiograms, partly due to their complex multi-view format. The essential first step toward comprehensive computer-assisted echocardiographic interpretation is determining whether computers can learn to recognize these views. We trained a convolutional neural network to simultaneously classify 15 standard views (12 video, 3 still), based on labeled still images and videos from 267 transthoracic echocardiograms that captured a range of real-world clinical variation. Our model classified among 12 video views with 97.8% overall test accuracy without overfitting. Even on single low-resolution images, accuracy among 15 views was 91.7% vs. 70.2–84.0% for board-certified echocardiographers. Data visualization experiments showed that the model recognizes similarities among related views and classifies using clinically relevant image features. Our results provide a foundation for artificial intelligence-assisted echocardiographic interpretation.

npj Digital Medicine (2018)1:6 ; doi:10.1038/s41746-017-0013-1

INTRODUCTION

Imaging is a critical part of medical diagnosis. Interpreting medical images typically requires extensive training and practice and is a complex and time-intensive process. Deep learning, specifically using convolutional neural networks (CNNs), is a cutting-edge machine learning technique that has proven "unreasonably"[1] successful at learning patterns in images and has shown great promise helping experts with image-based diagnosis in radiology, pathology, and dermatology, for example, in detecting the boundaries of organs in computed tomography and magnetic-resonance images, flagging suspicious regions on tissue biopsies, and classifying photographs of benign vs. malignant skin lesions.[2–4] However, deep learning has not yet been widely applied to echocardiography, a noninvasive, relatively inexpensive, radiation-free imaging modality that is an indispensable part of modern cardiology.[5]

A transthoracic echocardiogram (TTE) consists of scores of video clips, still images, and Doppler recordings measured from over a dozen different acquisition angles, offering complementary views of the heart's complex anatomy. The majority of the acquired information is represented as video clips; only pulsed-wave Doppler (PW), continuous-wave Doppler (CW), and m-mode recordings are represented exclusively as single images. Determining the view is the essential first step in interpreting an echocardiogram.[6] This step is non-trivial, not least because several views differ only subtly from each other. In principle, a CNN can be trained to classify views, requiring only a training set of labeled images from which to learn; given a new image, a well-trained model should then be able determine the view almost instantaneously. The versatility of training in deep learning represents a significant advantage over earlier machine-learning methods, which have sometimes been applied to echocardiography. Previous methods often require time-consuming and operator-dependent manual selection and annotation of features (e.g. manually tracing the outline of the heart) in each of a large number of training images, and are out-performed by deep learning on complex, high-dimensional problems, such as image recognition.[7–11]

To assist echocardiographers and improve use of echocardiography for precision medicine, we tested whether supervised deep learning with CNNs can be used to automatically classify views without requiring prior manual feature selection. We report a model that achieves nearly 98 percent overall test accuracy based on a variety of video and still-image view-classification tasks.

To achieve translational impact in medicine, novel computational models must not just achieve high accuracy but must also address clinical relevance. We did this in three main ways. First, we used randomly selected, real-world echocardiograms to train our model, including a variety of patient variables, echocardiographic indications and pathologies, technical qualities, and multiple vendors to ensure that our deep learning model would be clinically relevant. Second, deep learning approaches are often considered "data hungry;" we sought to achieve high accuracy on view classification with minimal data. Third, deep-learning models are sometimes considered "black boxes" because their internal workings are at first glance obscure. To address this issue, we used several methods to look inside our model to show that classification depends on human-recognizable clinical features within images.

Taken together, these results suggest that our approach may be useful in helping echocardiographers improve their accuracy,

[1]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Lab, California Institute for Quantitative Biosciences (QB3), University of California at Berkeley, 208A Stanley Hall Room 1762, Berkeley, CA 94720, USA; [2]Beth Israel Deaconess Medical Center, Harvard Medical School, 330 Brookline Avenue Dana 615, Boston, MA 02215, USA and [3]Cardiovascular Research Institute, University of California, 555 Mission Bay Blvd South Rm 484, San Francisco 94143, USA
Correspondence: Mohammad Mofrad (mofrad@berkeley.edu) or Rima Arnaout (rima.arnaout@ucsf.edu)

Received: 6 August 2017 Revised: 6 November 2017 Accepted: 17 November 2017
Published online: 21 March 2018

efficiency, and workflow and provide a foundation for high-throughput analysis of echocardiographic data.

RESULTS

Deep learning achieves expert-level view classification

We designed and trained a convolutional neural network (CNN) (Fig. 1) to recognize 15 different standard echocardiographic views, 12 from b-mode (video and still image) and three from pulsed-wave Doppler (PW), continuous-wave Doppler (CW), and m-mode (still image) recordings (Fig. 2), using a training and validation set of over 200,000 images (240 studies) and a test set of over 20,000 images (27 studies). To maintain sample independence, each echocardiogram was from a different patient, and training, validation and test sets did not overlap by patient or study (Fig. 1b). These images covered a range of natural echocardiographic variation with patient variables (Table 1) and indications for imaging (Table 2) that represented our overall clinical database, and they included differences in zoom, depth, focus, sector width, gain, chroma map, systole/diastole, angulation, image quality, and use of 3D, color Doppler, dual mode, strain, and LV contrast (Fig. 3). Clustering analyses showed that the neural network could sort heterogeneous input images into groups according to view (Fig. 4).

The model achieved an average overall test accuracy of 97.8 percent on videos (*F*-score 0.964 ± s.d. 0.035) and 100 percent accuracy on seven of the 12 video views (Fig. 5a). CW, PW, and m-mode categories, which always appeared in echocardiograms as still images, had 98, 83, and 99 percent accuracies, respectively (Fig. 5b). Classification of test images by the trained model took an average of 21 ms per image on a standard laptop (see section "Methods").

On single still images drawn from all 15 views, the model achieved an average overall accuracy of 91.7 percent (*F*-score 0.904 ± s.d. 0.058) (Fig. 5b), compared to an average of 79.4 percent (range, 70.2–84.0; $n = 4$ subjects) for board-certified echocardiograpers classifying a subset of the same test images (one-sample *t*-test, $p = 0.03$) (Fig. 5c). Associated areas under the curve (AUCs) for still-image model prediction by view category ranged from 0.985 to 1.00 (mean 0.996; Fig. 5f). For the 8.3 percent of test images that the model misclassified, its second-best guess —the view with the second-highest probability—was the correct one in 67.0 percent of cases (5.3 percent of test images; Fig. 5e). Therefore, 97.3 percent of test still-images were classified correctly when considering the model's top two guesses.

Accuracy was highest for views with more training data (e.g. apical four-chamber) and views that are most visually distinct from the others (e.g. m-mode). Accuracy was lowest for views that were clinically similar to other views, such as apical three-chamber (which can be confused for apical two-chamber) and apical four-chamber (vs. apical five-chamber), or views in which multiple view-defining structures can be seen in the same image, such as subcostal IVC vs. subcostal four-chamber. As expected, training on randomly labeled still images achieved an accuracy (6.9 percent) commensurate with random guessing (6.7 percent, the probability of guessing the correct one out of 15 views by chance).

Model classification is based on cardiac image regions

To understand whether classification is based on clinically relevant features, such as heart chambers and valves, or on confounding or statistical features that might be clearer to a machine than a human, such as fiducial markings, border regions, or fraction of white pixels, we performed occlusion experiments by measuring prediction performance on test images on which we masked clinically relevant features with different shapes. Overall test accuracy fell significantly with masking of the heart but not other parts of the image, consistent with this region being important to the model (Fig. 6a). In addition, saliency mapping, which identifies the input pixels that are most important to the model's assignment of a particular classification, revealed that structures that would be important to defining the view to a human expert were also the ones that contributed most to the model's classification (Fig. 6b).

DISCUSSION

View classification is the essential first step in interpreting echocardiograms. Previous attempts to use machine learning to assist with view classification required laborious manual annotation, failed to distinguish among more than a few views at a time, used only "textbook-quality" images for training, exhibited low accuracy, or were tied to a specific equipment vendor, limitations unsuitable for general practice.[7–13] In contrast, we report here a single, vendor-agnostic deep-learning model that correctly classifies all types of echocardiogram recordings (b-mode, m-

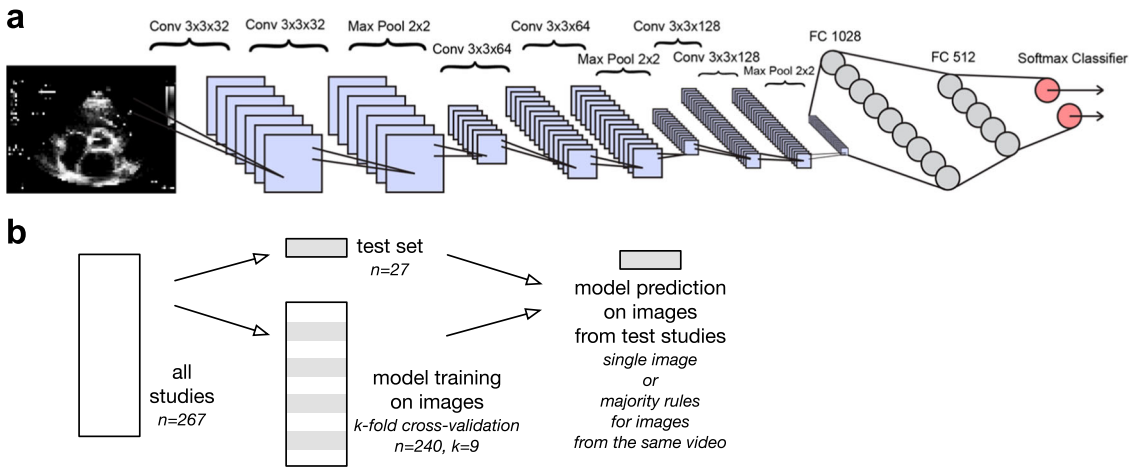

Fig. 1 Convolutional neural net architecture for image classification. **a** The neural network algorithm used for classification included six convolutional layers and two fully-connected layers of 1028 and 512 nodes, respectively. The softmax classifier (pink circles) consisted of up to 15 nodes, depending on the classification task at hand. **b** Training, validation, and test data were split by study, and test data was not used for training or validating the model. The model was trained to classify images, with video classification as a majority rules vote on related image frames. Conv convolutional layer, Max Pool max pooling layer, FC fully connected layer

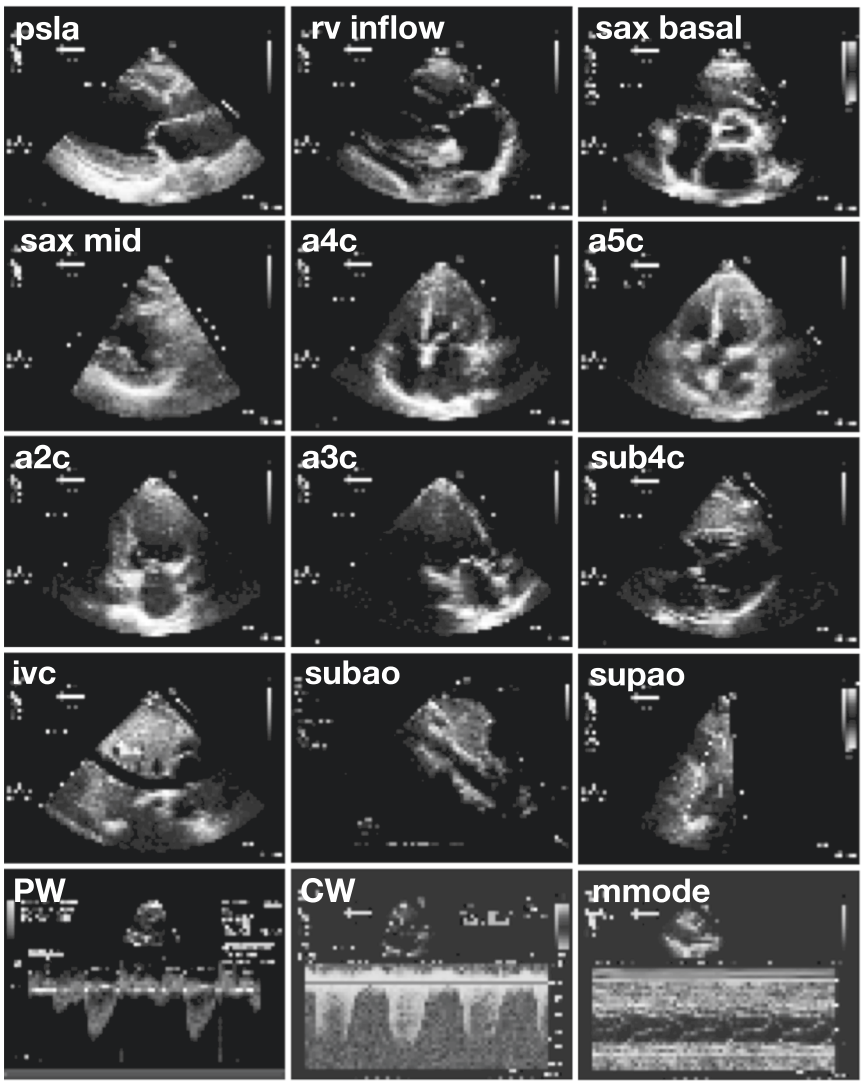

Fig. 2 Sample input images. Views classified included parasternal long axis (psla), right ventricular inflow (rv inflow), basal short axis (sax basal), short axis at mid or mitral level (sax mid), apical four-chamber (a4c), apical five chamber (a5c), apical two chamber (a2c), apical three chamber/apical long axis (a3c), subcostal four-chamber (sub4c), subcostal inferior vena cava (ivc), subcostal/abdominal aorta (subao), suprasternal aorta/aortic arch (supao), pulsed-wave Doppler (PW), continuous-wave Doppler (CW), and m-mode (mmode). Note that these images are the actual resolution of input data to the deep learning algorithm

mode, and Doppler; still images and videos) from all acquisition points relevant to a full standard transthoracic echocardiogram (parasternal, apical, subcostal, and suprasternal), at accuracies that exceed those of board-certified echocardiographers given the same task. Furthermore, the echocardiograms used in this study were drawn randomly from real echocardiograms acquired for clinical purposes, from patients with a range of ages, sizes, and hemodynamics; for a range of indications; and including a range of pathologies, such as low left ventricular ejection fraction, left ventricular hypertrophy, valve disease, pulmonary hypertension, pericardial effusion. Training data also included the natural variation in echocardiographic acquisition of each view, including variations in technical quality. By avoiding limited or idealized training subsets, our model is broadly applicable to clinical practice, although of course a larger training set would likely capture still more echocardiographic variability.

Because deep networks like CNNs usually include large numbers of (highly correlated) parameters (which describe the weights of connections among the nodes in the network), it is usually difficult to understand a model's decision-making by simple inspection. For life-or-death decisions, such as in medicine or self-driving cars, this issue can breed suspicion and has legal ramifications that can slow adoption. Occlusion testing and saliency mapping help address these concerns by getting inside the black box. In our model, these techniques show that classification depends on the same features that echocardiographers use to reach their conclusions. For example, the maps shown in Fig. 6b for a short-axis-mid view and a suprasternal aorta view, respectively, each trace the basic outlines of their corresponding input view. In the future, applying these approaches to intermediate layers may prove interesting to more precisely define the similarities, or differences, in how humans and models move from features to conclusions. For now, it is reassuring that our model considers the same features that human experts do in classifying views.

This similarity also explains the occasional misclassifications of single images, which most often involved views that can look similar to human eyes (Figs. 2e, f, g, h, j, k and 5). These include adjacent views in echocardiographic acquisition, where a slight difference in the angle of the sonographer's wrist can change

Table 1. Comparison of study sample characteristics to clinical echo database

Demographics	Study sample			Clinical Echo Database[a]			
	Mean	SD	IQR	Mean	SD	IQR	*p*-value[b]
Age (years)	56.1	16.6	22.5	58.5	16.8	23.0	0.5
Height (cm)	170	11.6	16.5	169	11.0	17.8	0.8
Weight (kg)	77.0	20.5	31.5	77.0	22.0	26.3	0.9
Systolic BP (mmHg)	127	19.0	20.3	126	22.0	28.0	1.0
Diastolic BP (mmHg)	70.0	123	13.3	70.0	12.0	17.0	0.5
MAP (mmHg)	88.9	13.4	18.3	88.6	13.9	13.8	0.7
BSA (m^2)	1.87	0.27	0.44	1.82	0.56	0.37	0.8
BMI (kg/m^2)	26.6	6.10	10.2	27.1	6.80	7.40	0.9

Demographics	Percent	*N*	Sample size	Percent	*N*	Sample size	*p*-value[c]
Female	50.6	135	267	49.5	79,460	159,503	0.7
Male	49.4	132	267	50.5	80,043	159,503	0.7
Obese	25.8	69	267	25.1	25,770	102,669	0.8

Pathology	Percent	*N*	Sample size	Percent	*N*	Sample size	*p*-value[c]
LVMI > normal (adjusted for sex)	32.8	67	204	39.2	34,056	86,878	0.06
LVEF < 55%	21.7	58	267	20.3	18,432	90,798	0.6
LVEDVI > normal (adjusted for sex)	16.9	45	238	46.8	10,677	90,375	0.3
RVSP > 40 mmHg	10.9	29	267	14.6	10,774	73,795	0.09
TAPSE < 1.6 cm	7.84	8	102	10.6	1768	16,679	0.4

BP blood pressure, *MAP* mean arterial pressure, *BSA* body surface area, *BMI* body mass index, *LVMI* left ventricular mass index (g/m^2), *LVEF* left ventricular ejection fraction, *LVEDVI* left ventricular end-diastolic volume index (ml/m^2), *RVSP* right ventricular systolic pressure, *TAPSE* tricuspid annular plane systolic excursion, *SD* standard deviation, *IQR* interquartile range
[a] *Ns* and sample size vary according to availability of different measurements
[b] Two-tailed Student's *t*-test, unequal variance
[c] Chi-squared test for comparison of proportions

the view, resulting in confusion of an apical three-chamber view for an apical two-chamber view or an apical five-chamber for apical four-chamber; as well as views in which two view-defining structures may be seen in the same image, such as the IVC seen in a subcostal four-chamber view. A low-velocity PW signal can look similar to a faint CW signal. In fact, the only misclassification made by our model without an obvious explanation of this sort was that of the right ventricular inflow view for short-axis basal; of note, the right ventricular inflow view was also very challenging for human echocardiographers to distinguish (with 51–57 percent accuracy). We note in the confusion matrices that misclassification of certain views for one another was non-symmetrical; for example, PW images were confused with CW, but CW images were almost never mistaken for PW (Fig. 5b). In this case, as mentioned above, this asymmetry makes clinical sense; however, more training and test data can be used to explore this phenomenon further and refine accuracies for these categories. Because classification of videos is based on multiple images, and error decays exponentially with the number of images, misclassification of videos was very rare (~2 percent; Fig. 5d). We also noted that the model's confidence in its choice (the probability assigned to a view classification for a particular image) affected performance; where confidence was higher, accuracy was also higher (Supplementary Fig. 1). Therefore, communicating the model's confidence for each classification should further benefit users.

Finally, our approach had two unexpected advantages related to efficiency, practicability, and cost-effectiveness. First was the perhaps surprising effectiveness of a simple majority vote in classification of videos. Video analysis can be a complex undertaking that involves non-trivial tasks, such as frame-to-frame color variation and object tracking. We have demonstrated that view classification, at least, can be done much more efficiently and cost-effectively, reducing coding and training time. Moving beyond view classification, it will be interesting to see what other clinically actionable information can be extracted from (collections of) still images. Second, in removing color and in standardizing the sizes and shapes of videos and still images for training, we discovered that we could downsample—i.e., shrink—images appreciably without losing accuracy. This allowed for a 96–99 percent savings in file size (vs. 300-by-400- to 1024-by-768-pixel images; Supplementary Fig. 2), and corresponding gains in the cost and speed of training and of classifying new samples at deployment. While human echocardiographers routinely classify views, they appear to require full-resolution, native video data to do so with high accuracy. With less input data, the model outperformed overall human accuracy (and speed: 32 s vs. hours to classify the same 1500-image test sample). We note the potential implications for telemedicine and global health, including in resource-poor regions of the United States, of requiring storage and transmission of smaller files (though decentralized use of the model can also come through transmission of the model, which is a small file), and of embracing older echo machines that may image with lower resolution.

Echocardiography is essential to diagnosis and management for virtually every cardiac disease. In this study, we have demonstrated the application of deep learning to echocardiography view classification that classified 15 major TTE views with expert-level quality. We purposely used a training set that reflected a wide

Table 2. Indications for study sample echocardiograms

Study sample indication	Percent	N
Heart failure/cardiomyopathy	24.0	64
Arrhythmia	11.6	31
Chemotherapy	10.9	29
Valve disease	10.5	28
Preoperative exam	7.9	21
Dyspnea	6.4	17
Coronary artery disease	6.0	16
Stroke	6.0	16
Syncope	5.2	14
Rule out endocarditis	4.9	13
Pulmonary HTN	4.5	12
Hypertension	3.7	10
Pericardial effusion	3.4	9
Murmur	3.0	8
Palpitations	3.0	8
Aortic aneurysm	2.6	7
Congenital heart disease	2.6	7
Lung disease	1.9	5
Edema	1.5	4
Hypotension	1.5	4
Cardiac arrest	0.4	1
Heart transplant	0.4	1

Number of normal studies[a]	Percent	N
Normal studies	10.9	29

[a] Defined by echo reports documenting normal four-chamber size and systolic/diastolic function, no chamber hypertrophy or wall motion abnormalities, normal valves with trace or less regurgitation, normal great vessels and estimated right atrial pressure, no pericardial effusion, RVSP < 40, and no other abnormalities, such as atherosclerosis, calcification, pleural effusion, ascites, prostheses, or catheters

range of clinical and physiological variations, demonstrating applicability to real-world data. We found that our model uses some of the same features in echocardiograms that human experts use to make their decisions. Looking forward, our model can be expanded to classify additional sub-categories of echocardiographic view (e.g. to distinguish among different CW, PW, and m-mode acquisitions), as well as diseases, work that has foundational utility for research, for clinical practice, and for training the next generation of echocardiographers.

METHODS

Dataset

All datasets were obtained and de-identified, with waived consent in compliance with the Institutional Review Board (IRB) at the University of California, San Francisco (UCSF). Methods were performed in accordance with relevant regulations and guidelines. Two-hundred sixty-seven echocardiographic studies from different patients and performed between 2000 and 2017 were selected at random from UCSF's clinical database. These studies included men and women (49.4 and 50.6 percent, respectively) ages 20–96 (median age, 56; mode, 63) with a range of body types (25.8 percent obese), which can affect technical quality of TTE (Table 1), and included indications and pathologies that are representative of the uses of echocardiography in current clinical practice (Table 2). Studies were carried out using echocardiograms acquired with equipment from several manufacturers (e.g., GE, Philips, Siemens).

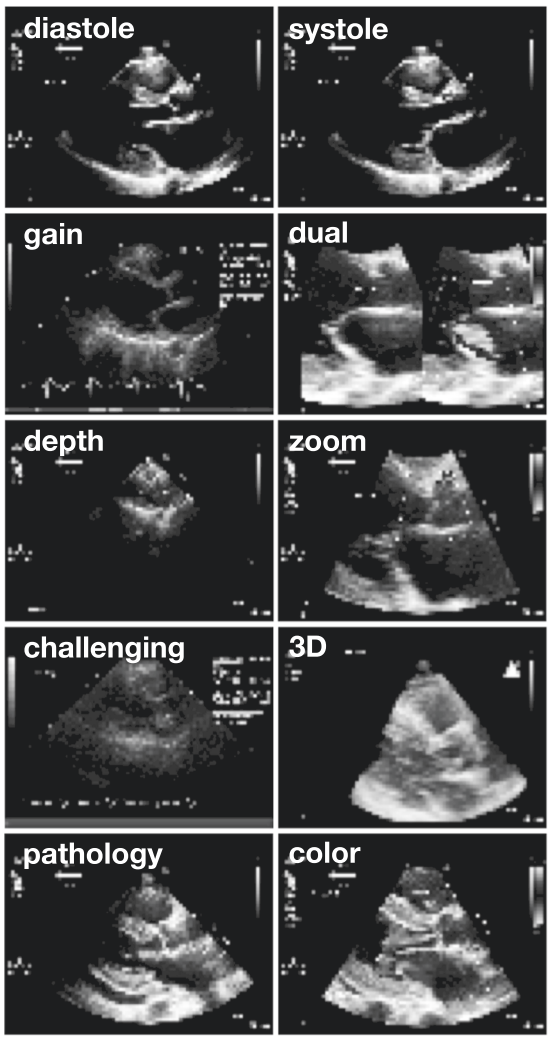

Fig. 3 Natural variations in input data. In addition to applying data augmentation algorithms, we included in each category a range of images representing the natural variation seen in real-life echocardiography. The parasternal long-axis view is shown here for example. Variations include a range of timepoints spanning diastole and systole, differences in gain or chroma map, use of dual-mode acquisition, differences in depth and zoom, technically challenging images, use of 3D acquisition, a range of pathologies (seen here, concentric left ventricular hypertrophy and pericardial effusion), and use of color Doppler, as well as differences in angulation, sector width, and use of LV contrast. Note that these images are the actual resolution of input data to the deep learning algorithm

Data processing

DICOM-formatted echocardiogram videos and still images were stripped of identifying metadata, anonymized by zeroing out all pixels that contained identifying information, labeled by view by a board-certified echocardiographer with access to native-resolution and video data, then split into constituent frames and converted into standardized 60 × 80-pixel monochrome images, resulting in 834,267 images. Fifteen views were selected for multi-category classification, covering the majority used in the field. Views classified included parasternal long axis, right ventricular inflow, basal short axis (aortic valve level), short axis at mid (papillary muscle) or mitral level, apical four-chamber, apical five chamber, apical two chamber, apical three chamber (apical long axis), subcostal four-chamber, subcostal inferior vena cava (IVC), subcostal abdominal aorta, suprasternal aortic arch, pulsed-wave Doppler, continuous-wave Doppler, and m-mode. For the purposes of this study, CW Doppler, PW Doppler, and m-mode recordings from different acquisition points were considered part of the same "view," e.g. m-mode of the aortic valve, mitral valve, left ventricle,

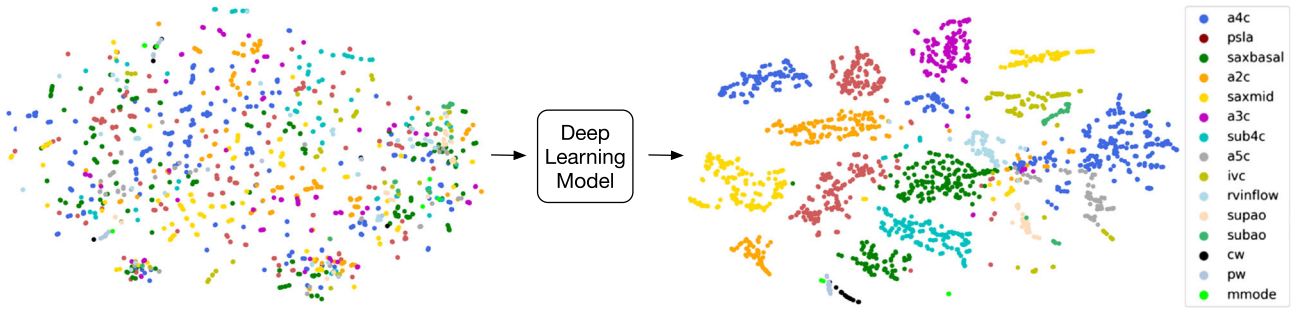

Fig. 4 Deep learning model simultaneously distinguishes among 15 standard echocardiographic views. We developed a deep-learning method to classify among standard echocardiographic views, represented here by t-SNE clustering analysis of image classification. On the left, t-SNE clustering of input echocardiogram images. Each image is plotted in 4800-dimensional space according to the number of pixels, and projected to two-dimensional space for visualization purposes. Different colored dots represent different view classes (see legend in figure). Prior to neural network analysis, input data does not cluster into clear groups. On the right, data as processed through the last fully connected layer of the neural network are again represented in two-dimensional space, showing organization into clusters according to view category. *Abbreviations*: a4c apical 4 chamber, psla parasternal long axis, saxbasal short axis basal, a2c apical 2 chamber, saxmid short axis mid/mitral, a3c apical 3 chamber, sub4c subcostal 4 chamber, a5c apical 5 chamber, ivc subcostal ivc, rvinflow right ventricular inflow, supao suprasternal aorta/aortic arch, subao subcostal/abdominal aorta, cw continuous-wave Doppler, pw pulsed-wave Doppler, mmode m-mode recording

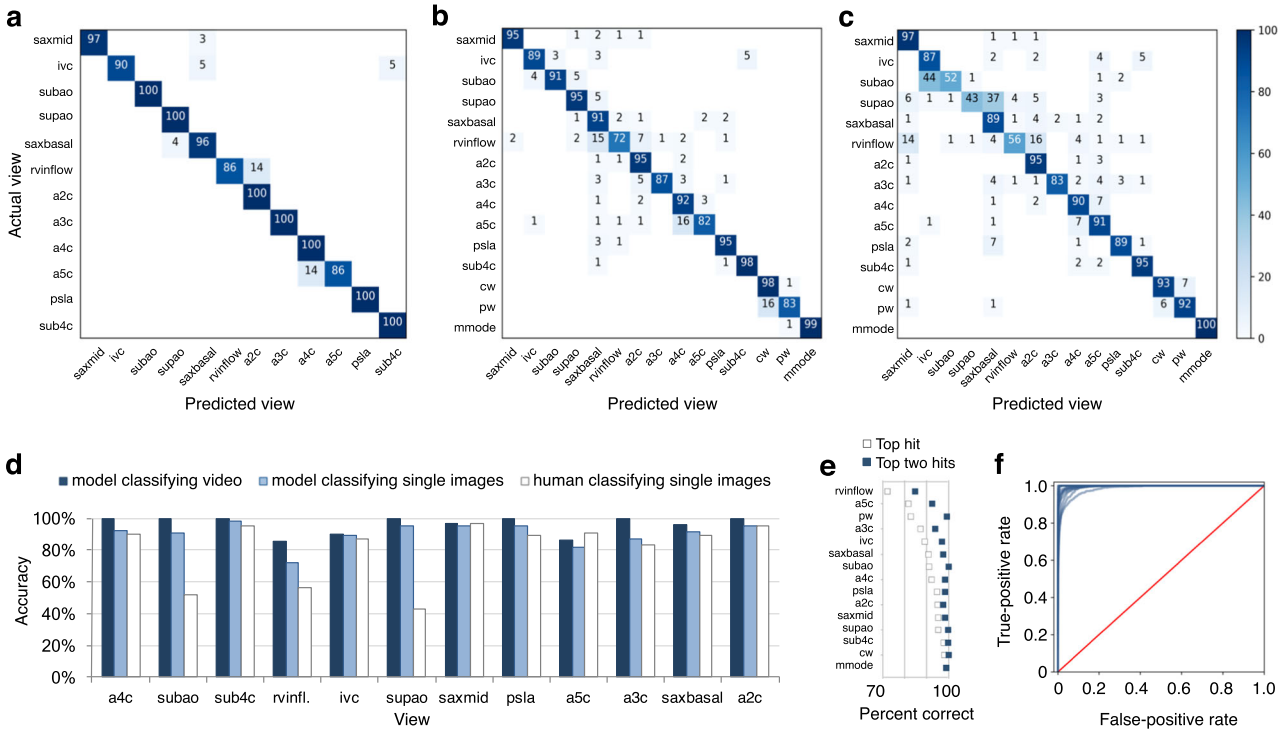

Fig. 5 Echocardiogram view classification by deep-learning model. Confusional matrices showing actual view labels on *y*-axis, and neural network-predicted view labels on the *x*-axis by view category for video classification (**a**) and still-image classification (**b**) compared with a representative board-certified echocardiographer (**c**). Reading across true-label rows, the numbers in the boxes represent the percentage of labels predicted for each category. Color intensity corresponds to percentage, see heatmap on far right; the white background indicates zero percent. Categories are clustered according to areas of the most confusion. Rows may not add up to 100 percent due to rounding. **d** Comparison of accuracy by view category for deep-learning-assisted video classification, still-image classification, and still-image classification by a representative echocardiographer. **e** A comparison of percent of images correctly predicted by view category, when considering the model's highest-probability top hit (white boxes) vs. its top two hits (blue boxes). **f** Receiver operating characteristic curves for view categories were very similar, with AUCs ranging from 0.985 to 1.00 (mean 0.996). Abbreviations: saxmid short axis mid/mitral, ivc subcostal ivc, subao subcostal/abdominal aorta, supao suprasternal aorta/aortic arch, saxbasal short axis basal, rvinflow right ventricular inflow, a2c apical 2 chamber, a3c apical 3 chamber, a4c apical 4 chamber, a5c apical 5 chamber, psla parasternal long axis, sub4c subcostal 4 chamber

and right ventricular annulus were all considered part of the m-mode view. For each view, we included images with a range of natural echocardiographic variation, such as differences in zoom, depth, focus, sector width, gain, chroma map, systole/diastole, angulation, image quality, and use of 3D, color Doppler, dual mode, strain, and left-ventricular (LV) contrast, to capture the range of variation normally seen by echocardiographers.

A subset of 223,787 images from 15 views were randomly split using Python into training, validation, and test datasets in approximately an 80:10:10 ratio. Each dataset contained images from separate echocardiographic studies, to maintain sample independence. The number of images in training, validation, and test datasets were 180,294, 21,747, and 21,746 images, respectively (corresponding to 213, 27, and 27 different studies in

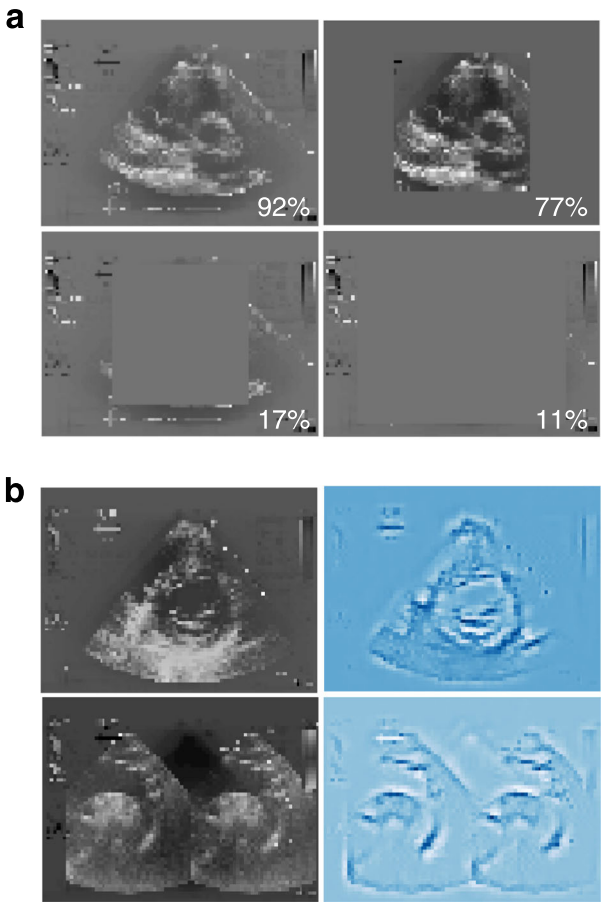

Fig. 6 Visualization of decision-making by neural network. **a** Occlusion experiments. All test images (a short axis basal sample image is shown here) were modified with grey masking of different shapes and sizes as shown, and test accuracy predicted for the test set based on each different modification. Masking that covered cardiac structures resulted in the poorest predictions. **b** Saliency maps. The input pixels weighted most heavily in the neural network's classification decision for two example images (left; suprasternal aorta/aortic arch and short axis mid/mitral input examples shown) were calculated and plotted. The most important pixels (right) make an outline of structures clinically relevant to the view shown

each set). The validation dataset was used for model selection and parameter fine-tuning. The test dataset was used for performance evaluation of the final trained and validated model. For training, 256-shade greyscale pixel values were scaled from [0.255] to [0.1] and the mean over the training data was subtracted from each dataset, as is standard in image-recognition tasks. Also as per standard practice, data were augmented at run-time by randomly applying rotations of up to 10 degrees, width and height shifts of up to a tenth of total length, zooms of up to 0.08, shears of up to 0.03, and vertical/horizontal flips. Training and validation datasets in which view labels were randomized were used as a negative control.

Model architecture and training

Our neural network architecture was designed in Python using the Tensorflow, Theano, and Keras packages, drawing inspiration from the VGG-16 network, which won the Imagenet challenge in 2014.[14–17] Our model utilized a series of small 3×3 convolutional filters connected with max-pooling layers over 2×2 windows. Dropout was utilized in training for both the convolutional and fully connected layers to prevent overfitting. In addition to dropout for regularization, batch normalization was used before neuron activations, which led to faster training and increased accuracy. Activation functions were mainly rectified linear units (ReLU) with the exception of the softmax classifier layer. Training was performed over 45 epochs using an adaptive learning-rate decay for RMSprop optimization. *k*-fold cross-validation ($k = 9$) was used to randomly vary which images were in the training and validation sets, to make use of all available data for training and to select the optimal weights at each epoch. Batches of 64 samples at a time were used for gradient calculation. Convergence plots of training and validation accuracy by epoch confirmed that the model was not overfitting. The training method was robust, with three separate trainings of the 223,787 images resulting in overall test accuracies above 97 percent. Training was performed on Amazon's EC2 platform with a GPU instance g2.2xlarge and took about 18 h. Testing was performed on a laptop computer (Intel i5-3320M CPU @ 2.60GHzx4 with 16 GB RAM); it took a total of 32 s to predict 1500 images, yielding an average of 21 ms per image. Code availability: VGG-16 is publically available on Github.

Model evaluation

Several metrics were used over the test dataset for performance evaluation. Overall accuracy was calculated as the number of correctly classified images as a fraction of the total number of images. Average accuracy was calculated as the average over all views of per-view accuracy. *F*-score was calculated in standard fashion as twice the harmonic mean of precision (positive predictive value) and recall (sensitivity). Receiver operator characteristic (ROC) curves were plotted in the standard way as true-positive fraction (*y*-axis) against false-positive fraction (*x*-axis) and the associated area under curve (AUC) was calculated. Confusion matrices were calculated and plotted as heat maps to visualize performance of multi-view classifiers and their associated errors. Single test images were classified according to the view with the highest probability. Test videos were classified by simple majority vote on multiple images from a given video.

The basis for the model's classification decisions was explored using t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction[18] of raw pixels and of the last fully connected layer output for each sample. Occlusion experiments were performed by masking test images with bounding boxes of different shapes, then submitting them to the model for label prediction. Saliency maps were created using guided backpropogation, which keeps the model weights fixed and computes the gradient of the model's output for a given image.

Comparison to human experts

Echocardiogram test-image classification by board-certified echocardiographers was approved by the UCSF Human Research Protection Program and Institutional Review Board. Each board-certified echocardiographer gave informed consent and was given a randomly selected subset of 1500 60-by-80 pixel images, 100 of each view, drawn from the same low-resolution test set given to the model, and performance compared using the relevant metrics above.

Data availability

The datasets generated during and/or analyzed in this study are available from rima.arnaout@ucsf.edu on reasonable request.

ACKNOWLEDGEMENTS

We thank the board-certified echocardiographers who scored images as human experts. R.A. was supported by NIH/NIAID K08AI114958, AHA 15GPSPG23830004. A. M. and M.M. were supported by AHA 16IRG27630014. R.A. was supported by NIH/NHLBI K08HL125945, AHA 15GPSPG23830004, AHA 17IGMV33870001.

AUTHOR CONTRIBUTIONS

R.A. conceived of the research study with critical input from R.A. and M.M. R.A. labelled data; R.A. and R.A. created and ran the data processing pipeline; A.M., R.A., and R.A. designed and evaluated the deep learning models. R.A., R.A., and A.M. wrote the manuscript with input from M.M.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Digital Medicine* website (https://doi.org/10.1038/s41746-017-0013-1).

Competing interests: The authors declare no competing financial interests.

REFERENCES

1. Karpathy, A. *The Unreasonable Effectiveness of Recurrent Neural Networks*. http://karpathy.github.io/2015/05/21/rnn-effectiveness/ (2015).
2. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
3. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
4. Litjens, G. et al. A survey on deep learning in medical image analysis. eprint at https://arxiv.org/pdf/1702.05747.pdf (2017).
5. Douglas, P. S. et al. ACCF/ASE/AHA/ASNC/HFSA/HRS/SCAI/SCCM/SCCT/SCMR 2011 appropriate use criteria for echocardiography. *J. Am. Coll. Cardiol.* **57**, 1126–1166 (2011).
6. Wharton, G. et al. A minimum dataset for a standard adult transthoracic echocardiogram: a guideline protocol from the British Society of Echocardiography. *Echo Res. Pract.* **2**, G9–G24 (2015).
7. Khamis, H. et al. Automatic apical view classification of echocardiograms using a discriminative learning dictionary. *Med. Image Anal.* **36**, 15–21 (2017).
8. Knackstedt, C. et al. Fully automated versus standard tracking of left ventricular ejection fraction and longitudinal strain: the FAST-EFs multicenter study. *J. Am. Coll. Cardiol.* **66**, 1456–1466 (2015).
9. Narula, S. et al. Machine-learning algorithms to automate morphological and functional assessments in 2D echocardiography. *J. Am. Coll. Cardiol.* **68**, 2287–2295 (2016).
10. Park, J., Zhou, S. K., Simopoulos, C. & Comaniciu, D. AutoGate: fast and automatic Doppler gate localization in B-mode echocardiogram. *Med. Image Comput. Comput. Assist. Interv.* **11**, 230–237 (2008).
11. Sengupta, P. P. et al. Cognitive machine-learning algorithm for cardiac imaging: a pilot study for differentiating constrictive pericarditis from restrictive cardiomyopathy. *Circ. Cardiovasc. Imaging* **9**, https://doi.org/10.1161/CIRCIMAGING.115.004330 (2016).
12. Gao, X. H., Li, W., Loomes, M. & Wang, L. Y. A fused deep learning architecture for viewpoint classification of echocardiography. *Inf. Fusion* **36**, 103–113 (2017).
13. Penatti, O. A. et al. Mid-level image representations for real-time heart view plane classification of echocardiograms. *Comput. Biol. Med.* **66**, 66–81 (2015).
14. Abadi, M. et al. *TensorFlow: Large-scale machine learning on heterogeneous systems* https://www.tensorflow.org/about/bib (2015).
15. Keras (GitHub, 2015).
16. Python Language Reference, version 2.7. Python Software Foundation (2017).
17. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision* **115**, 211–252 (2015).
18. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).