Machine Learning for Pediatric Echocardiographic Mitral Regurgitation Detection

Lindsay A. Edwards MD[a], Fei Feng BS[b], Mehreen Iqbal MD[c], Yong Fu BS[d], Amy Sanyahumbi MD[e], Shiying Hao PhD[f], Doff B. McElhinnney MD[g], X. Bruce Ling PhD[h], Craig Sable MD[i], Jiajia Luo PhD[j]

[a]Dept. of Pediatrics, Seattle Children's Hospital, 4800 Sand Point Way NE, Seattle, WA 98105, United States; lindsay.edwards@seattlechildrens.org
[b]University of Michigan-Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, 800 Dongchuan Rd., Minhang District, Shanghai 200240, China; feifeng@sjtu.edu.cn
[c]Dept. of Pediatrics, Stanford University School of Medicine, 300 Pasteur Drive Suite H310, Stanford, CA 94305, United States; miqbal@stanford.edu
[d]University of Michigan-Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, 800 Dongchuan Rd., Minhang District, Shanghai 200240, China; fuyong_eric@sjtu.edu.cn
[e]Dept. of Pediatrics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, United States; aesims@texaschildrens.org
[f]Dept. of Cardiothoracic Surgery, Stanford University School of Medicine, Falk Bldg, 870 Quarry Rd Extension, Palo Alto, CA 94304, United States; shiyingh@stanford.edu
[g]Dept. of Cardiothoracic Surgery, Stanford University School of Medicine, Falk Bldg, 870 Quarry Rd Extension, Palo Alto, CA 94304, United States; doff@stanford.edu
[h]Dept. of Surgery, Stanford University School of Medicine, Alway Building Ste M121, 300 Pasteur Drive, Stanford, CA 94305, United States; bxling@stanford.edu
[i]Dept. Of Pediatrics, Children's National Health System, 111 Michigan Ave NW, Washington, DC 20010, United States; csable@childrensnational.org
[j]Biomedical Engineering Dept., Peking University, 38 Xueyuan Rd, Haidian, Beijing 100191, China; jiajia.luo@pku.edu.cn

Address for correspondence:

Jiajia Luo, Ph.D.

Biomedical Engineering Department, Peking University, 38 Xueyuan Rd, Haidian, Beijing 100191, P.R. China

phone: +86-10-82805829, fax: +86-10-62010361, email: jiajia.luo@pku.edu.cn

Highlights

1) Many children undergo echocardiogram-based screening for valvular heart disease.
2) We built an AI-based view classification and mitral regurgitation detection model.
3) The model accurately identified view and mitral regurgitation of any severity.
4) With more research, automated pediatric valvular disease detection is feasible.

Background

Echocardiography-based screening for valvular disease in at-risk asymptomatic children can result in early diagnosis. These screening programs, however, are resource intensive, and may not be feasible in many resource-limited settings. Automated echocardiographic diagnosis may enable more widespread echocardiographic screening, early diagnosis, and improved outcomes. In this feasibility study, we sought to build a machine learning model capable of identifying mitral regurgitation (MR) on echocardiogram.

Methods

Echocardiograms were labeled by clip for view and by frame for the presence of MR. The labeled data were used to build two convolutional neural networks (CNNs) to perform the stepwise tasks of classifying the clips 1) by view and 2) by the presence of any MR, including physiologic, in parasternal long axis color Doppler views (PLAX-C). We developed the view classification model using 66,330 frames and evaluated model performance using a hold-out testing dataset with 45 echocardiograms (11,730 frames). We

developed the MR detection model using 938 frames and evaluated model performance using a hold-out testing dataset with 42 echocardiograms (182 frames). Metrics to evaluate model performance included accuracy, precision, recall, F1 score (average of precision and recall, 0 to 1 with 1 suggesting perfect precision and recall), and receiver-operating characteristic analysis.

Results

For the PLAX-C view, the view classification CNN achieved an F1 score of 0.97. The MR detection CNN achieved a testing accuracy of 0.86 and an area under the receiver operating characteristic curve of 0.91.

Conclusions

A machine learning model is capable of discerning MR on transthoracic echocardiography. This is an encouraging step toward machine learning-based diagnosis of valvular heart disease on pediatric echocardiograms.

Key words

Machine learning, deep learning, echocardiogram, mitral valve regurgitation

Abbreviations

AR = aortic valve regurgitation

ARF = acute rheumatic fever

AUC = area under the receiver operating curve

CAM = class activation mapping

CNN = convolutional neural network

ECG = electrocardiogram

MR = mitral valve regurgitation

PLAX-C = parasternal long axis with color Doppler

RHD = rheumatic heart disease

ROC = Receiver operating characteristic

t-SNE = t-distributed stochastic neighbor embedding

WHF = World Heart Federation

Introduction

Serial echocardiography to screen for valvular heart disease is an important component of routine care for several pediatric populations such as children with Marfan Syndrome, those exposed to mediastinal radiation, and those living in regions endemic for rheumatic heart disease (RHD). Early echocardiographic detection of pathologic mitral valve regurgitation (MR) may have important monitoring and/or treatment implications. For example, echocardiogram-based RHD screening is designed to detect latent disease (disease with echocardiographic evidence but no clinical symptoms). Screening asymptomatic children in low- and middle-income countries can result in early latent RHD diagnosis, potentially

providing opportunities for secondary prevention initiation and reduced morbidity and mortality.[1-5] The 2012 World Heart Federation (WHF) criteria are the gold standard for echocardiographic RHD diagnosis,[6] but screening with these criteria is not always feasible due to equipment costs and the complexity of the criteria for nonexpert scanners. Simplified single-view screening protocols with handheld ultrasound reduce cost, can be performed reliably by nonexpert scanners, and result in good sensitivity and specificity for RHD detection.[7-11]

Task-shifting of echocardiography to non-physician health care workers has been shown to increase access to screening,[7, 12-18] but still relies on significant expertise for interpretation. Automated diagnosis may obviate the need for a local physician, enabling more widespread screening. Several groups have demonstrated the feasibility of a cognitive machine-learning approach for pattern recognition in echocardiographic evaluation.[19-24] Martins et al published their preliminary experience with machine learning for echocardiographic RHD diagnosis; their 3D convolutional neural network achieved a diagnostic accuracy of 72.77%,[25] suggesting that while the task is feasible, further research is needed to improve accuracy of predictions.

The aim of the current study was to determine the feasibility of developing a fully automated, sequential machine learning model capable of view classification and MR detection (any MR severity). We hypothesize that a CNN will be capable of recognizing echocardiographic MR.

This manuscript describes our proof-of-concept model, the first step in developing a machine learning system for detection of pediatric mitral valve disease.

Materials and methods

Human Subjects Research

A waiver of approval was obtained from the Stanford University and Seattle Children's Hospital Institutional Review Boards. This study was approved by the Malawi National Health Sciences Research Committee.

Echocardiograms

All echocardiograms were performed by a pediatric cardiologist (AS) or a trained ultrasonographer in Malawi, Africa, as part of a screening program to detect latent RHD.[26] Asymptomatic children aged five to 16 years-old underwent school or community-based screening from 2014 to 2015. All echocardiographic studies were obtained using a Philips CX50 (Best, Holland) portable echocardiography machine with an S5-1 transducer probe. The studies adhered to an abbreviated RHD screening protocol consisting of parasternal long-axis, apical four-chamber, and apical five-chamber views with and without color Doppler.[6] If a concern for RHD emerged during the study, additional parasternal long axis, apical four-chamber, parasternal short-axis, and spectral Doppler clips were often obtained. Machine settings were consistent with WHF recommendations for echocardiographic diagnosis of RHD.[6] Echocardiograms that identified congenital heart disease were excluded. A total of

2,229 video clips and 66,330 still frames from 227 unique subject's echocardiograms were analyzed for model development.

Pipeline for the Automated Diagnosis of RHD

Our approach had two steps: 1) view classification, to identify parasternal long axis with color Doppler (PLAX-C) views from inputted 10-frame samples, and 2) MR detection, assessing for the presence of any MR in PLAX-C views (Figure 1) from inputted systolic frames. Two CNN models were developed, inspired by DenseNet[27] and ResNet[28], with hyperparameter tuning.

Echocardiographic Data Preprocessing and Labeling

We used the Python (v.2.7) and GDCM library (v.2.8.0) to extract metadata from the compressed DICOM echocardiogram data. Each clip was labeled by an investigator (LAE, MI) for view. Echocardiogram clips were divided into 10-frame samples (frames 1-10, 11-20, etc.) for the view classification CNN (10, 600, 800). Frame 1 standardly begins at the R wave on ECG tracing, corresponding to isovolumic contraction, but the first frame of subsequent samples did not correspond to a standard point in the cardiac cycle. All extraneous information (heart rate, scale, etc.) was stripped from the data, such that each frame included only the cropped 2-D, color Doppler, and/or spectral Doppler components. The final input size for the view classification CNN model was (10, 486, 486). The majority of echocardiograms included parasternal long axis, apical four-chamber, and apical five-

chamber views. There were many view iterations, depending on which structures the operator focused on, rotation or tilt of the probe, and presence or absence of color and spectral Doppler. We identified 10 unique view variations and divided the 10-frame samples into training, validation, and testing datasets by echocardiogram to ensure that samples from the same clip were not distributed amongst the groups.[29] The data distribution for presence of any MR among different views is summarized in Table 1 (more detail can be found in Supplemental Table 1). The training, validation, and testing datasets were divided by unique subject, with a ratio of 0.6:0.2:0.2 (145:48:48 subjects). As particular echocardiographic views were only obtained if RHD suspicion existed at the time, there was an imbalance of views. To overcome the difficulty of imbalanced data distribution in the training dataset, we utilized data augmentation techniques in the view categories with limited samples (color Doppler apical four and five chamber views and 2D and color Doppler parasternal short axis views) including image rotation, translation, and shear.[30] For data augmentation, we used the Keras (v.2.0.9) *DataGenerator* to resample the views mentioned above 5, 4, 3, 2, 20, and 1 time separately. Data augmentation techniques were not used in the validation and testing datasets.

MR presence can be detected only in frames captured in certain echocardiographic views with color Doppler and during systole in the cardiac cycle. For this feasibility study, the MR detection model used inputted data comprised of single systolic frames in the PLAX-C view, labeled individually for the presence of any MR, including physiologic. Systolic frames were

used for all labeled data (MR/No MR) to ensure otherwise similar data features. Therefore, only systolic frames were input into the MR detection model for training and testing. MR+ frames were individually identified and labeled throughout the PLAX-C clips. Amongst MR+ frames, we compared the first 10 frames of the clip to identify frames in which MR was most frequently detected. MR was most frequently identified in frame 4 followed by frame 5, thus these two frames were selected from MR- PLAX-C clips for the MR- input. The image was cropped to an input size was (300,300), focusing on the mitral valve region for MR detection. To ensure equal distribution of MR severity within the dataset, we evaluated each echocardiogram for the presence of MR in two views, a WHO-defined marker of more significant MR. We divided the frames into training, validation, and testing datasets by the presence of MR in two views (Supplemental Table 2).

Evaluation of the Convolutional Neural Network Models

To evaluate the performance of the two CNN models, several standard metrics were used as described below.[31]

Accuracy, precision, recall, and F1 score

Accuracy is the fraction of true predictions [accuracy= true predictions/ total predictions]. Precision, or positive predictive value, is the probability that a positive prediction is correct [precision=true positive predictions/ total positive predictions]. Recall, or sensitivity, is the probability that any given input generates the correct positive prediction [recall= true positive

predictions/(true positive + false negative predictions)]. F1 score, also known as F1-measure compute, is the harmonic average of precision and recall [F1 score= 2*(recall * precision)/ (recall + precision)]. F1 scores range from 0 to 1, with an F1 score of 1 suggesting perfect precision and recall of the model. While precision and recall can be manipulated by changing parameters in the model, hence introducing bias, the F1 score is generally regarded as a more impartial metric. Comparison of F1 scores across models is problematic (i.e. there is no standard cutoff for a "good" F1 score).

Confusion matrix

A confusion matrix allows for visualization of model performance by comparing actual (true label) versus predicted class.

Receiver operating characteristic (ROC) curve

ROC curves were constructed by plotting the true positive rate against the false positive rate. The area under the ROC curve (AUC) was used to measure the robustness of the model.

t-distributed stochastic neighbor embedding (t-SNE)

t-SNE is an unsupervised data dimension reduction method used to visualize high dimensional data in a two- or three-dimensional coordinate system.[32]

Class activation mapping (CAM)

The CAM technique activates the regions of the input data that most importantly contributed to the model's final prediction.[33]

Further methodology details are provided in the supplemental material.

Results

View Classification CNN Model

Figure 2a depicts the training curve for the view classification CNN. The model converged and achieved a high level of accuracy. We computed the final test accuracy for the view classification model as 0.98. The view classification CNN has an average testing F1 score of 0.97 across the ten views. The gap between the training and validation curves in the two plots reflects generalization error. For the PLAX-C view, the F1 score reached 0.97, suggesting near-perfect prediction (Table 2). With the exception of parasternal short axis color Doppler views, which suffered from inadequate training data, the F1 scores of all other views were above 0.96, suggesting excellent overall performance of the model.

t-SNE visualization (Figure 3a) depicts the view classification CNN's grouping of data. Overall, the model was able to separate the views well, excepting the parasternal short axis color Doppler views, which the model struggled to differentiate from PLAX-C. Figure 3b depicts the confusion matrix for view classification model performance.

After the model was trained, we used the class activation mapping (CAM) technique to visualize model learning (Figure 4).[33] The strongest activation signal indicates the part of the image relied on most heavily for the prediction output. The view classification CNN focuses on structural differences between imaging views.

MR Detection CNN Model

The MR detection CNN model promisingly differentiated any MR from No MR in PLAX-C views. Figure 2b depicts the training curve for the MR detection CNN. This model also converged and achieved a high level of testing accuracy at 0.86 on 182 testing samples, with the F1 score for MR equaling 0.90 and for No MR equaling 0.77 (Table 3).

The MR detection CNN's performance during the training process is depicted in the ROC in Figure 5. Our model achieved an AUC of 0.91, indicating a strong prediction ability. The CAM technique was also utilized to understand the MR detection model's learning process (Figure 6). The MR detection CNN relies heavily on the color Doppler signal.

Discussion

We present a machine learning-based automated pipeline for assessment of MR in a single echocardiographic view. This work represents the first step in the development of a model for diagnosis of mitral valve disease in a pediatric population. While the diagnosis of any MR, including physiologic, is not clinically useful, our model suggests the feasibility of a two-step

workflow for machine learning in the echocardiographic diagnosis of pediatric mitral valve disease. The models were developed using data from pediatric screening echocardiograms for RHD, but the proposed approach may translate well to other valvopathies, such as in children with Marfan's Syndrome or post mediastinal radiation.

The system employs two CNNs in series, the view classification CNN to identify PLAX-C views and a binary MR detection CNN in the PLAX-C view. The deep learning method is traditionally used as a black box; however, we favored two CNN models over a single-step end-to-end learning method for ease of conceptualization by clinicians and to benefit from cumulative model expertise. Both models were tested on individual hold-out testing datasets. For the PLAX-C view, the view classification CNN achieved near-perfect prediction (F1 score 0.97). The MR detection CNN achieved a testing accuracy of 0.86 and an AUC of 0.91. Our dataset consisted entirely of RHD screening echocardiograms obtained in the field, suggesting a high level of accuracy despite image variability due to differences in echocardiographic windows amongst subjects and technical variability amongst scanners.

Several groups have examined the potential of machine learning for echocardiographic view classification. Gao et al. found a fused CNN using both spatial and temporal echocardiographic information outperformed several well-known hand-crafted algorithms in differentiating echocardiographic views.[24] Our model achieved better view classification performance using a more efficient network and investigated color Doppler views. Khamis et

al. introduced an algorithm using Cuboid detection and supervised dictionary learning to discriminate amongst apical echocardiographic views.[20] This method focused on recognition of the mitral valve region in three similar apical views; our method included many more views, both 2D and color Doppler, and did not focus on a single anatomic structure, allowing for more broad views classification. Nascimento et al. performed automated view classification on a set of RHD screening echocardiograms collected through the PROVAR study in Brazil.[34] Our model achieved better view classification performance for more views. Gearhart et al used pediatric and young adult echocardiogram images to build a CNN to identify 27 different views with overall model accuracy of 90.3%.[35] While our test accuracy was superior, the comprehensiveness of our CNN was intentionally more limited given our goal of a multistep process for MR detection; we included fewer views, did not include spectral Doppler views, and did not include children with structurally abnormal hearts. A few groups have forayed into multistep models with view classification and automated pathology diagnosis. Madani et al. built a CNN capable of view classification with view segmentation and classification of left ventricular hypertrophy.[21, 22] Zhang et al.'s multistep model includes view classification, image segmentation, quantification of ventricular size and function, and disease classification of hypertrophic cardiomyopathy, pulmonary arterial hypertension, and amyloidosis.[23] Kusunose et al. also investigated using CNN method to achieved five-view echocardiographic data classification.[36] Our work differs in that color Doppler views were added, laying the foundation for automated MR diagnosis.

In recent years, some progress has been made using machine learning for echocardiographic assessment of MR severity [37]. Moghaddasi and Nourian used textural descriptors from echocardiographic images obtained at three ECG-determined points in the cardiac cycle to classify MR severity.[38] Zhang et al also developed a CNN to grade MR using apical 4-chamber color Doppler images with traced MR contours.[39] Our work differs from these studies in several important ways. Our model was built with pediatric data and focuses on MR identification rather than MR grading. Additionally, our machine learning pipeline includes comprehensive view classification, such that our algorithm will sort through a complete echocardiogram and select appropriate views before moving on to the MR detection step.

At this time, the MR detection CNN is limited in that it has only been trained in the PLAX-C view and does not differentiate MR severity. The presence of any MR is not useful, as trivial to mild MR is a nonspecific finding and can be present in healthy children. While this study was designed to test the feasibility of machine learning for echocardiographic MR detection, our next step is to train the CNN to detect pathologic MR. Pathologic MR is usually due to RHD in endemic populations and is the most common manifestation of RHD in children.[6] Simplified RHD screening criteria consist of the length of the MR jet (>2cm[7] or >1.5 cm[8]) and the presence of any AR. As we continue this project, we plan to detect a binary outcome of pathologic MR or no/non-pathologic MR, experimenting with MR jet length ≥ 1.5 cm[8,11] and ≥ 2 cm.[6] Zulkhe et al.[10] and Diamantino et al.[11] reported that screening protocols limited

to PLAX-C views and one measurement, MR jet length, are sensitive and specific for RHD diagnosis. The WHF differentiates pathologic MR from normal if seen in two views, so we will trial single versus two-view protocols as we continue to build our model.

Our methodology could be replicated to detect other echocardiographic features of RHD, such as AR, which has been used in addition to pathologic MR in simplified screening protocols,[7-9] or morphological valve features. Nunes et al.[40] proposed the first risk stratification score for RHD disease progression, which includes morphologic features of the mitral and aortic valve for RHD identification and prediction of disease progression. While incorporating morphologic features would increase model specificity, developing a training data set with morphological components is not feasible at present, as the individual morphological components lack sufficient inter-reviewer agreement to inform machine learning assessment of morphology; this may necessitate modification of guidelines for development of machine learning-based RHD algorithms that include 2D morphology. Rather than focus on specific echocardiographic features of RHD, Martins et al used 2D and color Doppler clips in numerous views from RHD+ (borderline and definite) and RHD- children to build their RHD diagnosis model and found that a 3D CNN with aggregated multiple clip input and random forest meta-classifier outperformed a 2D frame-based CNN, presumably due to the added temporal data.[25] However, 2D CNN has been successfully utilized for MR classification as well as other cardiac disease processes.[21,23,38]

Our approach has several additional limitations. The MR detection CNN shows higher generalization error than the view classification model, likely attributable to: 1) smaller training dataset size; 2) separability of features within the same view is more challenging than amongst different views; and 3) we did not include the effect of the error accumulation on precision, recall, or F1 score. Training data size is usually an important factor in the machine learning application and was limited in both of our CNNs. As parasternal short axis and spectral Doppler clips were not obtained in the RHD screening echocardiograms unless there was an RHD concern, fewer of these views are represented in our training, test, and validation sets, and those present introduce selection bias. Our model's performance in recognizing these views is predictably lower. As we are able to obtain and label more data, we anticipate improvement in model performance. Alternatively, there are some advanced techniques, such as a semi-supervised learning method, which could be explored. Technically, MR detection is a more fine-grained classification task and proved more challenging for the CNN than view classification. In addition, the view classification and MR detection models are launched in serial to reach the diagnosis, and, thus, the view classification model's error could reduce the MR detection model's performance.

An additional limitation to both models is that all echocardiograms were obtained in Malawi, where, anecdotally, most children have good echocardiographic windows. Studies were performed by a cardiologist and a sonographer, and it is unclear how our model would perform with echocardiograms performed by additional and potentially less skilled operators.

Furthermore, we do not have demographic or clinical information (e.g. age, body mass index, etc.) to understand if our model performed better in any subset of the Malawi population. In populations where obesity is prevalent, echocardiographic windows may be poor, and it is unclear how our model will generalize to these populations. Final limitations center around echocardiogram acquisition: 1) all clips were obtained with a Philips CX50 portable ultrasound, so it is unclear how our model will perform with other portable and handheld models, and 2) the scanner was actively interpreting the echocardiogram and obtained additional, potentially optimized, clips if there was an RHD concern, possibly skewing the quality of our positive data and potentially limiting the CNN's ability to detect MR on suboptimal clips.

Conclusions

We successfully built a machine learning model capable of view classification and MR detection. The early success of our approach suggests that automated MR detection for applications such as RHD screening is feasible. This study further demonstrates the potential of machine learning-based models in echocardiographic diagnosis of cardiac disease.

Availability of data and materials: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations of interest: none

Funding: F.F.'s doctoral studies are partially funded by NSFC General Program grant 31870942. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions: All authors participated in project conception and design. LE and MI labeled the data, and FF, YF, and JL developed and tested the model. LE, FF, and JL drafted the manuscript and all authors critically revised it. All authors approved the final version of this manuscript.

Acknowledgments: None

Table 1. Data Distribution for MR Presence and Different Views

View	Clip Distribution		
	MR†	No MR†	Total
Total	309	1920	2229
PLAX	0	429	429
PLAX-C	137	356	493
PSAX-AV	0	40	40
PSAX-AV-C	0	37	37
PSAX-MV	0	66	66
PSAX-MV-C	28	13	41
A4C	0	319	319
A4C-C	144	154	298
A5C	0	261	261
A5C-C	0	245	245

†MR and No MR headings in this table reflect the overall presence or absence of any MR in the clip. *MR: mitral regurgitation; PLAX: parasternal long axis; PLAX-C: parasternal long axis with color Doppler; PSAX-AV: parasternal short axis at the level of the aortic valve; PSAX-AV-C: parasternal short axis at the level of the aortic valve with color Doppler; PSAX-MV: parasternal short axis at the level of the mitral valve; PSAX-MV-C: parasternal short axis at the level of the mitral valve with color Doppler; A4C: apical four-chamber view; A4C-C: apical four-chamber view with color Doppler box over the mitral valve; A5C:*

apical five-chamber view (i.e. apical four-chamber view with anterior angulation); A5C-C: apical five-chamber view with color

Table 2. Performance of the View Classification CNN on Testing Data

View	Precision	Recall	F1-score	Testing sample
PLAX	1.00	1.00	1.00	200
PLAX-C	0.98	0.97	0.97	200
PSAX-AV	0.96	0.96	0.96	28
PSAX-AV-C	0.72	0.95	0.82	22
PSAX-MV	0.94	0.98	0.96	63
PSAX-MV-C	1.00	0.71	0.83	21
A4C	0.97	0.99	0.98	200
A4C-C	0.96	0.98	0.97	125
A5C	0.99	0.96	0.98	200
A5C-C	0.97	0.96	0.96	114

CNN: convolutional neural network; PLAX: parasternal long axis; PLAX-C: parasternal long axis with color Doppler; PSAX-AV: parasternal short axis at the level of the aortic valve; PSAX-AV-C: parasternal short axis at the level of the aortic valve with color Doppler; PSAX-MV: parasternal short axis at the level of the mitral valve; PSAX-MV-C: parasternal short axis at the level of the mitral valve with color Doppler; A4C: apical four-chamber view; A4C-C: apical four-chamber view with color Doppler box over the mitral valve; A5C: apical five-chamber view (i.e. apical four-chamber view with anterior angulation); A5C-C: apical five-chamber view with color Doppler box over the aortic valve.

Table 3. Performance of MR Detection CNN model on testing data

Class	Precision	Recall	F1-score	Testing sample
MR	0.93	0.86	0.90	130
No-MR	0.71	0.85	0.77	52

MR: any mitral regurgitation

Figure Legends

Figure 1. System construction pipeline. The system construction pipeline consisted of two separate convolutional neural networks for two tasks, view classification and mitral valve regurgitation (MR) detection.

Figure 2. Training and validation curves of the convolutional neural networks. The training and validation curves for the view classification model (*a*) converge at approximately 50 epochs (the number of passes of the training dataset the model has completed). Fluctuation is noted in the validation curve, likely reflecting variation within a single view amongst studies; this improves with increasing epochs, suggesting the model learns generalizable knowledge to inform its prediction. The final F1 score for the view classification model approached 1, suggesting near perfect precision and recall of the model. The test accuracy for the mitral valve regurgitation detection model (*b*) was also strong. The training curve and validation curves for this model converge after approximately 200 epochs. The mitral valve regurgitation detection model lacks sufficient training at this time to overcome generalization error and narrow the gap between training and validation curves and reduce fluctuation in the validation curve.

Figure 3. Visualization of the successful prediction of the view classification model. t-distributed stochastic neighbor embedding visualization (*a*) and confusion matrix (*b*). In the t-SNE plot (*a*), each point represents the data obtained from a 10-frame sample, and the distance between points reflects how similar the model found the data. The point's color reflects the true

label, i.e. the true echocardiographic view. In the confusion matrix (*b*), accuracy is represented by the color scale displayed at the right of the matrix. A diagonal line from top left to bottom right represents the correct prediction. The off-diagonal cells reflect wrong predictions; for example, A5C-C was erroneously predicted to be A4C-C on five occasions. Overall, both methods suggest the model was able to separate echocardiographic views well, the exception being parasternal short axis color Doppler views, which the model struggled to differentiate from parasternal long axis color Doppler views. *Parasternal long axis with (PLAX-C) and without (PLAX) color Doppler; parasternal short axis at the level of the aortic valve with (PSAX-AV-C) and without (PSAX-AV) color Doppler and at the level of the mitral valve with (PSAX-MV-C) and without (PSAX-MV) color Doppler; apical four-chamber view without color Doppler (A4C) and with color Doppler over the mitral valve (A4C-C); and apical five-chamber view without color Doppler (A5C) and with color Doppler over the aortic valve (A5C-C).*

Figure 4. Class activation mapping technique for the view classification model. The brightest activation signal indicates the part of the image relied on most heavily for the prediction output. For example, the model is relying on mitral inflow in the PLAX-C frame and on the presence of the 2D aortic root in the A5C and A5C-C frames; the A5C-C frame is in diastole, and the model is not relying on color aortic outflow in this particular frame. In the depicted PLAX frame, the model is focusing on something in the right ventricle. *Parasternal long axis with (PLAX-C) and without (PLAX) color Doppler; parasternal short axis at the level of the aortic valve with (PSAX-AV-C) and without (PSAX-AV) color Doppler and at the level*

of the mitral valve with (PSAX-MV-C) and without (PSAX-MV) color Doppler; apical four-chamber view without color Doppler (A4C) and with color Doppler over the mitral valve (A4C-C); and apical five-chamber view without color Doppler (A5C) and with color Doppler over the aortic valve (A5C-C).

Figure 5. Receiver operating characteristic curve for the blind testing of the mitral regurgitation detection model. The dashed line demonstrates chance. As the false positive rate approaches 0, the true positive rate nears 0.5, suggesting a high positive predictive value. The model achieved an area under the receiver operating curve of 0.91, also indicating a strong prediction ability.

Figure 6. Class activation mapping technique for the mitral regurgitation model. The brightest activation signals are the mitral regurgitation and aortic outflow color Doppler signals, indicating the model focused most heavily on these features to make its prediction.

References

1. Carapetis JR, Steer AC, Mulholland EK, Weber M. The global burden of group A streptococcal diseases. Lancet Infect Dis. 2005;5:685-94.
2. Bhaya M, Panwar S, Beniwal R, Panwar RB. High Prevalence of Rheumatic Heart Disease Detected by Echocardiography in School Children. Echocardiography. 2010;27:448-53.
3. Paar JA, Berrios NM, Rose JD, Cáceres M, Peña R, Pérez W, et al. Prevalence of Rheumatic Heart Disease in Children and Young Adults in Nicaragua. Am J Cardiol. 2010;105:1809-14.
4. Carapetis JR, Hardy M, Fakakovikaetau T, Taib R, Wilkinson L, Penny DJ, et al. Evaluation of a screening protocol using auscultation and portable echocardiography to detect asymptomatic rheumatic heart disease in Tongan schoolchildren. Nat Clin Pract Cardiovasc Med. 2008;5:411-7.
5. Marijon E, Ou P, Celermajer DS, Ferreira B, Mocumbi AO, Jani D, et al. Prevalence of Rheumatic Heart Disease Detected by Echocardiographic Screening. N Engl J Med. 2007;357:470-6.
6. Reményi B, Wilson N, Steer A, Ferreira B, Kado J, Kumar K, et al. World Heart Federation criteria for echocardiographic diagnosis of rheumatic heart disease—an evidence-based guideline. Nat Rev Cardiol. 2012;9:297-309.
7. Mirabel M, Bacquelin R, Tafflet M, Robillard C, Huon B, Corsenac P, et al. Screening for Rheumatic Heart Disease: evaluation of a focused cardiac ultrasound approach. Circ Cardiovasc Imaging. 2015;8:e002324.
8. Lu JC, Sable C, Ensing GJ, Webb C, Scheel J, Aliku T, et al. Simplified Rheumatic Heart Disease Screening Criteria for Handheld Echocardiography. J Am Soc Echocardiogr. 2015;28:463-9.
9. Godown J, Lu JC, Beaton A, Sable C, Mirembe G, Sanya R, et al. Handheld Echocardiography Versus Auscultation for Detection of Rheumatic Heart Disease. Pediatrics. 2015;135:e939-44.
10. Zühlke LJ, Engel ME, Nkepu S, Mayosi BM. Evaluation of a focussed protocol for hand-held echocardiography and computer-assisted auscultation in detecting latent rheumatic heart disease in scholars. Cardiol Young. 2016;26:1097-106.
11. Diamantino A, Beaton A, Aliku T, Oliveira K, Oliveira C, Xavier L, et al. A focussed single-view hand-held echocardiography protocol for the detection of rheumatic heart disease. Cardiol Young. 2018;28:108-17.
12. Sims Sanyahumbi A, Sable CA, Karlsten M, Hosseinipour MC, Kazembe PN, Minard CG, et al. Task shifting to clinical officer-led echocardiography screening for detecting rheumatic heart disease in Malawi, Africa. Cardiol Young. 2017;27:1133-9.
13. Colquhoun SM, Carapetis JR, Kado JH, Reeves BM, Remenyi B, May W, et al. Pilot study of nurse-led rheumatic heart disease echocardiography screening in Fiji – a novel approach in a resource-poor setting. Cardiol Young. 2013;23:546-52.
14. Ploutz M, Lu JC, Scheel J, Webb C, Ensing GJ, Aliku T, et al. Handheld

echocardiographic screening for rheumatic heart disease by non-experts. Heart. 2016;102:35-9.
15. Engelman D, Kado JH, Reményi B, Colquhoun SM, Watson C, Rayasidamu SC, et al. Teaching focused echocardiography for rheumatic heart disease screening. Ann Pediatr Cardiol. 2015;8:118-21.
16. Engelman D, Kado JH, Reményi B, Colquhoun SM, Carapetis JR, Donath S, et al. Focused cardiac ultrasound screening for rheumatic heart disease by briefly trained health workers: a study of diagnostic accuracy. Lancet Glob Health. 2016;4:e386-94.
17. Engelman D, Kado JH, Reményi B, Colquhoun SM, Carapetis JR, Wilson NJ, et al. Screening for rheumatic heart disease: quality and agreement of focused cardiac ultrasound by briefly trained health workers. BMC Cardiovasc Disord. 2016;16:30.
18. Beaton A, Nascimento BR, Diamantino AC, Pereira GTR, Lopes ELV, Miri CO, et al. Efficacy of a Standardized Computer-Based Training Curriculum to Teach Echocardiographic Identification of Rheumatic Heart Disease to Nonexpert Users. Am J Cardiol. 2016;117:1783-9.
19. Sengupta PP, Huang YM, Bansal M, Ashrafi A, Fisher M, Shameer K, et al. Cognitive Machine-Learning Algorithm for Cardiac Imaging: A Pilot Study for Differentiating Constrictive Pericarditis From Restrictive Cardiomyopathy. Circ Cardiovasc Imaging. 2016;9:e004330.
20. Khamis H, Zurakhov G, Azar V, Raz A, Friedman Z, Adam D. Automatic apical view classification of echocardiograms using a discriminative learning dictionary. Med Image Anal. 2017;36:15-21.
21. Madani A, Ong JR, Tibrewal A, Mofrad MRK. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. NPJ Digit Med. 2018;1:59.
22. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. NPJ Digit Med. 2018;1:6.
23. Zhang J, Gajjala S, Agrawal P, Tison Geoffrey H, Hallock Laura A, Beussink-Nelson L, et al. Fully Automated Echocardiogram Interpretation in Clinical Practice. Circulation. 2018;138:1623-35.
24. Gao X, Li W, Loomes M, Wang L. A fused deep learning architecture for viewpoint classification of echocardiography. Inform Fusion. 2017;36:103-13.
25. Martins JFBS, Nascimento ER, Nascimento BR, Sable CA, Beaton AZ, Ribeiro AL, et al. Towards automatic diagnosis of rheumatic heart disease on echocardiographic exams through video-based deep learning. J Am Med Inform Assoc. 2021;28:1834-42.
26. Sims Sanyahumbi A, Sable CA, Beaton A, Chimalizeni Y, Guffey D, Hosseinipour M, et al. School and Community Screening Shows Malawi, Africa, to Have a High Prevalence of Latent Rheumatic Heart Disease. Congenit Heart Dis. 2016;11:615-21.
27. Huang G, Liu Z, Maaten Lvd, Weinberger KQ. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2017;2261-69.
28. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE

Conference on Computer Vision and Pattern Recognition(CVPR). 2016;770-8.
29. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436-44.
30. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM. 2017;60:84-90.
31. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol. 2011;2:37-63.
32. Der Maaten LV, Hinton GE. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9:2579-605.
33. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). 2016;2921-9.
34. Nascimento BR, Meirelles ALS, Meira W, Pappa GL, do Carmo Nunes M, Sable C, et al. Computer Deep Learning for Automatic Identification of Echocardiographic Disease Screening: Data From The Atmosphere-Provar Study. J Am Coll Cardiol. 2019;73:1611.
35. Gearhart A, Goto S, Deo R C, et al. An Automated View Classification Model for Pediatric Echocardiography Using Artificial Intelligence. J Am Soc Echocardiogr. 2022. doi: https://doi.org/10.1016/j.echo.2022.08.009.
36. Kusunose K, Haga A, Inoue M, Fukuda D, Yamada H, Sata M. Clinically Feasible and Accurate View Classification of Echocardiographic Images Using Deep Learning. Biomolecules. 2020;10:665.
37. Nedadur R, Wang B, Tsang W. Artificial intelligence for the echocardiographic assessment of valvular heart disease. Heart. 2022;0:1–8.
38. Moghaddasi H, Nourian S. Automatic assessment of mitral regurgitation severity based on extensive textural features on 2D echocardiography videos. Comput Biol Med. 2016;73:47-55.
39. Zhang Q, Liu Y, Mi J, Wang X, Liu X, Zhao F, Xie C, Cui P, Zhang Q, Zhu X. Automatic Assessment of Mitral Regurgitation Severity Using the Mask R-CNN Algorithm with Color Doppler Echocardiography Images. Comput Math Methods Med. 2021;2021:2602688.
40. Nunes MCP, Sable C, Nascimento BR, Lima EMD, da Silva Jose LP, Diamantino AC, et al. Simplified Echocardiography Screening Criteria for Diagnosing and Predicting Progression of Latent Rheumatic Heart Disease. Circ Cardiovasc Imaging. 2019;12: e007928.