JMI Journal of Medical Imaging

RESEARCH PAPER

Challenges and solutions of echocardiography generalization for deep learning: a study in patients with constrictive pericarditis

Jiwoong Jeong,[a,*] Chieh-Ju Chao,[b] Reza Arsanjani,[c] Kihong Kim,[b] Melissa N. Pelkey,[b] Yi-Chieh Chen,[d] Raheel N. Ramzan,[c] Mohammad Elbahnasawy,[c] Mohamed Sleem,[c] Chadi Ayoub,[c] Juan Maria M. Farina,[c] Martha Grogan,[b] Garvan C. Kane,[b] Bhavik N. Patel,[e] Jae K. Oh,[b] and Imon Banerjee[a]

[a]Arizona State University, School of Computing and Augmented Intelligence, Tempe, Arizona, United States
[b]Mayo Clinic, Department of Cardiology, Rochester, Minnesota, United States
[c]Mayo Clinic, Department of Cardiology, Scottsdale, Arizona, United States
[d]Mayo Clinic Health System Austin, Department of Pharmacy, Austin, Minnesota, United States
[e]Mayo Clinic, Department of Radiology, Scottsdale, Arizona, United States

ABSTRACT. **Purpose:** The inherent characteristics of transthoracic echocardiography (TTE) images such as low signal-to-noise ratio and acquisition variations can limit the direct use of TTE images in the development and generalization of deep learning models. As such, we propose an innovative automated framework to address the common challenges in the process of echocardiography deep learning model generalization on the challenging task of constrictive pericarditis (CP) and cardiac amyloidosis (CA) differentiation.

Approach: Patients with a confirmed diagnosis of CP or CA and normal cases from Mayo Clinic Rochester and Arizona were identified to extract baseline demographics and the apical 4 chamber view from TTE studies. We proposed an innovative preprocessing and image generalization framework to process the images for training the ResNet50, ResNeXt101, and EfficientNetB2 models. Ablation studies were conducted to justify the effect of each proposed processing step in the final classification performance.

Results: The models were initially trained and validated on 720 unique TTE studies from Mayo Rochester and further validated on 225 studies from Mayo Arizona. With our proposed generalization framework, EfficientNetB2 generalized the best with an average area under the curve (AUC) of 0.96 ($\pm$0.01) and 0.83 ($\pm$0.03) on the Rochester and Arizona test sets, respectively.

Conclusions: Leveraging the proposed generalization techniques, we successfully developed an echocardiography-based deep learning model that can accurately differentiate CP from CA and normal cases and applied the model to images from two sites. The proposed framework can be further extended for the development of echocardiography-based deep learning models.

 [DOI: 10.1117/1.JMI.10.5.054502]

Keywords: deep learning; transthoracic echocardiography; restrictive cardiomyopathy

Paper 23055GRR received Feb. 28, 2023; revised Sep. 11, 2023; accepted Sep. 19, 2023; published Oct. 12, 2023.

*Address all correspondence to Jiwoong Jeong, jjeong35@asu.edu

2329-4302/2023/$28.00 © 2023 SPIE

1 Introduction

Transthoracic echocardiography (TTE) is one of the most widely available imaging modalities in clinical cardiology, and it has been used as the first-line screening tool in various cardiac conditions.[1–5] In addition to its availability, TTE has the advantages of having a high temporal resolution and being radiation free. Despite its importance in TTE echo studies for clinical phenotyping, there is also significant variance in the human interpretation of echocardiogram images that could impact diagnosis and clinical care. Formalized training guidelines for cardiologists recognize the value of experience in interpreting echocardiogram images, and basic cardiology training might be insufficient for interpreting echocardiograms at the highest level. The reading of the studies is difficult based on not only the reader variations but also the positioning of the probe, length of the acquisition, breathing cycle, and variation in image reconstructions by different devices.

As an ultrasound-based technology, echocardiography images come with intrinsic limitations such as a smaller field of view and low signal-to-noise ratio.[6] Furthermore, the background noise and intensity distribution can vary between studies, which can be changed with the ultrasound machine vendors as well as the settings at the time of scanning. The above conditions can also significantly limit the generalization of echocardiography-based artificial intelligence (AI) models, especially when using cross-institution data for external validation. However, most of the recent echocardiography-based AI studies emphasized model performance,[7–9] and detailed step-by-step preprocessing or generalization instructions for addressing the common obstacles were not available.

Previous works[10] on developing more generalizable deep learning models on echocardiograms have used traditional data augmentation techniques, which include geometric transformations such as flipping and rotation as well as adjusting the contrast and saturation of the image. For example, Madani et al.[11] applied standard data augmentation during training with up to 10 deg rotation and 10% height and weight shifts and obtained 91.2% accuracy for binary left ventricular hypertrophy classification on the on hold-out internal dataset. Yu et al.[12] also applied traditional image augmentation to the echocardiograms, including random shifts of contrast, brightness, or saturation, with or without horizontal flips, for classifying left ventricular hypertrophy. But even though these models showed encouraging performance on the internal datasets, the generalization capability was never validated on an external dataset from another center. Silva et al.[13] developed a 3D convolutional neural network (CNN) model for classification of ejection fraction, but no particular data augmentation scheme was mentioned in the study, and no generalization was shown for other center data. Østvik et al.[14] applied more realistic augmentation designed for ultrasound (US), e.g., Gaussian shadowing, haze artifacts, depth attenuation and speckle reduction, and demonstrated good motion estimation generalization for different echo vendors.

In this study, we choose the challenging constrictive pericarditis (CP) patients as the study population because these echocardiography studies contain most of the common issues that one could encounter in echocardiography-based AI studies. A dilemma in training models for CP is how to preserve the characteristic septal bounce feature, an abnormal motion of the interventricular septum termed interventricular septal shift that is suggested to be one the three most useful clinical criteria for diagnosis of CP,[15] in spatio-temporal relationships and avoid overfitting to small training samples due to the rarity of the disease. In addition, these studies usually include burned-in respirometers to investigate the characteristic septal bounce, which can be a potential target leak of CP cases.

2 Materials and Methods

2.1 Population Selection

The study was approved by the Mayo Clinic Institutional Review Board (protocol #19-009303). Patients who underwent a TTE study at Mayo Clinic from January 01, 2003, to December 31, 2021, were reviewed to identify the cases of CP and cardiac amyloidosis (CA) [as the representative of restrictive cardiomyopathy (RCM)]. Specifically, the diagnosis of CP was confirmed by surgery, and the diagnosis of CA was established by endomyocardial

biopsy or advanced imaging modalities.[14] Cases with normal echocardiography were used as the control group.

Patients were excluded if any of the following conditions were present in the echocardiographic study: (1) inadequate echocardiographic images; (2) greater than or equal to moderate aortic/mitral regurgitation or aortic/mitral stenosis; (3) significant pericardial effusion; (4) prosthetic valve; (5) mitral/tricuspid valve annuloplasty; (6) conduction delay [greater than or equal to first degree atrioventricular (AV) block, left bundle branch block, or AV dissociation]; (7) intracardiac device such as a pacemaker; (8) cardiac resynchronization therapy device or implanted cardioverter device; and (9) increased respiratory effort (i.e., chronic obstructive lung disease, severe obesity), patients with significant pericardial effusion, patients with atrial fibrillation/flutter, and patients with severe right ventricular (RV) dysfunction. If any of the three parameters (hepatic vein, mitral inflow, medial e′) was not available, patients were excluded. We primarily selected Mayo Clinic Arizona data as our internal training and test set and Mayo Clinic Rochester data as the external test set. Even though both centers are within the Mayo enterprise, the practice pattern and acquisition devices are different between the two centers. We summarize the cohorts in Table 1, and Sec. 2.2 describes the performance of the models only on the hold-out internal and external test data.

2.2 Challenges Faced During the Model Design and Implementation

Our primary objective is to devise an efficient and generalizable model for differentiating CP, CA, and normal cases based on the standard apical 4 chamber (A4C) view from TTE studies. However in Secs. 2.2.1–2.2.3, we list interesting challenges that we observed while implementing a deep learning approach for TTE studies.

2.2.1 *Burned-in information about respiratory line entanglement*

Additional objects in the echo images can bias the model prediction—especially when subsamples from a particular class contain the prominent object. The CNN model learns to recognize those prominent objects as the discriminative marker of the class rather than anatomical features, e.g., burned-in text information in the chest x-ray images.[16] We observed that many CP images (not all) in our training dataset contain respiratory lines, highlighted in Fig. 1, which were specifically used to record variations over the respiratory cycle in the workup of CP cases.[15] These inconsistent respiratory lines are placed across the AP4 view images with high contrast against the background heart. Early in our experimentations, the existence of these respiratory lines became heavily correlated with the CNN discrimination of CP versus normal cases as seen by the gradient-weighted class activation mapping (GRADCAM) activation heatmaps focusing on the bottom right corners in Fig. 1. During experimentation, even if we removed the bottom ⅓ of the images (to remove most of the lines without removing too much of the image that held the pertinent information), the model still focused on the bottom corners of the images as the CP and normal cases still focused mostly on the bottom right corner where some of the respiratory lines were still present.

2.2.2 *Lack of available data for the rare diagnosis*

Ideally, the deep learning-based CNN model requires a significantly large amount of data to train the model. As a general rule of thumb, the size of a dataset should be at least about 10× trainable parameters of the model. Echo studies generate a sequence of frames as video and need sequential processing of frames for capturing the heart motion, which requires either a sequential model [e.g., recurrent neural network (RNN)] or a 3D CNN model for video processing. Given the limitation of data availability for the rare clinical cases, it is impractical to design a 3D CNN model targeted for video processing (trainable parameters >10M) with such a limited amount of data for clinical cases. In particular, in our targeted task, CP studies from tertiary referral centers are extremely rare.[15,17–19] The collection of an unbalanced dataset with mostly normal TTE cases may not provide enough power to the model for training the discriminative task of determining the subtle differences of the anatomy.

Table 1 Cohort characteristics: demographics and comorbidities. Statistics are shown for both internal and external datasets.

Characteristics	Internal train and test	External test
No. of patients (studies)	720	225
No. of frames	Total: 65,031	Total: 14,502
	Train: 38,202	
	Validation: 13,637	
	Test: 13,192	
Age (mean ± std)	54.9 ± 17.4	62.6 ± 17.5
Gender	Male: 53%	Male: 23%
	Female: 47%	Female: 77%
Race	White: 85%	White: 91.7%
	Asian: 13.7%	Asian: 4.3%
	Black: 1.3%	Black: 4.0%
Ethnicity	Hispanic or Latino: 12%	Hispanic or Latino: 8%
	Not Hispanic or Latino: 88%	Not Hispanic or Latino: 92%
Comorbidities at the time of TTE	Afib: 69 (10%)	Afib: 23 (10%)
	Cancer: 13 (2%)	Cancer: 11 (5%)
	Hypertension: 152 (21%)	Hypertension: 91 (40%)
	Chronic kidney disease: 83 (12%)	Chronic kidney disease: 9 (4%)
	Coronary artery disease: 76 (11%)	Coronary artery disease: 12 (5%)
	Diabetics (Type I and Type II): 47 (7%)	Diabetics (Type I and Type II): 24 (11%)
Diagnosis		
CA	CA: 197 (27.4%)	CA: 165 (73.3%)
CP	CP: 184 (25.6%)	CP: 16 (7.1%)
Normal	Normal: 339 (47.1%)	Normal: 44 (19.6%)
Device	GE (92.5%)	GE (100%)
	Philips (5.4%)	
	ACUSON (2.1%)	

2.2.3 *Issue with model generalization: image intensity differences between the internal and external data*

Generalization refers to the AI model's ability to adapt to new, previously unseen data. Traditionally, internal validation or holdout tests were performed to observe the generalizability of deep learning models. However, unique characteristics of medical datasets is the variety of manufacturers and the difference in acquisition protocols across institutions. For example, in echo devices alone, there are many manufacturers such as general electric (GE), Philips, Samsung, and Toshiba to name a few, with each manufacturer having many different machines, various probes, and acquisition parameters available. In addition, there are acquisition differences across institutions such as the A4C views being left-right flipped at Mayo Clinic

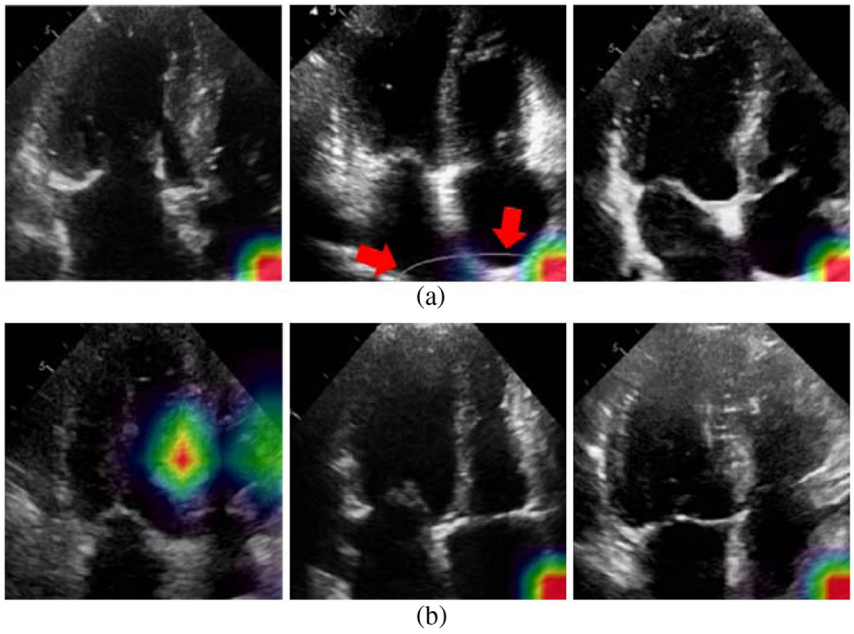

Fig. 1 Randomly sampled GRADCAM activation heatmaps overlaid on single frames of three different exams: (a) full input images and (b) bottom third removed input images. The red arrows highlight the burned-in respiratory line present in CP TTEs.

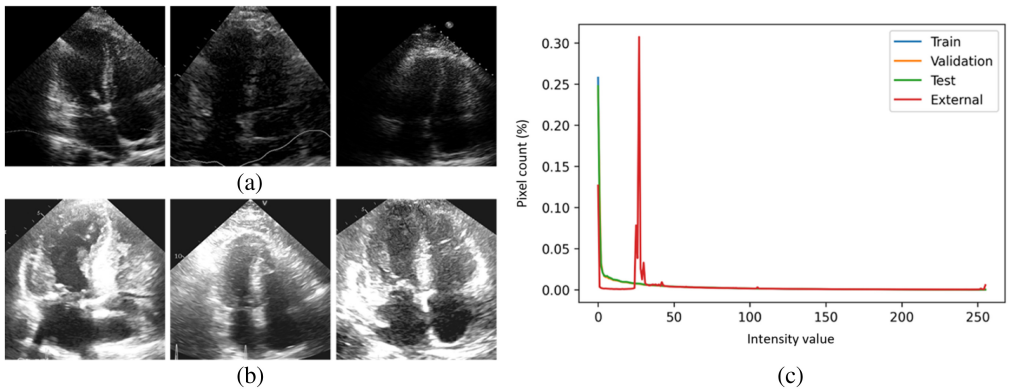

Fig. 2 Randomly sampled images of the internal and external sets and a graph of the intensity histogram distribution. (a) Random samples of the internal set, (b) random samples of the external set, and (c) internal (train, validation, test) and external set's normalized intensity histogram.

(known as "Mayo format") versus other institutions. With these factors in mind, we cannot reliably conclude that models are generalizable just by looking at their performance within only a holdout test set that is selected from the internal dataset and has homogeneous acquisition parameters. External data, e.g., data collected from different institutions, are necessary to judge the generalizability of the models across different manufacturers and acquisition protocols. We simply plot the differences between our internal and external datasets by comparing the differences of the intensity histogram between internal train, validation, test, and external test sets. Figure 2 shows that, although the internal sets are very similar to each other, to the point that they overlap, the intensity distribution of the external set was significantly different, making the generalization of our models difficult.

3 Methodology

Figure 3 shows a diagrammatic representation of the proposed framework that highlights the core processing blocks, namely the preprocessing block, shown in more detail in the Supplementary

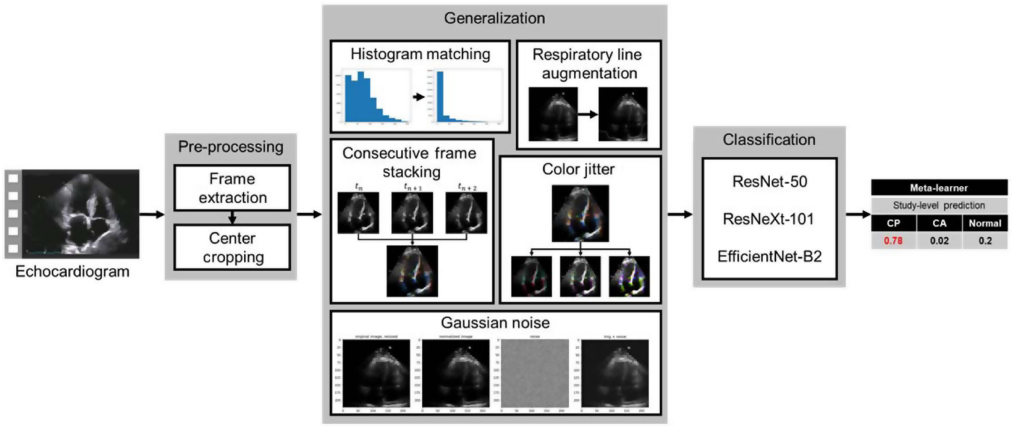

Fig. 3 Architecture of the proposed model. Blocks represent the different modules.

Material, generalization block, framewise classification block, and examwise diagnosis block. The trained framework is designed to directly read the video clips and produce a probabilistic diagnosis at the exam level. The preprocessing block is similar to the preprocessing done in our previous paper[20] in which the raw digital imaging and communications in medicine format (DICOM) file of the A4C TTE video clips is loaded and individual frames are separated and saved as.png files for further processing. The classification block tested several different families of model architectures (ResNet, ResNeXt, and EfficientNet) that were deemed to be relevant to the task at hand. Finally, a meta-learner block consolidates the framewise predictions into one examwise prediction.

3.1 Innovative Solutions for the Challenges

As mentioned in the introduction and literature review, there were no echocardiogram specific augmentations, only general computer vision augmentations such as geometric transformations and contrast adjustments. However, as mentioned in Sec. 2.2, common augmentation techniques could not resolve the problems of the model not generalizing well on the external datasets that have different intensity distributions and those with respiratory lines absent. As such, in our proposed framework, we integrate the following innovative solutions for the TTE study challenges mentioned above. The code used to preprocess and train the images with the solutions is available in a Github repository.

3.1.1 *Augmentation of artificial respiratory line templates*

To disentangle the burned-in respiratory lines from the model's decision-making criteria, we introduced a random augmentation of artificial respiratory lines in model training. We generated 10 artificial templates of the respiratory lines observed in our training data (see Fig. 4). From the set of 10 templates, the model randomly selects 1 artificial respiratory line and replaces the original pixel value in the image with the line template value. We coded a 50% chance of augmentation during model training. With the random addition of these artificial respiratory lines during model training, the model does not focus on the lines and extrapolates information much better from the heart image itself for its decision making as seen in the bottom row of Fig. 8.

3.1.2 *Assimilating heart motion*

As the echocardiograms are diagnosed by viewing them in motion, a 3D-video classifier (two dimensions of the echocardiogram image and one dimension of time) would be an ideal model architecture for this task. Given the rarity of the condition, we obtained a total of 720 patients, and only 184 of them were CP cases. Although this is a decent number compared with the previous literature,[13] the dataset is relatively small for 3D convolutional neural network model training. To train a deep learning algorithm to classify these videos, we increased the number of samples available by extracting the individual frames of the echocardiogram video. By training

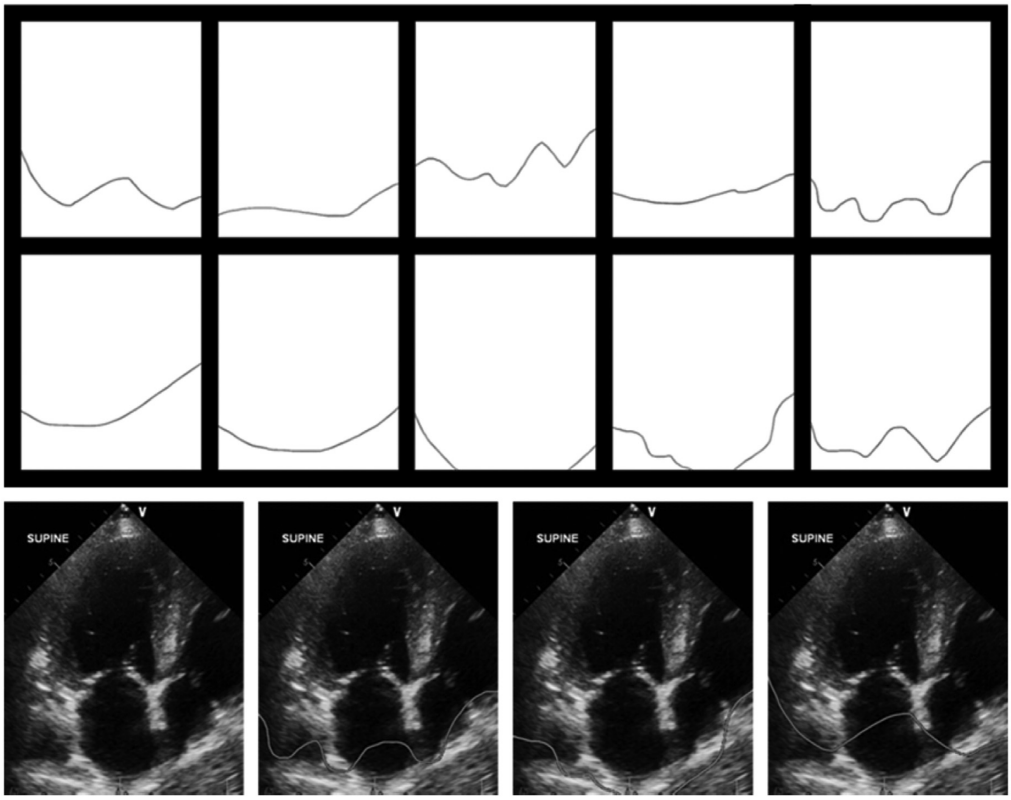

Fig. 4 Ten hand-drawn respiratory lines that were randomly sampled to augment the dataset. The last row contains the original frame without respiratory lines (far left) and three randomly drawn respiratory lines applied on the original image.

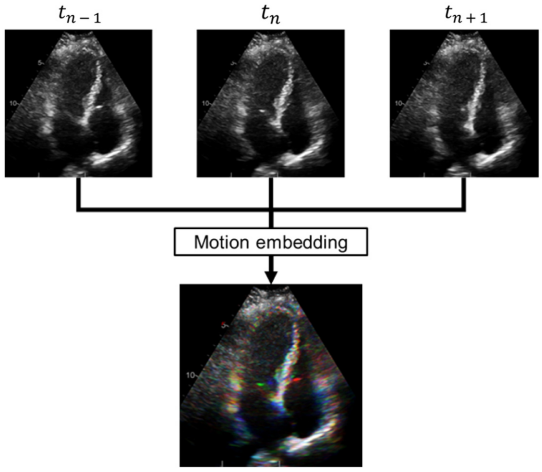

Fig. 5 Motion embedding.

a 2D deep learning model on individual frames, we increased the sample size from 720 to 65,031. However, CP is a disease in which the pericardium becomes stiff and interferes with the heart's pumping ability and motion information is very valuable and unfortunately removed from individual frame information. As mentioned above, individual frames do not have any motion information that is important for this task. To embed motion information into the model training, we combined 3 consecutive frames into a single RGB image as $\text{Image}_{\text{RGB}} = \sum_{n=1}^{3} \text{frame}_{t_{n-1}}$, where t_i is the index of the individual frame with the non-grayscale color representing the motion information, as seen in Fig. 5. The comparisons of a single frame versus motion-embedded RGB image are given in Sec. 4 (Table 4).

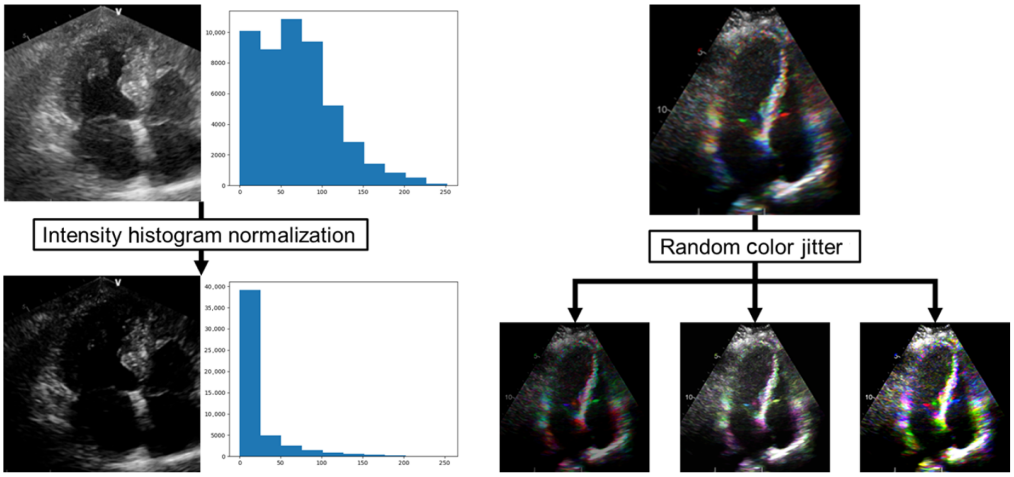

Fig. 6 Preprocessing and augmentations: (a) image histogram matched to the training data and (b) several examples of an image with color jitter.

3.1.3 *Generalization techniques*

As shown in Fig. 2, generalizability of our models on external data was a concern due to the difference in basic image characteristics, e.g., the intensity differences between the internal and external data and the possibility of overfitting with deep learning models. To mitigate such issues, we preprocessed the images by intensity histogram matching to normalize the intensity of the training and testing set and augment the training dataset with Gaussian noise and color jitter, as seen in Fig. 6. Matching the average intensity during preprocessing allows the model to see consistent images that are within the intensity distribution that the model expects, whereas the Gaussian noise mimics the specular nature of the ultrasound and allows the model to be more robust to ultrasound noise. Finally, the color jittering allows for the model to be robust to the different exposure of the ultrasound that may differ across institutions and technicians. In addition, to select the most generalizable model that does not overfit, we tested three families of deep model architectures focusing on the generalization aspect, ResNet, ResNeXt, and EfficientNet, which are discussed in Sec. 3.1.4.

3.1.4 *Model selection*

During the selection of an optimal deep learning architecture for an image classification task, one needs to take into consideration the heterogeneity of the input data, the complexity of the task, and the level of precision desired for the task. In general, these considerations become trade-offs between simple, lightweight models that are fast, moderately precise, and easily generalizable versus complex, deep models that are slow, very precise, and prone to overfitting. However, in medical imaging tasks, models must be sufficiently complex and deep to extract the underlying imaging marker from both high and low levels, and it is a matter of selecting a specific family of models with attributes that are appropriate for the target task. As such, we selected three different families of models: ResNet (residual block), ResNeXt (residual block with a split-transform-merge function), and EfficientNet (inverting residual blocks and scaling the resolution of the network), to test and determine which model performs the best and test their generalization on external data. ResNet[21] is a family of architectures that provides the advantages of a residual block: allowing networks to be deeper (more complex relations are extracted) while retaining pertinent image information. The key idea of the residual block is that, through skip connections and identity mapping, it allows the model/block to reference its original input rather than the encoded function. In simpler terms, it allows the model to refer to the original image when it makes its decision rather than some abstraction of the original image. This in turn allows deeper models to perform better than shallow ones (solving the degradation problem) and avoids the vanishing gradient problem in which redundant information confounds the model. ResNeXt[22]

architecture expands on ResNet's residual blocks by introducing a split-transform-merge strategy defined as the model's cardinality. The cardinality of the model determines the number of transformations applied to a lower-dimensional embedding, which are then summed. Each transformation extracts different features from the image and allows the model to get a comprehensive picture by viewing the problem from many different angles. EfficientNet[23] builds on the ResNet backbone in two different ways: inverting residual blocks and effectively scaling the depth, width, and resolution of the model. The inverted residual blocks from MobileNetV2[12] architecture "invert" the basic residual block by making skip connections between narrow-wide-narrow layers instead of the classic wide-narrow-wide layers, which is considerably more memory efficient. The depth, width, and resolution of the network are widely known parameters that contribute to the extraction of more rich and complex features, fine grain features along with easier training, and fine-grained patterns, respectively. However, each parameter comes at the cost of the difficulty of training due to vanishing gradients for depth, inability to capture higher level features for width, and diminishing returns for resolution. As such, EfficientNet systematically and uniformly scales each parameter with a compound coefficient, which achieves much better accuracy and efficiency and allows for fast and efficient models to be deployed in clinical practice.

3.1.5 *Model training and hyperparameter tuning*

We applied a 60:20:20 (train:validation:test) split on the internal dataset to train and internally test the model. The external dataset is only used for model testing. The validation split is used to tune the optimal hyperparameters using a grid search method among the following parameters: learning rate from [0.0002 to 0.000001] and weight decay from [0.0 to 0.5]. These hyperparameter ranges were determined from our previous paper that dealt with echocardiograms.[20] The hyperparameters used for the consecutive frame classification method were a batch size of 32, a learning rate of 0.000001, and a weight decay of 0.3 with cross-entropy loss for 200 epochs. The models were trained on an Nvidia RTX A5000 GPU. To prevent data leakage, we generated the split at the study level, so no images from the same study are mixed between the train and test.

4 Results

4.1 Comparative Analysis of Various Augmentation Methods

We evaluated the effect of the various augmentations in a stepwise manner to compare using the same training and test set: (1) state-of-the-art models with no augmentation, (2) state-of-the-art models with common augmentation strategies such as geometric augmentations, and (3) state-of-the-art models with our proposed method in Table 2. The table shows notable improvements in

Table 2 Tabular data showing the overall performance of the various model architectures and augmentation methods on the internal test set in terms of precision, recall, and *F*1-score. Optimal performance is highlighted in bold. 95% confidence interval added for framewise results.

	Average internal test performance								
	ResNet50			ResNeXt101			EfficientNetB2		
Class	Precision	Recall	*F*1-score	Precision	Recall	*F*1-score	Precision	Recall	*F*1-score
No augmentation	0.843 ± 0.008	0.836 ± 0.012	0.836 ± 0.012	0.876 ± 0.008	0.873 ± 0.010	0.873 ± 0.010	0.856 ± 0.009	0.851 ± 0.013	0.851 ± 0.013
Traditional augmentation	0.879 ± 0.007	0.874 ± 0.011	0.874 ± 0.010	0.873 ± 0.007	0.869 ± 0.010	0.869 ± 0.009	0.862 ± 0.010	0.855 ± 0.015	0.855 ± 0.014
Proposed method	**0.906 ± 0.007**	**0.903 ± 0.008**	**0.904 ± 0.008**	**0.892 ± 0.006**	**0.890 ± 0.006**	**0.890 ± 0.006**	**0.868 ± 0.010**	**0.864 ± 0.013**	**0.864 ± 0.013**

performance from no augmentation, common augmentations, and our proposed method. The table shows the best performance with a ResNet50 architecture with our proposed method. However, as shown later in the analysis, ResNet50 and ResNeXt101 seem to overfit with the performance improvements, whereas EfficientNetB2 has modest improvements and generalizes better.

4.2 Comparative Analysis of Various Model Architectures

We evaluated the quantitative performance of the model architectures on the same internal and external test sets in terms of area under the receiver-operating characteristic curve (AUROC) (see Fig. 7) and standard statistical metrics—precision, recall, and $F1$-score (internal test: Table 3 and external test: Table 4). Our models compute a framewise prediction, and finally, we aggregate the framewise prediction results to a studywise performance by averaging the prediction probabilities across all frames.

In Fig. 7 and Table 3, we do not observe any significant performance difference between ResNet50, ResNeXt101, and EfficientNetB2 on the internal test set. However, on the external dataset, EfficientNetB2 achieved the highest generalization performance (AUROC for CA 0.88, CP 0.82, and normal 0.79), which could be due to effectively scaling the depth, width, and resolution of the model on the validation data. We observe a significant drop in studywise performance for the CP class and normal on the external dataset (Table 4); we only obtained 16 CP cases and among them 3 cases are misclassified as normal, and among 42 normal cases, 12 are misclassified as CP, which were manually reviewed to be subtle or borderline cases as described in more detail in Sec. 5. Of the misdiagnosed external cases, a manual review by (Chieh-Ju Chao and Reza Arsanjani) showed that these cases were on subtle or borderline cases. Of the three misdiagnosed CP cases, two were identified as borderline cases with very subtle septal shifts, and the last one showed borderline constriction features that required catheters to confirm.

4.3 Comparison of Single Frame Versus Motion-Embedded RGB Image

We formed RGB images with three sequential gray-scale frames to capture the pericardium motion (see Sec. 3.1.2). To evaluate the efficiency of the proposed formation, we compared the performance of the optimal EfficientNetB2 model side-by-side using single-frame input, e.g.,

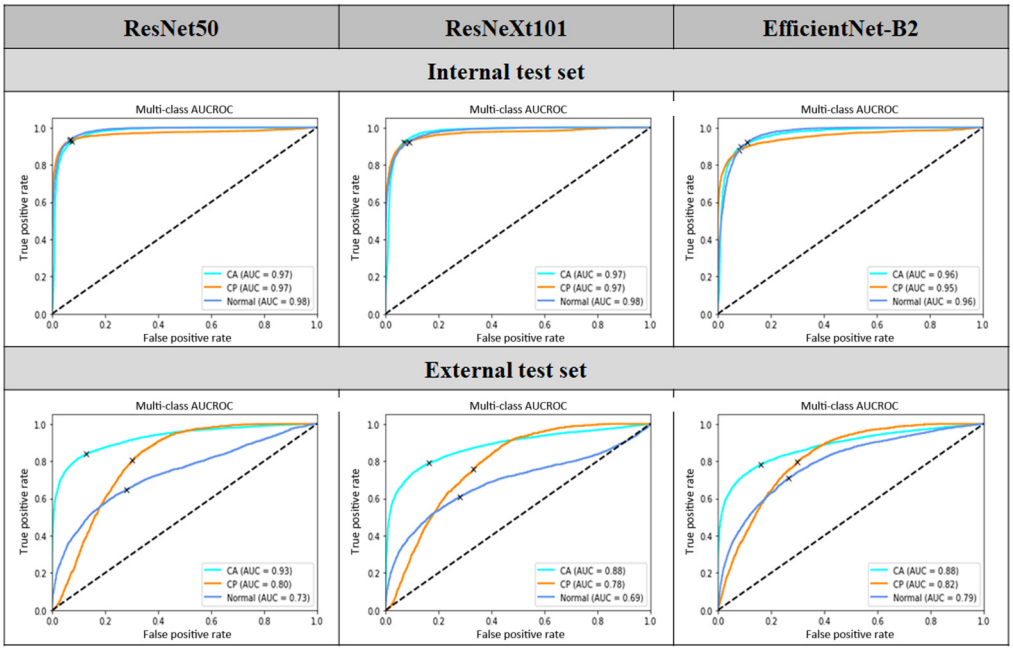

Fig. 7 Area under the receiver operating curve (AUCROC) curves for all three models were tested framewise. Each column denotes the different models, and each row denotes the internal and external test sets training and testing AUCs as well as the confusion matrix for each model is given in the Supplementary Material.

Table 3 Tabular data showing the quantitative performance of the various model architectures on the internal test set in terms of precision, recall, and *F*1-score. Optimal performance is highlighted in bold. 95% confidence interval added for both framewise and studywise results.

	ResNet50			ResNeXt101			EfficientNetB2		
Class	Precision	Recall	*F*1-score	Precision	Recall	*F*1-score	Precision	Recall	*F*1-score
Framewise internal test									
CA	**0.933 ± 0.002**	**0.934 ± 0.002**	**0.934 ± 0.002**	0.922 ± 0.002	0.922 ± 0.002	0.922 ± 0.002	**0.933 ± 0.002**	**0.934 ± 0.002**	**0.933 ± 0.002**
CP	0.916 ± 0.002	0.916 ± 0.002	0.916 ± 0.002	**0.951 ± 0.002**	**0.951 ± 0.002**	**0.951 ± 0.002**	0.912 ± 0.002	0.912 ± 0.002	0.912 ± 0.002
Normal	**0.931 ± 0.002**	**0.931 ± 0.002**	**0.931 ± 0.002**	0.929 ± 0.002	0.927 ± 0.002	0.927 ± 0.002	0.879 ± 0.002	0.874 ± 0.003	0.875 ± 0.003
Studywise internal test									
CA	0.914 ± 0.054	0.898 ± 0.046	**0.905 ± 0.038**	**0.918 ± 0.051**	**0.902 ± 0.043**	**0.909 ± 0.034**	0.911 ± 0.055	0.850 ± 0.053	0.878 ± 0.040
CP	**0.969 ± 0.030**	**0.922 ± 0.041**	**0.944 ± 0.027**	0.940 ± 0.042	0.921 ± 0.042	0.930 ± 0.031	0.911 ± 0.056	0.894 ± 0.048	0.901 ± 0.040
Normal	**0.938 ± 0.035**	**0.987 ± 0.013**	**0.961 ± 0.020**	0.920 ± 0.043	0.957 ± 0.026	0.938 ± 0.026	0.906 ± 0.049	0.971 ± 0.020	0.937 ± 0.029

Table 4 Tabular data showing the quantitative performance of the various model architectures on the external test set in terms of precision, recall, and *F*1-score. Optimal performance is highlighted in bold. 95% confidence interval added for both framewise and studywise results.

	ResNet50			ResNeXt101			EfficientNetB2		
Class	Precision	Recall	*F*1-score	Precision	Recall	*F*1-score	Precision	Recall	*F*1-score
Framewise external test									
CA	0.829 ± 0.002	0.764 ± 0.003	0.770 ± 0.003	0.842 ± 0.002	0.798 ± 0.003	0.803 ± 0.003	**0.850 ± 0.002**	**0.808 ± 0.003**	**0.813 ± 0.003**
CP	0.851 ± 0.003	0.686 ± 0.004	0.739 ± 0.003	0.884 ± 0.002	0.785 ± 0.003	0.817 ± 0.003	**0.859 ± 0.003**	**0.787 ± 0.003**	**0.814 ± 0.003**
Normal	**0.787 ± 0.004**	**0.800 ± 0.003**	**0.791 ± 0.004**	0.787 ± 0.004	0.798 ± 0.003	0.791 ± 0.003	0.769 ± 0.004	0.718 ± 0.004	0.735 ± 0.003
Studywise external test									
CA	**0.983 ± 0.012**	**0.824 ± 0.027**	**0.896 ± 0.017**	0.971 ± 0.020	0.708 ± 0.031	0.819 ± 0.022	**0.982 ± 0.014**	0.728 ± 0.033	0.836 ± 0.022
CP	0.207 ± 0.082	0.688 ± 0.106	0.311 ± 0.098	0.177 ± 0.075	0.752 ± 0.104	0.279 ± 0.097	**0.312 ± 0.108**	**0.804 ± 0.097**	**0.440 ± 0.117**
Normal	0.560 ± 0.133	0.429 ± 0.072	0.479 ± 0.078	**0.580 ± 0.136**	0.457 ± 0.073	0.504 ± 0.080	0.480 ± 0.132	**0.682 ± 0.068**	**0.553 ± 0.099**

Table 5 Comparison[a] between single frame and multi-frame input for the EfficientNetB2 models using the internal test set. Optimal performance is highlighted in bold[b].

	Single-frame test			Multi-frame test		
	Precision	Recall	*F*1-score	Precision	Recall	*F*1-score
CA	0.904 ± 0.002	0.906 ± 0.002	0.904 ± 0.002	**0.933 ± 0.002**	**0.934 ± 0.002**	**0.933 ± 0.002**
CP	**0.923 ± 0.002**	**0.921 ± 0.002**	**0.919 ± 0.002**	0.912 ± 0.002	0.912 ± 0.002	0.912 ± 0.002
Normal	0.863 ± 0.003	0.851 ± 0.003	0.851 ± 0.003	**0.879 ± 0.003**	**0.874 ± 0.003**	**0.875 ± 0.003**

[a]Performance metrics reported are bootstrap confidence intervals with 1000 iterations and a minimum sample size of 25% of the data.
[b]Bolded results are significantly different than their counterparts with at least one standard deviation of difference.

using one frame for all RGB channels, and multi-frame input, e.g., using consecutive frames for each RGB channels (Table 5). Although the model was able to extract information about the pericardium from single frame images and perform well on the test set, our experiments showed that adding the motion information significantly improved the performance of the model for CA and normal cases with a small reduction in CP cases on the internal test set, but the performance difference is minimal.

4.4 GRADCAM and Ablation Studies

GRADCAMs are often used to interpret the performance of the model by localizing the activations of the final convolution layer of the model during the investigation.[24] In Fig. 8, we show the activations of the model overlaid on top of the input image to show where the model was focusing when it made its classification on correctly and incorrectly classified images. In Fig. 8, the top row shows the correctly classified studies in which the model correctly focuses on the region of septal thickness and large atrium for classifying CP [Fig. 8(a)] and wall thickening for classifying CA [Fig. 8(b)]. Among the three wrongly classified CP cases on the external dataset, we visualized two representative samples. Figure 8(c) is particularly an RV focused view; it loses the left ventricle, the lateral annular motion seems okay, and atrium is not enlarged as much. Given these qualities and anatomical specification, we believe that the model predicted this case as normal instead of CP. Figure 8(d) is another external CP case that was classified as CA. From

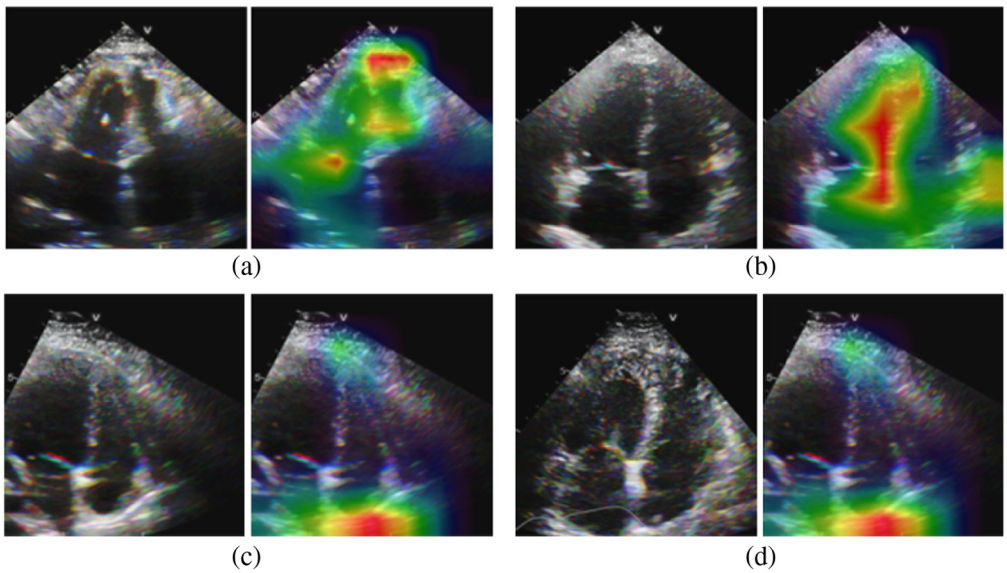

Fig. 8 GRADCAM images overlaid on the original echo frames: (a) correctly classified CP case; (b) correctly classified CA case; and (c), (d) wrongly classified CP cases.

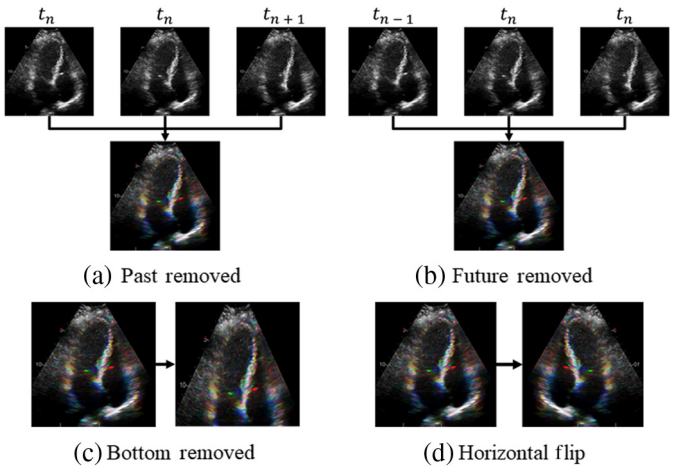

Fig. 9 Different types of ablations: (top row) time ablation, (bottom row) frame portion ablation, and image flip ablation. (a) Past removed, (b) future removed, (c) bottom removed, and (d) horizontal flip.

the GRADCAM, it seems that the model is looking at the right area, but the septal bounce for this case is not pronounced, which is one of the common characteristics for CP; thus it could be easily mistaken.

In the ablation study design, we focus on two parallel strategies: (1) time ablation and (2) frame portion ablation as shown in Fig. 9. We designed time ablation to understand the

Table 6 Side-by-side framewise and examwise performance for all ablation settings.

	Framewise performance			Examwise performance		
Class	Precision	Recall	*F*1-score	Precision	Recall	*F*1-score
			Ablate past frame			
CA	0.929 ± 0.002	0.929 ± 0.002	0.929 ± 0.002	0.913 ± 0.055	0.874 ± 0.050	0.892 ± 0.040
CP	0.898 ± 0.002	0.899 ± 0.002	0.897 ± 0.002	0.962 ± 0.035	0.788 ± 0.064	0.865 ± 0.042
Normal	0.873 ± 0.003	0.866 ± 0.003	0.866 ± 0.003	0.863 ± 0.061	0.986 ± 0.013	0.919 ± 0.036
			Ablate future frame			
CA	0.928 ± 0.002	0.928 ± 0.002	0.928 ± 0.002	0.911 ± 0.056	0.848 ± 0.054	0.877 ± 0.042
CP	0.905 ± 0.002	0.905 ± 0.002	0.903 ± 0.002	0.930 ± 0.053	0.814 ± 0.057	0.867 ± 0.042
Normal	0.880 ± 0.003	0.871 ± 0.003	0.872 ± 0.003	0.872 ± 0.060	0.986 ± 0.013	0.924 ± 0.035
			Ablate bottom third frame			
CA	0.932 ± 0.002	0.933 ± 0.002	0.932 ± 0.002	0.939 ± 0.044	0.876 ± 0.046	0.905 ± 0.033
CP	0.917 ± 0.002	0.918 ± 0.002	0.917 ± 0.002	0.934 ± 0.049	0.869 ± 0.055	0.899 ± 0.040
Normal	0.891 ± 0.002	0.882 ± 0.003	0.883 ± 0.003	0.910 ± 0.049	1.000 ± 0.000	0.952 ± 0.027
			Horizontal flip			
CA	0.927 ± 0.002	0.928 ± 0.002	0.926 ± 0.002	0.933 ± 0.047	0.775 ± 0.065	0.845 ± 0.047
CP	0.912 ± 0.002	0.913 ± 0.002	0.912 ± 0.002	0.935 ± 0.048	0.894 ± 0.046	0.913 ± 0.034
Normal	0.880 ± 0.002	0.865 ± 0.003	0.866 ± 0.003	0.856 ± 0.069	0.986 ± 0.014	0.915 ± 0.042

importance of the past and future frame reference. In time ablations, we replace the past frame or future frame using the current frame, which may cause some information removal of the cardiac motion. In the frame ablation, we removed the bottom part of the image that the model was focusing on for respiratory lines (see Fig. 1). To ensure generalizability to other institutions that may have reverse side acquisition, we also applied horizontal flipping to the test images and computed the results.

As can be observed from Table 6, the model performance reduces after dropping only the past and future frame for all of the targeted classes—(ablation of the past frame) CA $F1$-score: −0.004, CP $F1$-score: −0.015, normal: −0.009; (ablation of the future frame) CA $F1$-score: −0.005, CP $F1$-score: −0.009, normal: −0.003. This observation shows that the strategy of adding multiple frames helps the model's discrimination ability. Given our innovative augmentation of artificial respiratory lines, ablating the bottom third of the frame does not have any drop on the performance and even improved the performance for the CP and normal classes—CA $F1$-score: −0.001, CP $F1$-score: +0.005, normal: +0.008. Horizontal flipping was applied to see how the model would perform on a non-Mayo standard format, with the Mayo format being a horizontally flipped TTE. It also minimally affects the performance and thus shows generalization to our modeling strategy for other institutions.

5 Discussion

Echocardiography has been considered the first line diagnostic tool for CP.[15,25] However, the accuracy of an echo-based diagnosis largely depends on the quality of the echocardiography study, which requires skilled sonographers and experienced interpretation physicians.[25] Moreover, it is known that medical image analysis is impeded by a lack of machine learning and deep learning model generalizability and the ability of a model to predict accurately on varied data sources not included in the model's training dataset. In this retrospective study, we proposed some innovative and simplistic image preprocessing and augmentation techniques for increasing the generalization of the machine learning (ML) and deep learning (DL) models for echo images on external datasets. Leveraging the generalization techniques, we successfully developed an echocardiography-based deep learning model that can accurately differentiate CP from CA (as a representative of RCM) and normal cases and applied the model on both internal and external datasets even though the quality of the external dataset was significantly different from the internal ones. Our proposed techniques can be easily applied to other TTE image case studies.

CP is an uncommon but reversible cause of heart failure with preserved ejection fraction if identified correctly and sufficiently early. Common causes of CP included viral-induced, post cardiac surgery, or secondary to prior chest radiation; however, many cases are idiopathic.[15,18,26] The clinical presentation of CP can be similar to that of other myocardial diseases and may easily be confused with other causes of heart failure with preserved ejection fraction, or what is generally termed RCM. CA is another uncommon condition that involves protein infiltration of the myocardium, causing thickening of the cardiac walls and heart failure, and it may be used as a prototype for RCM. As such, it would be clinically useful to develop a model that identifies and differentiates CP from CA, two conditions that are challenging to diagnose clinically but with courses that may be positively altered with early diagnosis and therapy. Importantly, with the emergence of point-of-care ultrasound (POCUS) applications, the proposed framework can become a primer to facilitate the generalization between datasets obtained from traditional echocardiography and POCUS devices.[26] Furthermore, we hope our approach can facilitate future echo-based studies for other uncommon conditions with a limited training dataset.

The major strengths of this work include—(1) assimilating cardiac motion in a 2D image-based model using a sequential frame-based approach, which provides a computationally efficient option, avoids overfitting, and preserves the spatial-temporal relationship in the meantime. Our multiframe approach outperformed the single frame image models (Table 5). (2) Innovative augmentation techniques to deal with burned-in respiratory line entanglement strategy provide an efficient template-based way to confuse the model against learning the burned-in image information. In our external cohort, many of the CP images were obtained with a respirometer reflecting septal bouncing over the respiratory cycle as part of the diagnosis criteria.[17] The template-based augmentation strategy significantly reduces the cautionary image preprocessing

steps to deal with the burned-in information and help to preserve the original image quality. The GRADCAM and ablation analysis (Table 6) also demonstrated that the model was not relying on the respirometer curves for decision making. (3) The generalization technique for the external dataset with histogram normalization and color jitter helps to minimize the image variations for the external dataset. The proposed techniques can be easily extendable for other echo image case studies.

5.1 Limitations

The study is limited by its retrospective nature. Although the overall training sample size is relatively small from a machine learning perspective, our cohort contains one of the largest CP series available. The frame-based approach also provides a reasonable method of data augmentation and counters the issue of overfitting. Compared with a video-based approach, our frame-based approach may lose certain information about spatio-temporal relationships, but our model achieves an overall superior performance compared with the single frame performance. An online calculator of this model is currently under work, and we plan to release models and trained weights with the Massachusetts Institute of Technology open-source license to benchmark the performance.

6 Conclusions

In differentiating CP, CA, and normal cases using only the standard A4C view, the overall model performance using the generalization techniques on the internal dataset is AUC > 0.95 and on the external dataset is 0.83. We foresee the potential of this pipeline to enable an automated clinical workflow to improve the quality of interpretation and facilitate the diagnosis of CP. We foresee this model being used in daily echocardiography lab practice to improve the initial triage of CP and CA cases. This model will be especially useful for labs at institutions with limited diagnostic and/or therapeutic resources for the above conditions. Specifically, with the high accuracy in identifying CA and CP cases, this model will largely facilitate the early recognition of these patients and potentially improve prognosis. Furthermore, we hope our approach can facilitate future echo-based studies for other uncommon conditions such as CP with limited training datasets.

Disclosures

The authors have no conflicts of interest to declare.

Code, Data, Materials Statement

The archived version of the code can be accessed at https://github.com/jeong-jasonji/TTE_generalization/tree/main. The data and materials used in this paper contain PHI and are not publicly available for viewing.

Acknowledgments

The authors have no acknowledgments or funding sources to declare.

References

1. E. A. Amsterdam et al., "2014 AHA/ACC guideline for the management of patients with non–ST-elevation acute coronary syndromes: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines," *J. Am. Coll. Cardiol.* **64**(24), e139–e228 (2014).
2. I. Banerjee et al., "Developing an echocardiography-based, automatic deep learning framework for the differentiation of increased left ventricular wall thickness etiologies," Authorea Preprints (2022).
3. A. J. DeGrave, J. D. Janizek, and S. I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nat. Mach. Intell.* **3**(7), 610–619 (2021).
4. G. Duffy et al., "High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning," *JAMA Cardiol.* **7**(4), 386–395 (2022).
5. S. Gandhi et al., "Automation, machine learning, and artificial intelligence in echocardiography: a brave new world," *Echocardiography* **35**(9), 1402–1418 (2018).

6. K. He et al., "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. And Pattern Recognit. (CVPR)*, pp. 770–778 (2016).
7. M. Imazio et al., "Risk of constrictive pericarditis after acute pericarditis," *Circulation* **124**(11), 1270–1275 (2011).
8. M. M. Kittleson et al., "Cardiac amyloidosis: evolving diagnosis and management: a scientific statement from the American Heart Association," *Circulation* **142**(1), e7–e22 (2020).
9. F. P. Li et al., "Towards CT enhanced ultrasound guidance for off-pump beating heart mitral valve repair," *Lect. Notes Comput. Sci.* **8090**, 136–143 (2013).
10. L. H. Ling et al., "Constrictive pericarditis in the modern era: evolving clinical spectrum and impact on outcome after pericardiectomy," *Circulation* **100**(13), 1380–1386 (1999).
11. A. Madani et al., "Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease," *NPJ Digit. Med.* **1**(1), 1–11 (2018).
12. X. Yu et al., "Using deep learning method to identify left ventricular hypertrophy on echocardiography," *Int. J. Cardiovasc. Imaging* **38**(4), 759–769 (2022).
13. J. F. Silva et al., "Ejection fraction classification in transthoracic echocardiography using a deep learning approach," in *IEEE 31st Int. Symp. Comput.-Based Med. Syst. (CBMS)*, IEEE, pp. 123–128 (2018).
14. A. Østvik et al., "Myocardial function imaging in echocardiography using deep learning," *IEEE Trans. Med. Imaging* **40**(5), 1340–1351 (2021).
15. J. K. Oh et al., "Diagnostic role of Doppler echocardiography in constrictive pericarditis," *J. Am. Coll. Cardiol.* **23**(1), 154–162 (1994).
16. C. M. Otto et al., "2020 ACC/AHA guideline for the management of patients with valvular heart disease: executive summary: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines," *J. Am. Coll. Cardiol.* **77**(4), 450–500 (2021).
17. M. Sandler et al., "MobileNetV2: inverted residuals and linear bottlenecks," in *IEEE/CVF Conf. Comput. Vis. And Pattern Recognit. (CVPR)*, pp. 4510–4520 (2018).
18. M. Schwefer et al., "Constrictive pericarditis, still a diagnostic challenge: comprehensive review of clinical management," *Eur. J. Cardio-thorac. Surg.* **36**(3), 502–510 (2009).
19. R. R. Selvaraju et al., "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 618–626 (2017).
20. P. P. Sengupta et al., "Cognitive machine-learning algorithm for cardiac imaging: a pilot study for differentiating constrictive pericarditis from restrictive cardiomyopathy," *Circulation Cardiovasc. Imaging* **9**(6), e004330 (2016).
21. M. Tan and Q. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, PMLR, pp. 6105–6114 (2019).
22. T. D. Welch et al., "Echocardiographic diagnosis of constrictive pericarditis: Mayo Clinic criteria," *Circulation Cardiovasc. Imaging* **7**(3), 526–534 (2014).
23. S. Xie et al., "Aggregated residual transformations for deep neural networks," in *IEEE Conf. Comput. Vis. And Pattern Recognit. (CVPR)*, pp. 1492–1500 (2017).
24. J. Zhang et al., "Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy," *Circulation* **138**(16), 1623–1635 (2018).
25. W. A. Zoghbi et al., "Recommendations for noninvasive evaluation of native valvular regurgitation: a report from the American Society of Echocardiography developed in collaboration with the Society for Cardiovascular Magnetic Resonance," *J. Am. Soc. Echocardiogr.* **30**(4), 303–371 (2017).
26. M. P. T. Le et al., "Comparison of four handheld point-of-care ultrasound devices by expert users," *Ultrasound J.* **14**(1), 27 (2022).

Biographies of the authors are not available.