ORIGINAL RESEARCH

Classification of Echocardiogram View using A Convolutional Neural Network

Hannah Ornstein*, Dan Adam

Biomedical Engineering, Technion Institute of Technology, Haifa, Israel

Received: July 07, 2021 **Accepted:** August 18, 2021 **Online Published:** September 13, 2021
DOI: 10.5430/air.v11n1p1 **URL:** https://doi.org/10.5430/air.v11n1p1

ABSTRACT

The standard views in echocardiography capture distinct slices of the heart which can be used to assess cardiac function. Determining the view of a given echocardiogram is the first step for analysis. To automate this step, a deep network of the ResNet-18 architecture was used to classify between six standard views. The network parameters were pre-trained with the ImageNet database and prediction quality was assessed with a visualization tool known as gradient-weighted class activation mapping (Grad-CAM). The network was able to distinguish between three parasternal short axis views and three apical views to 99% accuracy. 10-fold cross validation showed a 97%-98% accuracy for the apical view subcategories (which included apical two-, three-, and four- chamber views). Grad-CAM images of these views highlighted features that were similar to those used by experts in manual classification. Parasternal short axis subcategories (which included apex level, mitral valve level, and papillary muscle level) had accuracies of 54%-73%. Grad-CAM images illustrate that the network classifies most parasternal short axis views as belonging to the papillary muscle level. Likely more images and incorporating time-dependent features would increase the parasternal short axis view accuracy. Overall, a convolutional neural network can be used to reliably classify echocardiogram views.

Key Words: View Classification, Echocardiography, Grad-CAM, Convolutional Neural Network

1. INTRODUCTION

1.1 Echocardiography

Echocardiography, or ultrasound imaging of the heart, can be used to assess both cardiac structure and function and to detect diseases.[1] In echocardiography, the ribs present an obstacle for imaging, as they attenuate most of the incoming waves, thereby blocking transmission of the beam to the heart.[2] To avoid imaging through the ribcage, sonographers will place the transducer at certain imaging windows.[3] These windows include the suprasternal, subcostal, left parasternal, right parasternal, and apical windows. The suprasternal window sits above the rib cage, while the subcostal window sits below it. For the left parasternal, right parasternal, and apical windows, the transducer is placed in between the ribs.[4]

During an examination, a sonographer will rotate the transducer at different windows in order to get different cross sections of the heart. These cross-sections allow the doctor to visualize different heart structures at different angles.[3] Short axis views from the left parasternal window and long axis views from the apical windows constitute two categories of these cross-sections that will be explored in this paper. While many of the same cardiac structures are shown in both categories, long axis apical views and parasternal short axis

*Correspondence: Hannah Ornstein; Email: ornsteinh@gmail.com; Address: Biomedical Engineering, Technion Institute of Technology, Haifa, Israel.

views are orthogonal from one another.[5] Long axis views taken from the apical imaging window include the apical two chamber (A2C), apical three chamber (A3C), and apical four chamber views (A4C). All three of these apical views include the left ventricle, left atrium, and mitral valve. The right ventricle, right atrium, and tricuspid valve can be visualized in the A4C view, while the aortic valve is shown in the A3C view. Short axis views from the left parasternal imaging window include the mitral (PSAX -MV level), papillary (PSAX -PM level), and apex (PSAX -AP level) views. All three of these views show the left ventricles and partially the right ventricle walls whereas the mitral valve is included in the PSAX -MV level view and the papillary muscles are shown in the PSAX -PM level view.[3]

The typical echocardiogram analysis pipeline starts with view identification. The identified view can then be segmented appropriately to identify specific cardiac structures present in the images. For example, the left ventricle can be segmented from the A4C view. From these segmentations, measurements can be calculated related to cardiac structure (such as left ventricle volumes), indices relevant to cardiac function (such as left ventricle ejection fraction) can be measured, and diseases (such as hypertrophic cardiomyopathy) can be detected.[1]

New technologies, such as portable ultrasound devices and cloud storage, have increased the need for automated echocardiogram analyses. For portable devices, automatic analysis is needed for non-expert users, and cloud capabilities have allowed for the storage of large databases where manually intensive analyses aren't a viable option. As the analyses of echocardiograms become automatic, there is an increasing need for automatic view classification, as it is the first step in the analysis pipeline.

Automatic classification of echo views is a challenging problem because of the great variability in these clips.[6,7] Different views can look similar to one another; for example, apical view subcategories (such as A2C, A3C, and A4C) are similarly oval shaped, whereas parasternal short axis (PSAX) subcategories (such as PSAX -AP level, PSAX -MV level, and PSAX -PM level) are similarly circular shaped. Additionally, a single clip may fluctuate between two different views if the transducer was placed close to the border between them. Even echos belonging to the same view can appear differently, as differing patient body composition, sonographer expertise, ultrasound machine, and speckle noise can introduce variations within the image quality and background data.

1.2 Deep Learning

1.2.1 *Convolutional Neural Network*

A possible solution for this challenging problem, is a deep learning approach, specifically a convolutional neural network. Convolutional neural networks (CNNs) are commonly used as the target function for image processing with machine learning.[8] An important element of this target function is convolution which is a mathematical procedure that is commonly used for feature extraction in image processing.[9] For example, a vertical edge filter can be convolved with an input image to extract all the vertical lines in an image. This output image of vertical lines is called an activation map because it highlights the parts of the image activated by the given feature. In a convolutional neural network, the target function learns what kernels (features) are most useful for the given task. While filters can describe horizontal/ vertical lines, blobs or other low dimensional features, deep networks have stacked convolutions so that learned filters describe more high dimensional features. In stacked convolutions, filters are convolved over an input image, and the resulting activation maps are stacked along their depth dimensions. Filters are then convolved along these stacked volumes. These filters describe combinations of lower dimensional features. Therefore, filters in deeper convolutional layers generally describe higher dimensional features.[10] For example, in a network trained to classify cars, features at the lower levels may describe vertical and horizontal edges, where those at convolutional layers towards the middle could describe circles and grid patterns, and finally those towards the end layers may describe wheels and tires. Common convolutional neural networks include the Residual Network (ResNet),[11] which is an architecture created by Microsoft research that classifies natural images extremely well and won the ImageNet challenge in 2015 with only a 3.57% error.

1.2.2 *Transfer Learning*

In deep networks, there are millions of parameters that need to be optimized in the target function. Training the network to find the most optimal features would theoretically require millions of examples to train.[12] This is an issue especially in the medical image processing domain where the datasets are of limited size. A possible solution is transfer learning, which is a technique where a model developed for task a is reused as the starting point for the model on task b. Usually task a is trained on a huge dataset whereas task b has only a limited sized dataset. The purpose is to apply the knowledge learned from a related task as the initialization for another task of limited data.[12] ImageNet,[13] which is a well-known (and very large) database of natural images, is commonly used as the dataset in "task a" for a smaller image classification "task b".

1.2.3 *Grad-CAM*

Deep network results can be deceiving, as high accuracies do not always indicate reliable networks. For example, a network could have high classification accuracy but base its decision on "incorrect" features due to a bias in the input dataset. In Selvaraju et al., a network was tasked to identify between nurses and doctors.[14] While it performed with high accuracy on the training set, the net was using "incorrect" features such as face and hair to distinguish between nurses and doctor. This was due to a bias in the dataset, as the majority of nurses were female and the majority of doctors were male. For applications in the medical field, it is important to understand how the network makes decisions, and why, or why not, it produces high results. Gradient-weighted class activation mapping (Grad-CAM) is a relatively new visualization tool that can be used to analyze learned networks.[14] Grad-CAM produces coarse heat-maps that highlight the region of an image recognized as a given class. These class discriminative coarse heat-maps provide a visual explanation for how a model made its decision. In short, Grad-CAM uses the gradients flowing back into a convolutional layer to create its heat maps. The gradient size determines which feature maps most greatly affect the score prediction. The feature maps that most affect the score are weighted more heavily. When all the weighted feature maps in the convolutional layer are combined, only the pixels that have a positive influence on the class are kept since these pixels increase the class prediction.

Depending on the convolutional layer chosen, Grad-CAM can highlight different features of interest. While the lower layers are known to have lower dimensional features and small receptive fields, deeper layers have higher dimensional features.[10] Grad-CAM can be used at different layers to assess both low and high dimensional features.

1.3 Previous Works

Previous works that have attempted to automatically classify echocardiogram views can be divided into two main categories. The first approach is used in earlier works and involves manually picked features as input to a machine learning classifier. For example, Khamis et al.[7] extracts a region with important spatio-temporal features using a cuboid detector and uses supervised dictionary learning as the classifier. Using this method, the A2C, A3C, and A4C views were classified with 95% accuracy. Many other papers use a support vector machine (SVM) as their machine learning classifier of choice such as in Refs.[15–20]. The reader should refer to Penatti et al.[21] for a review on this approach.

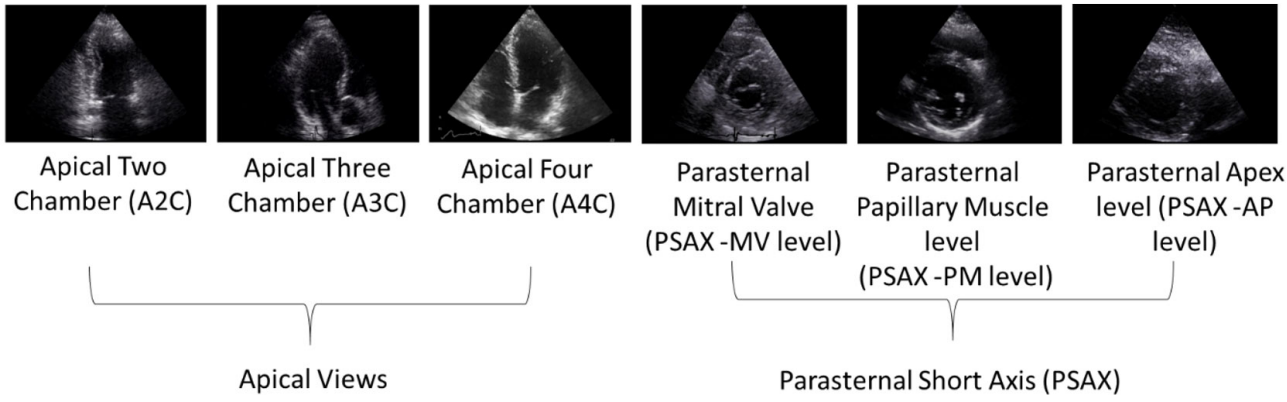


Figure 1. Examples of frames from the six views. A2C, A3C, A4C, PSAX –MV level, PSAX –PM level, PSAX –AP level comprise the views categorized in this paper. These views belong to the Apical and PSAX categories as detailed in the figure. Other articles may refer to A3C as Apical long axis (ALAX).

Manually chosen features may not be optimal for classification. In the second approach, which is found in newer works, deep networks attempt to learn the "optimal" discriminative features. For example, Zhang, Jeffrey, et al.[1] classified six echo views as part of an analysis pipeline using a VGG network and achieved accuracies of 93%-97% for apical views. Zhang later expanded on this initial study to include twenty-three subcategories which included views containing occlusions.[22] Madani, Ali, et al.[23] also utilized a VGG network to classify fifteen echo views and found that low resolution images were classified by the network better than an expert. They achieved frame accuracies of 87%-95% for the PSAX -PM and PSAX -MV levels (which they included as one group), A2C, A3C, and A4C views. Ostvick et al.[24] classified seven echo views using an original CNN architecture and had access to a very large database of images. They achieved 98%-99% accuracies for A2C and A4C clips. Gao, Xioahong, et al.[25] used both hand crafted and network

learned features to achieve an overall accuracy of 92.1%. They fused two networks to classify eight views; one learned spatial information from echo clips while the other applied the feature of acceleration (derived from optical flow images) to incorporate temporal information.

In this study, we automatically classify an echocardiogram database into six common views including three apical views (A2C, A3C, A4C) as well as three parasternal views (PSAX -AP level, PSAX -MV level, and PSAX -PM level) (see Figure 1). The A2C, A3C, and A4C views image the heart lengthwise, while the MV, PM, and AP views image the heart along its cross-section. We use a deep learning approach and chose ResNet as our architecture due to its high accuracy in image classification. Additionally, we utilize transfer learning and initialize the network's weights with those trained on ImageNet. We train our network in a ten-fold cross validation scheme and assess the results using Grad-CAM.

2. METHODS

2.1 Database

The echocardiogram database was composed of clips collected from Assaf Harofeh Medical Center, Carmel Medical Center, Emek Medical Center, Shaare Zedek Medical Center, Soroka Medical Center, Rambam Medical Center, Hillel Yaffe Medical Center, University of Leipzig, Kaplan hospital, Assaf Harofeh Medical Center, and Hadassah Hospital - Har HaTsofim. The database included healthy and abnormal echocardiograms of varying image quality. Depending on the source, data was saved in either DICOM, AVI, JPEG, or MAT files and was manually labeled by one or two experts. The database consisted of 681 echo clips divided among the six classes as follows: 170 clips or 23,351 frames for the A2C views, 180 clips or 23,844 frames for the A3C views, 163 clips or 23,005 frames for the A4C views, 86 clips or 10,822 frames for the PSAX -PM level view, 41 clips or 5,268 for the PSAX -MV level view, and 41 clips or 5,118 for the PSAX -AP level view. The clips' frames were of size 636X434.

2.2 Data Preprocessing

For pre-processing, the periphery information in the images were removed in a procedure similar to that in Zhang, Jeffrey et al.[1] In short, pixels that remained static throughout the entire clip were removed. Additionally, colored pixels from each frame were removed by masking out pixels with large standard deviations (>9) over their color channels. A histogram was used to find the standard deviation value that distinguished between gray and colored pixels. Preprocessing was conducted using MATLAB software.[26]

2.3 Model Architecture and Training

The network parameters were initialized by weights pretrained on the ImageNet database.[13] During training, updating occurred on all the parameters in the network. For stochastic optimization, the ADAM optimizer was used.[27] The initial learning rate was 0.001 and had a mini-batch size of 95 frames.

The average accuracy from 10-fold cross validation runs was used to approximate the model's accuracy. Classes were split for each fold so that roughly 90% of a specific class was in the training fold, and 10% was in the validation fold. A weighted cross-entropy loss was used to account for the imbalanced dataset. The weights chosen for each class equaled the number of samples in the largest class divided by the number of samples in a given class and were: A2C) 1.02 A3C) 1 A4C) 1.037 PSAX -PM level) 2.20 PSAX -AP level) 4.536 PSAX -MV level) 4.66 respectively. Frames served as inputs to the network and those from the same video belonged to either training or validation and were not mixed between the two categories. Videos were classified to the label predicted for the majority of its frames.

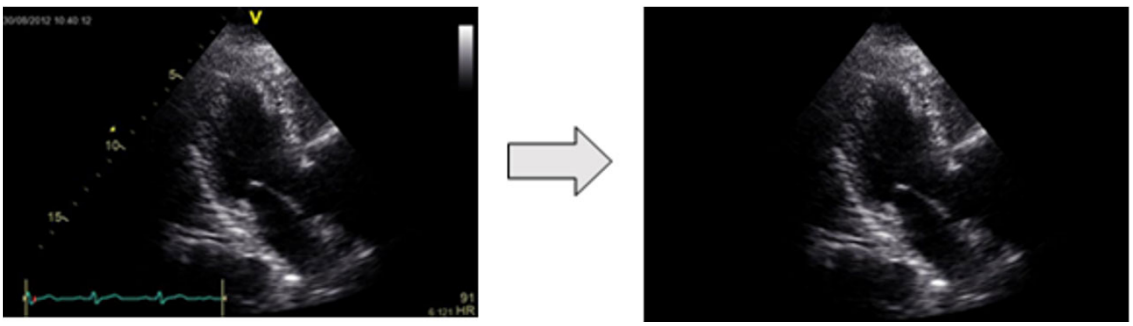


Figure 2. Echocardiogram after preprocessing. Original image on left is preprocessed to remove metadata in background. Image on the right is the resulting image

Grad-CAM images were used to visualize the features learned by the network. To view features at different layers, the layer and basic block for the Grad-CAM classification were specified (see Figure 3). Additionally, the class label was specified in the Grad-CAM code to observe features for different classes.

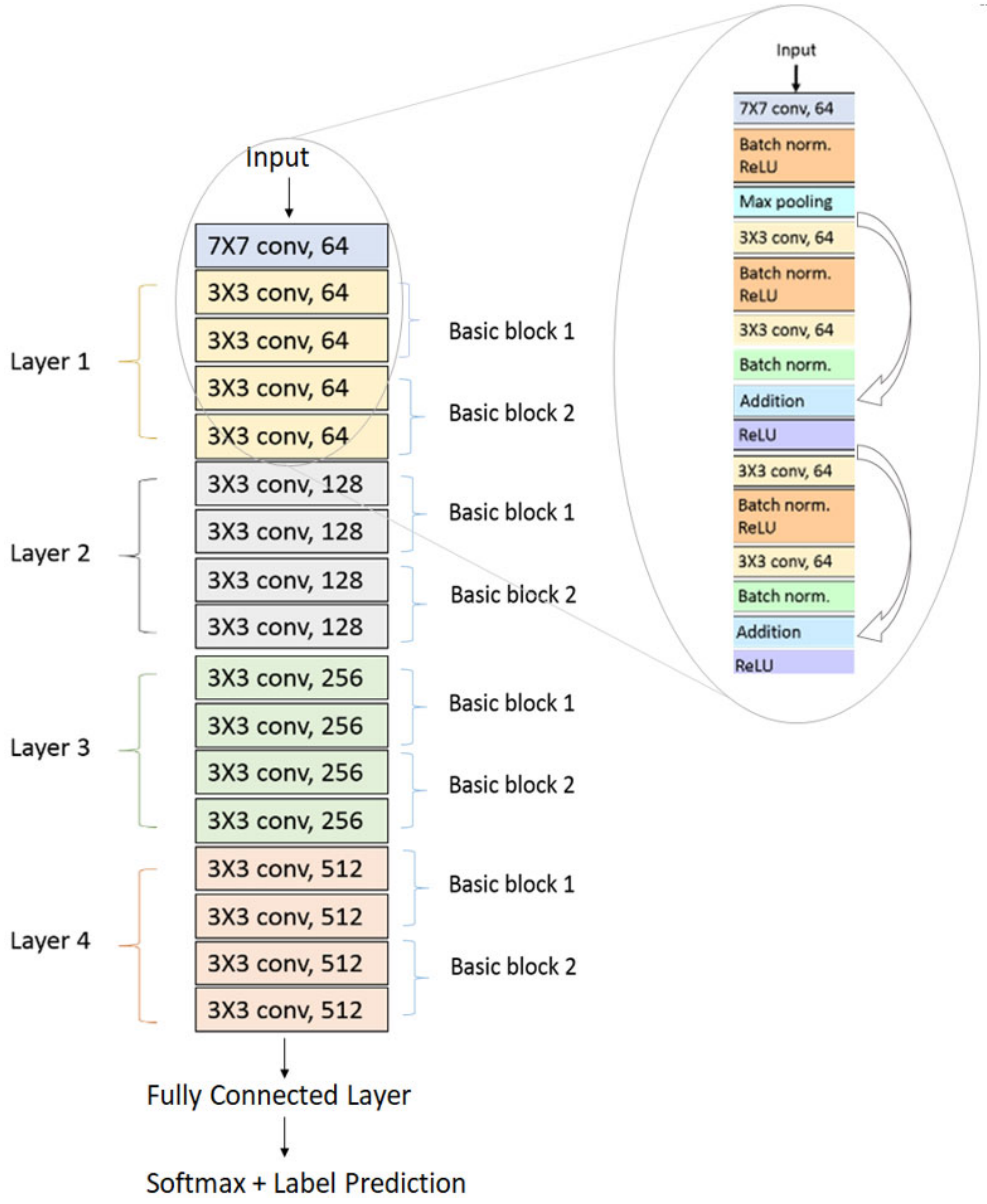


Figure 3. Labels for convolutional layers in ResNet-18. Specific convolutional layers for ResNet-18 are labeled according to their "Layer" and "Basic block". Each of the four "Layers" consist of four convolutional layers with an equal number of output filters (i.e. 64 output filters in "Layer 1"). The "Basic block" label pairs the convolutional layers according to the presence of skip connections (see enlarged figure on the right for illustration of complete "Layer 1" with inclusion of skip connections).

2.4 Data Transformations

Transformations to the training dataset simulated changes in gain, zoom, and rotation. Transformations to each image included: conversion to grayscale, a randomly placed crop of 224 x 224, an adjustment of brightness by a random factor between 0 and 2, a random rotation within the ranges of [-1,1] degrees, and a horizontal flip 50% of the time. Gray channel values were repeated over three channels to accommodate the ResNet architecture. Images were normalized to ImageNet's mean: [0.485, 0.456, 0.406] and standard deviation: [0.229, 0.224, 0.225]. For the validation set, the 636 x 434 validation images were rescaled to 256 x 256 and then center-cropped

to 224 × 224. This crop ensured that the frame had the right dimensions for ResNet and that the resulting image included most of the frame information.

2.5 Software and Hardware

The pytorch library with python 3 was used for creating and training the networks. The ResNet-18 architecture, pre-trained ImageNet weights, and data transformations were accessed using the torchvision package for PyTorch.[28] Code used to update weights from certain layers was based on Ref.[29] The Grad-CAM script was based on code written by Kazuto Nakashima and Utku Ozbulak[30,31] The GPU used was a Tesla K20m which had 5120 MB of memory.[32]

3. RESULTS

A confusion matrix of the averaged results from the 10-fold cross validations show high accuracies >97% for apical (A2C, A3C, A4C) view classification (see Figure 4). Additionally, the confusion matrix indicates that the network rarely classified (<1%) parasternal short axis subcategories (PSAX -PM level, PSAX -MV level, PSAX -AP level) as apical subcategories. Most incorrectly classified parasternal short axis views were predicted as PSAX -PM level.

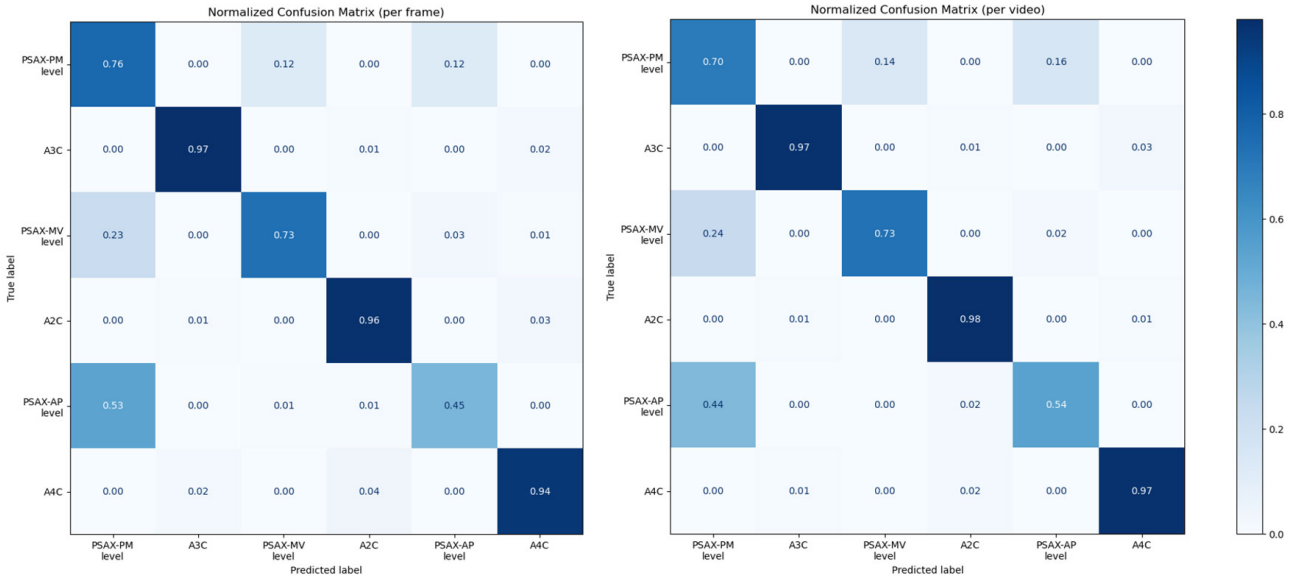
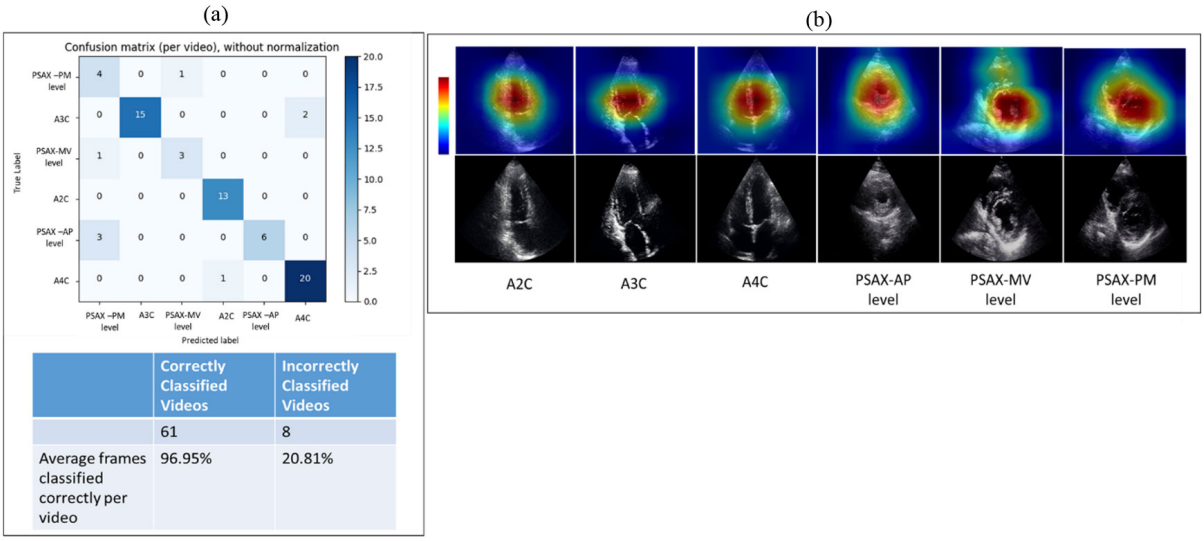


Figure 4. Confusion matrices averaged over 10-fold cross validation results. Accuracies are calculated on frames (left) or videos (right). Values on the diagonal line indicate matches between true labels and predicted labels. Apical window subcategories (A2C, A3C, A4C) show accuracies of 97%, and 98% (video classification). Parasternal short axis subcategories (PSAX -PM level, PSAX -MV level, PSAX -AP level) show accuracies of 54%-73% (video classification). Results are normalized with respect to the number of frames/videos in each class.

For a given network from the 10-fold cross validations, Grad-CAM heatmaps highlighted main regions of clinical interest (see Figure 5). Particularly, for the parasternal short axis subcategories, heatmaps were centered around the left ventricle.

For a correctly predicted apical four chamber view, Grad-CAM heatmaps from earlier layers of the net learned lower-level features such as edges recognizing the septums, valves, and myocardium. While, from the second and third layers of the net, Grad-CAM heatmaps highlighted regions pertaining to the left ventricle walls, the atrial septum, the valve level, and the atrium. In the deepest layers, Grad-CAM heatmaps highlighted clinically relevant features such as the left ventricle and ventricular septum strongly and more exclusively (see Figure 6a). This correctly classified view was predicted as A3C with .002% probability, A2C with .001% probability, and belonging to one of the parasternal short axis levels with 0% probability. The corresponding Grad-CAM images highlight some relevant features for the different apical views but show an almost inverted heatmap for the parasternal short axis views when given the corresponding labels. By using Grad-CAM in earlier layers, the valve level was highlighted strongly for the correct class, weakly for the A3C view, and very negatively for the A2C view (see Figure 6b).



	Correctly Classified Videos	Incorrectly Classified Videos
	61	8
Average frames classified correctly per video	96.95%	20.81%

Figure 5. GradCAM analysis for correctly classified views (a) Confusion matrix results from one of the 10-fold cross validation networks. All frames shown here were from the validation set from one of the 10-fold cross validation networks. Results in the confusion matrix were not normalized and represent the number of videos labeled in a particular category. (b) Grad-CAM heatmaps were calculated from "Layer 4" in ResNet-18. The original frames are included below the Grad-CAM images for reference. A red to blue scale indicates areas of highest to lowest importance for classification.

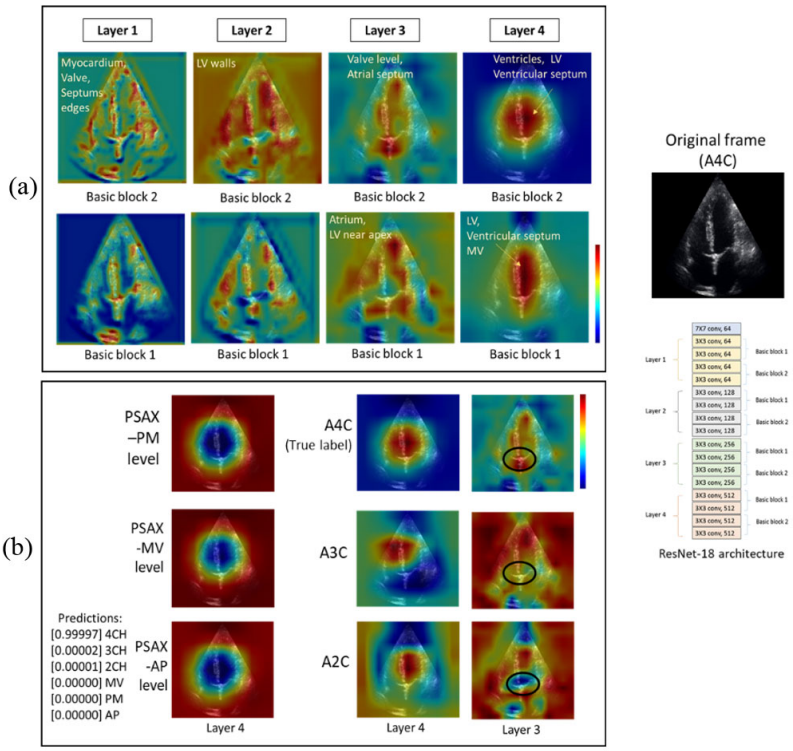


Figure 6. GradCAM analysis for a correctly classified frame of apical four chamber view. Original frame and ResNet-18 architecture are provided for reference. (a) GradCAM algorithm was performed to different "Layers" and "Basic Blocks" of ResNet-18. Different anatomical markers were recognized by the network as useful features for classification. Some of these clinically relevant features are labeled in the figure. (b) GradCAM heatmaps generated from all possible six labels. Heatmaps generated from "Layer 4" for the three parasternal short axis labels show an almost inverted heatmap to the correct apical four chamber view label. The valve-level feature is circled in the heatmaps generated from "Layer 3" for the three apical view labels. The network's predictions for the original frame is included for reference.

The few misclassified apical views in the given network were either of poor quality, mislabeled, or had frames which fluctuated between two different views (see Figure 7). For misclassified parasternal short axis videos, Grad-CAM images are centered over the left ventricle for the papillary muscle level, but do not highlight clinically relevant features for either the mitral valve or apical levels (see Figure 8).

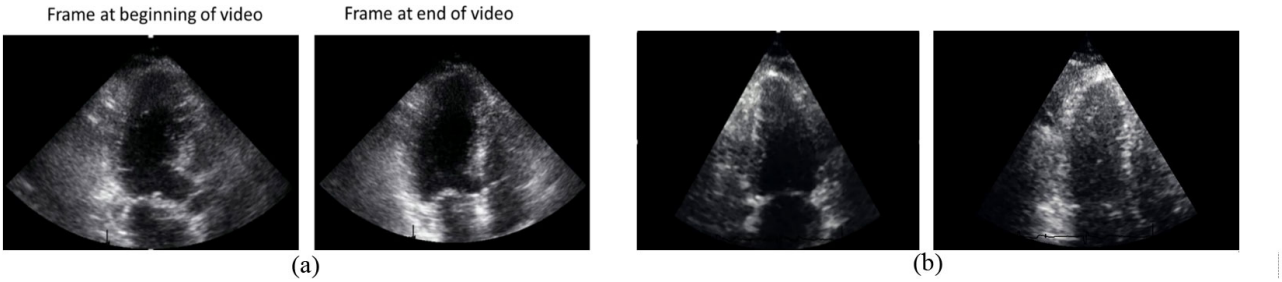


Figure 7. Misclassified apical view clips. (a) A3C clip misclassified as A4C. The clip was labeled as a A4C view but fluctuates between the A3C view and A4C view due to transducer position and heart motion. The network predicted the first half of the video frames as a A3C view and the second half as a A4C view. Still frames taken from the beginning and end of the video do appear as a A3C and A4C view respectively. (b) A4C clip.

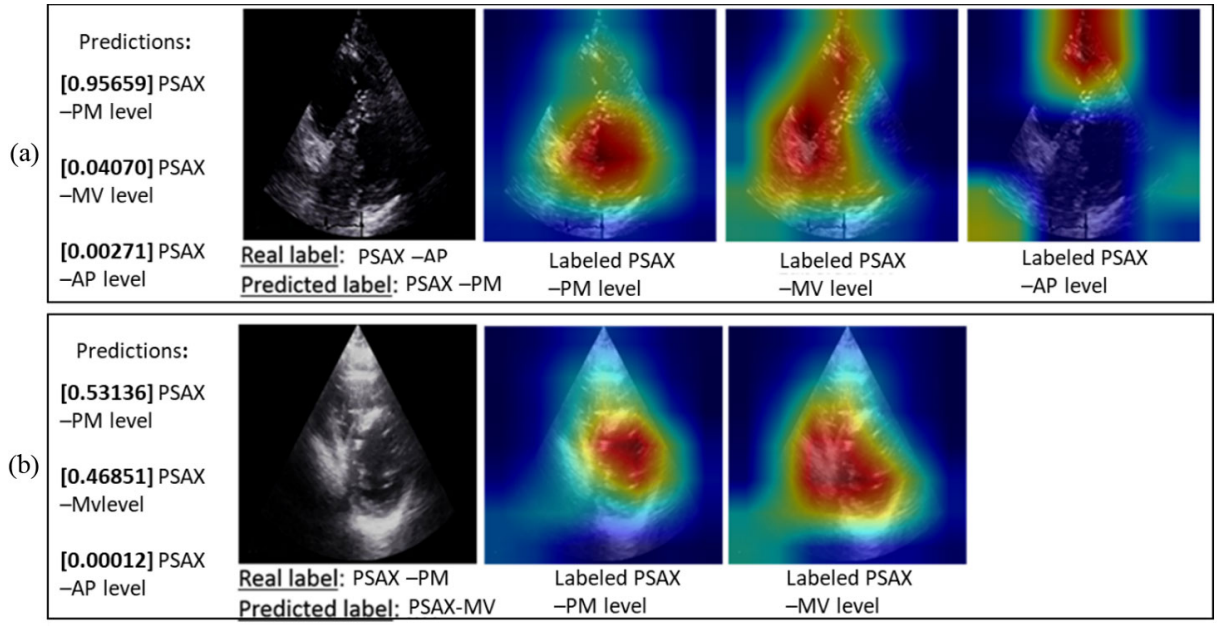


Figure 8. Misclassified parasternal short axis clips. The network's predictions for the given frame are included for reference. (a) Grad-CAM from a misclassified apical view clip. Clip was misclassified as a parasternal papillary muscle level view. Grad-CAM heatmap with PSAX -PM level label was centered over the cross section of the heart, whereas Grad-CAM heatmap with PSAX -AP label was very off-center. (b) GradCAM from a misclassified parasternal papillary muscle level clip. Clip was classified as a parasternal mitral valve level view, though this particular frame was correctly classified. Grad-CAM heatmap with PSAX -PM level label is centered over the left ventricle, whereas the Grad-CAM heatmap with PSAX -MV level label is off-center.

4. Discussion

4.1 Classification of Parasternal Short Axis versus Apical views

The network distinguished between parasternal short axis views and apical views extremely well (with >99% accuracy). Corresponding to this high accuracy, given Grad-CAM heatmaps showed no overlap between the learned parasternal short axis and apical features (see Figure 6b). The two categories were easily distinguishable likely because the classification was "simpler" to learn as these views are quite

different from one another; parasternal short axis views are more circular while apical views are more oval. It is important to note that many previous works do not distinguish between the parasternal short axis subcategories; labeling them as one category,[5,6,16,18,33,34] or not including them at all.[1,7,24] The ease of classification between the parasternal short axis category and other apical views should be taken into account when comparing our results to their reported accuracies.

4.2 Classification of Apical views

The trained network was able to classify the three apical views very well (to 98% accuracy) with minimal preprocessing (see Figure 4). These reported accuracies are a reliable estimation of the "true" accuracies as they are not merely a consequence of an easy to classify test set; rather, they are an average of the ten network accuracies from the cross validations. In a given network, all misclassified apical views had reasonable explanations, further illustrating the network's reliability (see Figure 7). Due to the mislabeling, fluctuations, and low-quality videos, the "true" accuracy values for apical views may be even higher than those reported. Misclassified clips that were of low quality and/or mislabeled looked more like their predicted view than their manually labeled view. Those clips would likely be unusable for analyses requiring the view corresponding to their manually given labels. Fluctuating clips, such as that in Figure 7a, could be removed from the database since the frame classifications are split between two different classes. A cut-off could be set to determine the percent of frames needed for clip classification to remove fluctuating videos instead of using a majority vote. More experiments would be necessary to determine the appropriate cutoff.

Grad-CAM images showed that the network recognized class-specific features for apical views similar to those used by an expert (see Figure 5). The heat maps for the apical views were centered a bit above the valves, showing a slight preference for the ventricles as opposed to the atrium. Similarly, in Zhang et al.,[22] occluding the left ventricle impaired classification more than occlusion of the left atrium. Figure 6a illustrates that the network recognized clinically relevant landmarks. As is expected in deep networks, features learned towards the end of the architecture were more high level than those learned in the earlier layers. Interestingly, the valve level area was highlighted by the Grad-CAM as an important feature for distinguishing between apical subcategories (see Figure 6b) which is similar to the findings in Khamis et al.[7]

4.3 Classification of Parasternal Short Axis views

While classification accuracies of the parasternal short axis subcategories were low in comparison with the apical subcategories (see Figure 4), Grad-CAM images showed that the learned features were centered over the "correct" regions of interest for accurately classified videos (see Figure 5). Theoretically, when given infinite data and the proper network architecture, the network should learn the optimal features which may or may not correspond to those features that we believe to be "correct". With that in mind, we make the reasonable assumption that the "correct" region of interest for parasternal short axis views is the area within the left ventricle as it contains either the mitral valves (for PSAX -MV level), papillary muscles (for PSAX -PM level), or an empty space (for PSAX -AP level). To reiterate, the Grad-Cam generated heat maps for correctly classified parasternal short axis views were centered around this region of interest.

Unlike the apical views, parasternal short axis videos that were misclassified by the network included those of good quality (see Figure 8). Misclassified parasternal short axis views were usually labeled as parasternal papillary muscle level (see Figure 4) and Grad-CAM heat maps suggested that the network recognized features relevant for the parasternal papillary muscle level regardless of the correct class (see Figure 8). The over classification of PSAX -PM level views may be a consequence of our imbalanced dataset; the network was trained on double the amount of PSAX -PM level views (in comparison to PSAX -AP level and PSAX -MV level views). More parasternal papillary level views may have caused the network to learn features pertaining to this class or to classify more frames as this class. Possibly, optimizing the weights for the cross entropy would balance out the effect of the different sized classes, although, the network likely needs to train on more parasternal short axis images. Assuming that clips were evenly distributed among the classified views, other deep networks for echo classification (excluding Gao et al.[25] who incorporated manually extracted features) had a minimum of 13,000 frames per class.[1,23,24] This is greater than the number of frames for all parasternal short axis subcategories in our database.

Since parasternal short axis views resemble one another and are more challenging to visually classify (even manually) than the apical views, it may be necessary to incorporate information from movement instead of relying solely on structure. The current network considered spatial features only, but time may be an important feature for the classification of parasternal short axis views. Gao et al.[25] showed that after incorporating a convolutional network using optical flow-based images (representing acceleration), accuracies for PSAX -MV level images increased from 68.8% to 75%,

though accuracy for PSAX -PM level views did not change. Time dependent features could be incorporated by extracting our network features to use as input to a long short-term memory network (LSTM) over a given clip. Likely movement based features are less important for apical view classification as the network was able to recognize the valve level feature which Khamis et al.[7] found to be an important spatial and temporal feature. Additionally Kumar et al.[17] achieved very high accuracies (88%) for parasternal short axis subcategories and very low accuracies for apical subcategories (65%) when using motion features as input to a learned dictionary and SVM.

4.4 Additional Comparisons to Other Works

A limitation of this study is that Grad-CAM provides a qualitative analysis and its results are only shown for one of the cross validation networks. Nevertheless, the Grad-CAM images support the results of the averaged 10-fold net accuracy and help promote the believability of the network. To our knowledge, we are the only paper of an echo classification network which visualizes within the layers. Other deep network echo classification papers may use t-SNE[1,23] or very general occlusion experiments[23] that do not attempt to explain what distinguishing features the network has learned. As previously stated, these deep network papers (excluding those which include manual features) either ignore the parasternal short axis view[1,24] or include all subcategories as one class.[23] An exception is Zhang et. al, who does include the three parasternal short axis subcategories and achieved 76% accuracy for the PSAX -PM level and PSAX -MV level views (no PSAX-AP level views were in the test set).[22] While these are still not high accuracies, their improved results may be a result of a more varied dataset. For training, they extracted frames from a dataset which contained approximately four times the amount of PSAX -PM level and PSAX -MV level videos. In comparison to papers which use machine learning classifiers, our high accuracies were obtained without incorporating ECG information (unlike in Ref.[17]) and without using hand crafted features (unlike in Ref.[16]) or dictionary learning (like in Ref.[7]) that would make adaptation to other views difficult.

5. Conclusion

In conclusion, a deep learning approach was shown to classify echocardiogram views with minimal preprocessing. Accuracies for apical views (A2C, A3C, A4C) were particularly high. Grad-CAM images were used to analyze the believability of the network. These images illustrated that the network learned features similar to what an expert would use to classify the views. Unlike in previous works, Grad-CAM was used to analyze features in the lower layers, and showed that the network recognized features with anatomical significance. Misclassified apical views were typically mislabeled, of low quality, or fluctuated between views while misclassified parasternal short axis views were typically labeled as parasternal papillary muscle level views. Classification of parasternal short axis views would likely improve with more training images.

Acknowledgements

This study was partially funded by the Israel Innovation Authority KAMIN program. The authors are grateful for this support.

Conflicts of Interest Disclosure

The authors certify they have no known conflict of interest that may influence the study presented in this manuscript.

References

[1] Zhang J, Gajjala S, Agrawal P, et al. A Computer Vision Pipeline for Automated Determination of Cardiac Structure and Function and Detection of Disease by Two-Dimensional Echocardiography. 2017 Jun 22 [cited 2019 Mar 6]; Available from: http://arxiv.org/abs/1706.07342

[2] Azhari H. Basics of Biomedical Ultrasound for Engineers. Basics Biomed Ultrasound Eng [Internet]. 2010 Apr 9 [cited 2021 Aug 15]. Available from: https://onlinelibrary.wiley.com/doi/book/10.1002/9780470561478

[3] Henry WL, DeMaria A, Gramiak R, et al. Report of the American Society of Echocardiography Committee on Nomenclature and Standards in Two-dimensional Echocardiography. Circulation. 1980; 62(2): 212–7. PMid:7397962. https://doi.org/10.1161/01.CIR.62.2.212

[4] Mohamed AA, Arifi AA, Omran A. The basics of echocardiography. J Saudi Hear Assoc. 2010 Apr; 22(2): 71–6. PMid:23960599. https://doi.org/10.1016/j.jsha.2010.02.011

[5] BF W, CP T, JD S, et al. Tomographic views of normal and abnormal hearts: the anatomic basis for various cardiac imaging techniques. Part I. Clin Cardiol [Internet]. 1990 Nov[cited 2021 Aug 15]; 13(11): 804–12. PMid:2272138. https://doi.org/10.1002/clc.4960131111

[6] Park JH, Zhou SK, Simopoulos C, et al. Automatic cardiac view classification of echocardiogram. In: Proceedings of the IEEE International Conference on Computer Vision [Internet]. IEEE; 2007 [cited 2019 Mar 6]. p. 1–8. Available from: http://ieeexplore.ieee.org/document/4408867/

[7] Khamis H, Zurakhov G, Azar V, et al. Automatic apical view classification of echocardiograms using a discriminative learning dictionary. Med Image Anal [Internet]. 2017 Feb [cited 2019 Mar 6]; 36: 15–21.

PMid:27816858. https://doi.org/10.1016/j.media.2016.10.007

[8] Pak M, Kim S. A review of deep learning in image recognition. Proc 2017 4th Int Conf Comput Appl Inf Process Technol CAIPT 2017. 2018: 1–3.

[9] Sharma N, Jain V, Mishra A. An Analysis Of Convolutional Neural Networks For Image Classification. Procedia Comput Sci. 2018 Jan 1; 132: 377–84. https://doi.org/10.1016/j.procs.2018.05.198

[10] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013; 35(8): 1798–828. PMid:23787338. https://doi.org/10.1109/TPAMI.2013.50

[11] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition [Internet]. [cited 2021 Jan 18]. Available from: http://image-net.org/challenges/LSVRC/2015/

[12] Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data 2016 31 [Internet]. 2016 May 28 [cited 2021 Aug 15]; 3(1): 1–40. Available from: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0043-6

[13] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. In Institute of Electrical and Electronics Engineers (IEEE); 2010. p. 248–55.

[14] Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Int J Comput Vis [Internet]. 2016 Oct [cited 2021 Jan 18]; 128(2): 336–59. https://doi.org/10.1007/s11263-019-01228-7

[15] Ebadollahi S, Chang SF, Wu H. Automatic view recognition in echocardiogram videos using parts-based representation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2004.

[16] Otey M, Bi J, Krishna S, et al. Automatic view recognition for cardiac ultrasound images. In: MICCAI: International Workshop on Computer Vision for Intravascular and Intracardiac Imaging [Internet]. 2006 [cited 2019 Mar 6]. p. 187–94. Available from: http://www.engr.uconn.edu/~jinbo/doc/MICCAIworkshopCVII.pdf

[17] Kumar R, Fei Wang, Beymer D, et al. Echocardiogram view classification using edge filtered scale-invariant motion features. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition [Internet]. IEEE. 2010 [cited 2019 Mar 6]. p. 723–30. Available from: http://ieeexplore.ieee.org/document/5206838/

[18] Agarwal D, Shriram KS, Subramanian N. Automatic view classification of echocardiograms using Histogram of Oriented Gradients. In: Proceedings - International Symposium on Biomedical Imaging. 2013. p. 1368–71.

[19] Wu H, Bowers DM, Huynh TT, et al. Echocardiogram view classification using low-level features. In: Proceedings - International Symposium on Biomedical Imaging [Internet]. IEEE; 2013 [cited 2019 Mar 18]. p. 752–5. Available from: http://ieeexplore.ieee.org/document/6556584/

[20] Qian Y, Wang L, Wang C, et al. The synergy of 3D SIFT and sparse codes for classification of viewpoints from echocardiogram videos. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) [Internet]. Springer, Berlin, Heidelberg; 2013 [cited 2019 Mar 18]. p. 68–79. https://doi.org/10.1007/978-3-642-36678-9_7

[21] Penatti OAB, Werneck R de O, de Almeida WR, et al. Mid-level image representations for real-time heart view plane classification of echocardiograms. Comput Biol Med [Internet]. 2015 Nov 1 [cited 2019 Mar 18];66:66–81.PMid:26386547. https://doi.org/10.1016/j.compbiomed.2015.08.004

[22] Zhang J, Gajjala S, Agrawal P, et al. Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy. Circulation. 2018; 138(16): 1623–35. PMid:30354459. https://doi.org/10.1161/CIRCULATIONAHA.118.034338

[23] Madani A, Arnaout R, Mofrad M, et al. Fast and accurate view classification of echocardiograms using deep learning. npj Digit Med [Internet]. 2018 Dec[cited 2019 Mar 6]; 1(1): 6. Available from: http://www.nature.com/articles/s41746-017-0013-1

[24] Østvik A, Smistad E, Aase SA, et al. Real-Time Standard View Classification in Transthoracic Echocardiography Using Convolutional Neural Networks. Ultrasound Med Biol [Internet]. 2019 Feb [cited 2019 Mar 6]; 45(2): 374–84. PMid:30470606. https://doi.org/10.1016/j.ultrasmedbio.2018.07.024

[25] Gao X, Li W, Loomes M, et al. A fused deep learning architecture for viewpoint classification of echocardiography. Inf Fusion [Internet]. 2017 Jul [cited 2019 Mar 6]; 36: 103–13. https://doi.org/10.1016/j.inffus.2016.11.007

[26] MATLAB and Statistics Toolbox, Natick, Massachusetts, United States: The MathWorks, Inc.

[27] Kingma DP, Ba JL. Adam. A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings [Internet]. International Conference on Learning Representations, ICLR; 2015 [cited 2021 Jan 18]. Available from: https://arxiv.org/abs/1412.6980v9

[28] torchvision.transforms - PyTorch master documentation [Internet]. [cited 2018 Oct 18]. Available from: https://pytorch.org/docs/stable/torchvision/transforms.html

[29] Important Pytorch Stuff [Internet]. [cited 2018 Oct 18]. Available from: https://spandan-madan.github.io/A-Collection-of-important-tasks-in-pytorch/

[30] Nakashima K. grad-cam-pytorch [Internet]. GitHub repository; 2017. Available from: https://github.com/kazuto1011/grad-cam-pytorch/blob/master/grad_cam.py

[31] Ozbulak U. PyTorch CNN Visualizations [Internet]. GitHub repository. 2019. Available from: https://github.com/utkuozbulak/pytorch-cnn-visualizations

[32] NVIDIA Tesla K20m Specs | TechPowerUp GPU Database [Internet]. [cited 2018 Oct 18]. Available from: https://www.techpowerup.com/gpu-specs/tesla-k20m.c2029

[33] Aschkenasy SV, Jansen C, Osterwalder R, et al. Unsupervised image classification of medical ultrasound data by multiresolution elastic registration. Ultrasound Med Biol. 2006 Jul 1; 32(7): 1047–54. PMid:16829318. https://doi.org/10.1016/j.ultrasmedbio.2006.03.010

[34] Roy A, Sural S, Mukherjee J, et al. State-based modeling and object extraction from echocardiogram video. IEEE Trans Inf Technol Biomed [Internet]. 2008 May [cited 2019 Mar 6]; 12(3): 366–76. PMid:18693504. https://doi.org/10.1109/TITB.2007.910352

[35] Snare SR, Aase SA, Mjølstad OC, et al. Automatic real-time view detection. In: Proceedings - IEEE Ultrasonics Symposium [Internet]. IEEE; 2009 [cited 2019 Mar 6]. p. 2304–7. Available from: http://ieeexplore.ieee.org/document/5441530/