Echo Doppler Flow Classification and Goodness Assessment with Convolutional Neural Networks

1st Ghada Zamzmi
National Library of Medicine, LHNCBC
National Institutes of Health
Bethesda, MD, USA
alzamzmiga@mail.nih.gov

2nd Li-Yueh Hsu
National Heart, Lung, and Blood Institute
National Institutes of Health
Bethesda, MD, USA
lyhsu@nhlbi.nih.gov

3rd Wen Li
National Heart, Lung, and Blood Institute
National Institutes of Health
Bethesda, MD, USA
liwh@nhlbi.nih.gov

4th Vandana Sachdev
Heart, Lung, and Blood Institute
National Institutes of Health
Bethesda, MD, USA
sachdevv@nhlbi.nih.gov

5th Sameer Antani
National Library of Medicine, LHNCBC
National Institutes of Health
Bethesda, MD, USA
santani@mail.nih.gov

***Abstract*—Doppler Echocardiography is critical for measuring abnormal cardiac function and diagnosing valvular stenosis and regurgitation. The current practice for assessing and interpreting Doppler echo images is time-consuming and depends highly on the experience of the operator. The limitations of this practice can be mitigated using fully automated intelligent systems. Essential first steps toward comprehensive computer-assisted Doppler echocardiographic interpretation include automatic classification into view/flow categories and goodness assessment of these flows. In this paper, we propose a deep learning-based method for Doppler flow classification and goodness assessment. The method has been trained on labeled images representing a wide range of real-world clinical variation. Our method, when evaluated on unseen data, achieved overall accuracies of 91.6% and 88.9% for flow classification and goodness assessment, respectively. While further research is needed, these results are encouraging and prove the feasibility of using fully automated intelligent systems for analyzing and interpreting Doppler echo images.**

***Index Terms*—Echocardiography, Image Classification, Assessment, Deep Learning, Point-of-Care Ultrasound.**

I. Introduction

Cardiovascular disease (CVD) is the leading cause of mortality in the United States [1]. CVD can be diagnosed using several imaging techniques, such as echocardiography (echo), cardiac resonance imaging (CMR), and computed tomography (CT). Of these techniques, echo is the most commonly used as it is non-invasive, portable, inexpensive, and widely available [2]. Transthoracic echocardiogram (TTE), a very safe and common type of echocardiogram [2], involves using a transducer to transmit high frequency (2-12 MHz) ultrasound waves to the heart and converting the reflected waves (echoes) into images. Different imaging modes, including M-mode, B-mode, and Doppler, can be measured using TTE with different acquisition angles and configurations.

Spectral Doppler echocardiography is essential to cardiology. It uses the frequency shift in reflected waves to visualize the blood flow as a graph that shows the velocity of blood flow (Y axis) over time (X axis). It is routinely performed using Pulsed Wave Doppler (PWD) or Continuous Wave Doppler (CWD). PWD utilizes a single transducer element to send and receive an ultrasound wave. By sending and receiving pulses, PW Doppler has the ability to measure the velocity of flow at a specific cardiac region (a.k.a., sample volume, see Figure 1). PWD Doppler is a powerful method for providing site-specific information. A variant of PWD is the Tissue Doppler (TD), which allows to measure the velocity of myocardial tissue movements. A major limitation of PWD is its inability to display high velocities accurately [3]. CWD, on the other hand, can accurately measure high blood velocities using two dedicated transducer elements for continuously sending and receiving ultrasound waves. CWD signal represents the sum of all signals from moving objects (not site-specific) [3].

Doppler indices (e.g., peak velocity) play a critical role in measuring abnormal cardiac function as well as diagnosing valvular stenosis and regurgitation. Prior to computing indices, the view of the recorded images (e.g., PW or CW) as well as the type of the acquired blood flow (e.g., Mitral Valve [MV] flow) need to be determined. In addition, unacceptable or unmeasurable flows need to be excluded from further analysis. For example, the images shown in the second column of Figure 2 should be excluded from further analysis since the peaks are ambiguous or not measurable. The manual determination of view/flow types and the assessment of image goodness suffers from intra- and inter-observer variability [4]. Further, it is time-consuming and requires expertise as some flows differ subtly from each other. Cardiological expertise is a heavily burdened resource and often unavailable in low-resource settings. To assist echocardiographers and allow the efficient utilization of echocardiography in low-resource settings, we propose to exploit supervised deep learning methods for automatically detecting different blood flows and assessing their quality.

978-1-7281-4550-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMLA.2019.00283

A. Overview of Existing Works

Existing automated methods for analyzing medical images are broadly divided into handcrafted-based and deep learning-based methods. Handcrafted methods depend highly on human expertise to manually design and select relevant features from the image. The challenge of manually designing handcrafted descriptors and extracting the best set of features has motivated researchers to use deep learning-based methods for medical image analysis. These methods can learn and extract relevant features, at multiple levels of abstraction, directly from the source data or images. They outperformed the traditional handcrafted methods in many clinical applications. In this section, we briefly review existing deep learning-based methods for classifying Doppler echo images and assessing their quality.

1) Doppler Image Classification: This task can be defined as detecting the view (e.g., CW or PW) or the flow type (e.g., Tricuspid Regurgitation [TR]) of the acquired Doppler images. Madani et al. [5] are the first to propose a deep learning-based method for view classication of Doppler images. Specifically, the proposed method was used to distinguish 15 different echo views: 12 views from B-mode (e.g., PLAX and A4C), m-mode view, and two Doppler views (CW and PW). The proposed network, which was inspired from the well-known VGG Convolutional Neural Network [6], consists of six convolutional layers (3×3) followed by max-pooling layers (2×2) and two fully-connected layers with 1028 and 512 nodes, respectively. The final layer performs classification using Softmax function with 15 nodes. The network was trained using RMSprop optimization over 45 epochs. CW and PW Doppler had overall test accuracies of 98% and 83%, respectively. We are not aware of any automated method for blood flow (e.g., TR or MV) classification.

2) Flow Quality Assessment: As the acquisition of echocardiography images is not automated, the quality of the acquired Doppler images depends highly on the technician's knowledge, expertise, and other settings. Different automatic methods [7]–[10] were recently proposed to measure the quality of acquired 2D (B-mode) echo videos/images in real-time. These methods can aid during data acquisition by providing real-time feedback and automatically rejecting low-quality images. We are not aware of any automated method that assesses the quality or measure the goodness of the acquired Doppler flows.

B. Contributions and Roadmap

In this paper, we investigate the use of Convolutional Neural Networks (CNNs) for Doppler flows classification and assessment. The main contributions of the paper can be summarized as follows:

- We proposed a deep learning-based method for classifying three types of blood flows: Tricuspid Regurgitation (TR), Mitral Valve (MV), and Mitral Annular (MA). Figure 1 shows TR, MV, and MA flows obtained using CW, PW, and Tissue Doppler modes, respectively.
- We assessed the quality of these flows using a deep learning-based method. Specifically, we classified each flow category into low- or good-quality (e.g., low- or good-quality TR). Figure 2 presents examples of good- and low-quality cases.
- We used randomly selected real-world echocardiograms recorded from normal patients and patients with different pathologies. The data were acquired, under different configuration, using multiple vendors. Such setting ensures that our deep learning models would be clinically relevant.

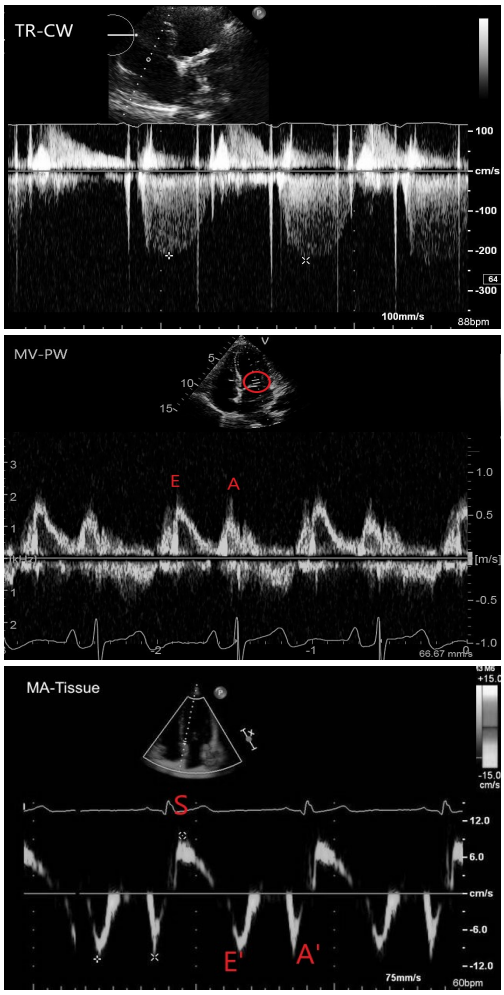

Fig. 1. Examples of different Doppler flows. First row: TR profile with a peak velocity recorded in CW Doppler mode. Second row: MV measured by PW Doppler with a sample volume (red circle in the anatomical region) parallel to the direction of the flow; MV has two triangles with two peaks corresponding to E velocity (early diastole) and A velocity (late diastole). Third row: MA flow recorded in Tissue Doppler mode. This flow has three velocities, S' (systolic velocity), E' (early diastolic velocity), and A' (late diastolic velocity). The thin curve above (TR-CW) and below (MV-PW and MA-Tissue) the Doppler region is the electrocardiogram (ECG) signal.

Section II presents the dataset we utilized to build our models. Section III describes our method for flow classification and assessment. Section IV contains the experimental results and discussion of these results. Finally, Section V concludes the paper and lists several directions for future research.

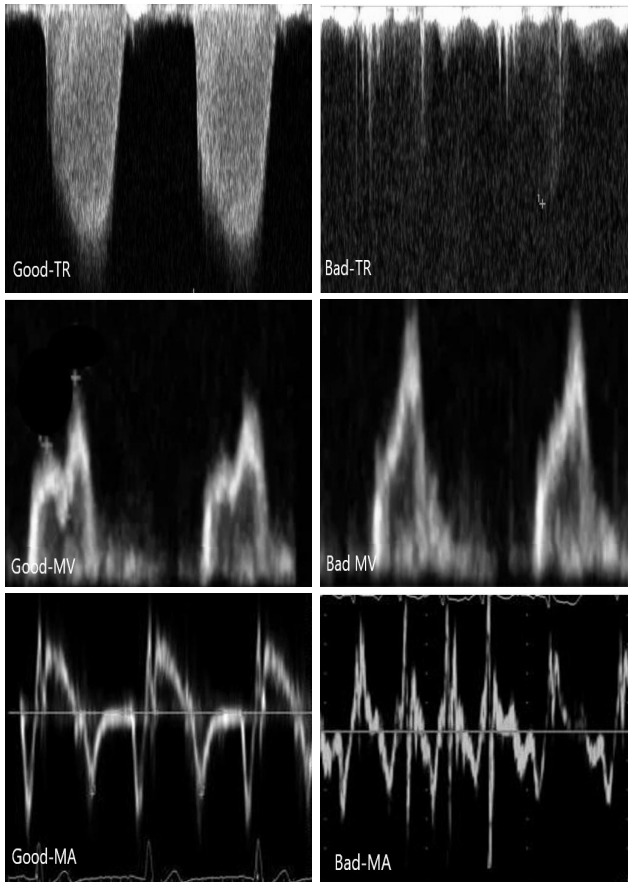

Fig. 2. Examples of good- and low-quality cases for TR, MV, and MA flows. In the first column, the peaks are clear and measurable. In the second column, the peaks are not measurable; i.e., unclear spectral envelope (TR), overlapping between two peaks (MV), or ambiguity (MA).

II. Data Collection

Doppler echo images were collected from 100 patients, who were referred for echocardiographic examination in the echo lab at the Clinical Center of the National Institutes of Health (NIH), USA. This project was reviewed by the NIH Office of Human Subjects Research Protections (OHSRP, ID#18-NHLBI-00686). De-identified images were used, and this project was determined to be "not human subjects research". The Doppler traces of the mitral valve (PW), mitral annular (Tissue), and tricuspid regurgitation (CW) were acquired using commercially available echocardiography systems including the Phillips iE33, GE Vivid95, and GE Vivid9. Each of the acquired Doppler images has two labels: a flow type label (TR, MV, or MA) and a quality label (low- or good-quality). These labels were provided by an expert technician and further verified by an expert cardiologist.

III. Methodology

This section provides a description of our method, which depends on VGG-16 and ResNet-50 architectures, for classifying different Doppler flows and assessing their quality.

As is well known, training CNNs requires large and well-annotated datasets (e.g., ImageNet - approx. 1.2 million images and 1000 classes). In practice, it is restively rare, especially in the medical domain, to access large and well-annotated datasets. Therefore, transfer learning technique was introduced to handle the lack of data. This technique allows to use the state-of-the-art CNN architectures, which were trained from scratch using significantly large datasets, as initialization or starting point.

VGG-16 [6], trained using ImageNet dataset, is one of the state-of-the-art CNNs. This network passes an input RGB image with size 224×224 to a stack of 13 convolutional layers, each uses a small filter of size 3×3 with 1 stride and 1 padding. Five max pooling with 2×2 window and stride 2 are used after each block of the convolutional layers. The stack of convolutional and max pooling layers is followed by three fully connected layers and a Softmax layer. The fully connected layers have 4096, 4096, and 1000 units, respectively. The number of units in the last layer corresponds to the number of classes in ImageNet dataset (1000 classes). All the hidden layers are equipped with ReLU function. VGG-16 is trained using 138 million parameters [6].

We fine-tuned VGG-16 using images from our dataset; i.e., we used ImageNet trained weights in the lower layers and only changed the upper layers parameters using Doppler images. We changed the layer parameters of VGG-16 and added dropout of 0.5 after each of the fully connected layers to reduce over-fitting. The final layer performs classification using Softmax function with 3 (classification) or 6 (assessment) nodes. We trained the network over 100 epochs with early stopping. For optimization, we used the ADAM optimizer with an initial learning rate of 1×10^{-3} and mini-batch size of 32. For regularization, we applied a weight decay of 1×10^{4}. We used 10-fold cross-validation to randomly vary which images were in the training and validation sets. The plots of training and validation loss by epoch confirmed that the model was not over-fitting.

ResNet-50 [11] is another state-of-the-art CNNs. The network takes as input 224×224 RGB image and passes the given image to a 7×7, 64 convolutional layers with stride 2 followed by 3×3 max pooling. The output is then sent to a stack of 48 convolutional layers distributed over four blocks. Each block starts with a convolutional layer that has a filter size of 1×1 followed by a convolutional layer that has a filter size of 3×3 and ends with a convolutional layer that has 1×1 filter size. The stack of the convolutional layers (the blocks) is followed by average pooling layer, a fully connected layer with 1000 units (1000 classes), and a Softmax layer. ResNet-50 was trained using augmented images (scale and color augmentation) from ImageNet dataset with 25.6 million parameters [11].

We fine-tuned ResNet-50 using images from our dataset as follows. We used global average pooling to obtain the base model output from the lower layers of ResNet-50. We also added a dropout of 0.5 after the global average pooling to reduce over-fitting and changed the number of classes in

the last classification layer. We trained the network over 100 epochs using mini-batch size of 32 and cyclical learning rate [12] that ranges from 1×10^{-3} to 0.1. We used 10-fold cross-validation to randomly vary which images were in the training and validation sets. The plots of training and validation loss by epoch conrmed that the model was not over-fitting.

As our training set (images of 70 patients) is relatively small and unbalanced, we rotated each image by 10 degrees followed by ipping the rotated image. It has been verified, by an expert, that the applied rotation/flipping does not degrade the clinical quality of the image. This process enlarges the set and increases the number of images in the minority class. All implementations (augmentation, training, and evaluation) were done using the PyTorch library [13].

IV. Experimental Results and Discussion

In this section, we describe the pre-processing and model evaluation used to produce the results. We then present the performance of flow classification model and goodness assessment model. The first model classifies Doppler images into TR, MV, or MA. The second model assesses the goodness of flows by classifying them into low- or good-quality.

A. Pre-processing

Prior to the application of our models, we localized the Doppler region from the raw images using a deep learning-based method. Specifically, we adopted and fine-tuned a well-known deep learning detector, known as Faster R-CNN [14], to locate the Doppler, ECG, and anatomical image regions in echo images with different levels of noise, flow types, and shape variations. We fine-tuned the upper layers of Faster R-CNN and used the weights of the network in the lower layers. We fine-tuned the upper layers of Faster R-CNN using Adam optimizer with a learning rate of 1×10^{-3} and batch size of 50, and modified the last layer of the original architecture to handle 3 ROIs. Our localization model achieved 0.97, 0.88, and 0.96 precision in detecting Doppler, ECG, and anatomical regions, respectively. We used the detected Doppler regions as input to the classification model.

The localization of ECG region was used to divide the Doppler region into individual beats as follows. We applied edge detection-based method to delineate the ECG signal curve from the detected ECG region followed by finding the start and end points of the cardiac cycle. These points are used to segment the Doppler images into individual beats ($>$2000) as shown in Figure 3. All beat images were re-sized to 224×224 to accommodate with VGG-16 and ResNet-50 image size requirement.

B. Model Evaluation

We used several metrics for performance evaluation. These metrics are overall accuracy, precision, recall (sensitivity), and F1-score. We also present the confusion matrices to visualize the performance of all classiers. The performance was reported on the test set, which contains 30 patients out of the 100 patients in our dataset. The data of the remaining

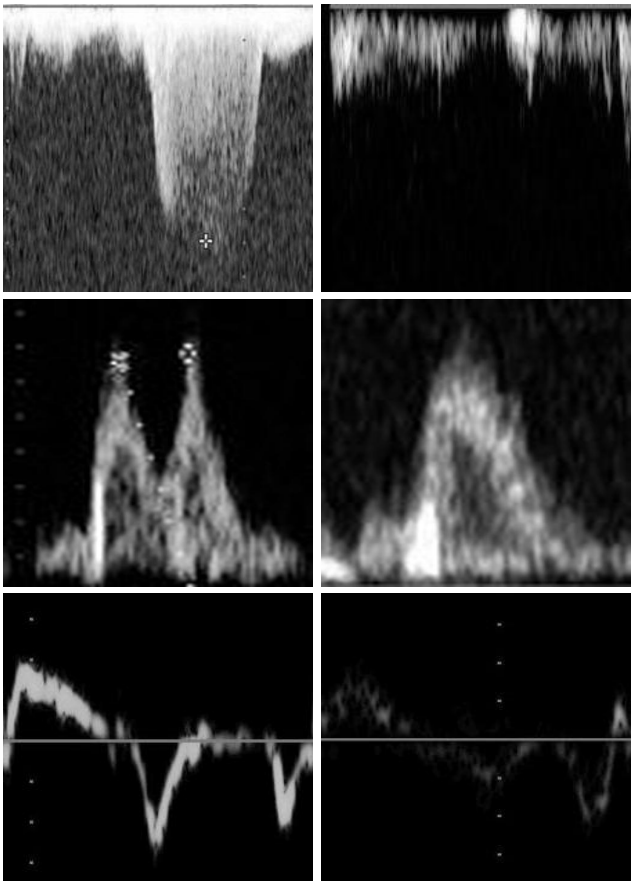

Fig. 3. Individual beats after segmentation based on ECG signal. The first column has good-quality beats for TR, MV, and MA flows, respectively. The second column has low-quality images for TR, MV, and MA flows, respectively.

70 patients were used to build the models. The training set was further divided into training and validation using 10-fold cross-validation.

C. Doppler Flow Classification

Our flow classification model achieved 91.6% overall accuracy. Table I and Table II present the performance of the model. Table I shows the precision, recall, and F1-score of all three classes using VGG-16 and ResNet-50. As can be seen, TR class has the highest performance. VGG-16 achieved better than ResNet-50 in most cases. The confusion matrix of VGG-16 is presented in Table II. As shown in the matrix, TR class has the highest correct classification rate (true positive) while MV has the lowest rate. This can be attributed to two reasons. First, the number of samples of MV in the original dataset is smaller than TR class. Our experimental results showed that augmenting the minority class images to obtain a relatively balanced dataset improve the performance. However, we believe our model is biased to predict TR class as it has the highest number of original data. Second, MV images with bad shape (low-MV) can be similar to the TR images as shown

TABLE I
PERFORMANCE OF FLOW CLASSIFICATION FROM DOPPLER ECHO USING VGG-16 AND RESNET-50

	VGG-16			ResNet-50		
	Precision	Recall	F1-score	Precision	Recall	F1-score
TR-CW	0.91	0.99	0.95	0.92	1.0	0.96
MV-PW	0.97	0.71	0.82	0.95	0.69	0.80
MA-Tissue	0.87	0.98	0.92	0.83	0.93	0.88

TABLE II
NORMALIZED CONFUSION MATRIX OF FLOW CLASSIFICATION (VGG-16)

	TR-CW	MV-PW	MA-Tissue
TR-CW	**0.99**	0.01	0.00
MV-PW	0.18	**0.71**	0.11
MA-Tissue	0.03	0.00	**0.97**

TABLE III
PERFORMANCE OF GOODNESS ASSESSMENT USING VGG-16 AND RESNET-50

	VGG-16			ResNet-50		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Good-quality TR	0.85	0.97	0.90	0.87	0.96	0.91
Low-quality TR	0.98	0.89	0.94	0.96	0.87	0.91
Good-quality MV	0.88	0.83	0.86	0.89	0.83	0.86
Low-quality MV	0.62	0.72	0.67	0.57	0.75	0.65
Good-quality MA	0.89	0.87	0.88	0.82	0.85	0.83
Low-quality MA	0.76	0.73	0.75	0.72	0.70	0.71

TABLE IV
NORMALIZED CONFUSION MATRIX FOR ASSESSMENT MODEL (VGG-16)

	Good-TR	Low-TR	Good-MV	Low-MV	Good-MA	Low-MA
Good-TR	**0.97**	0.02	0.01	0.00	0.00	0.00
Low-TR	0.08	**0.89**	0.01	0.02	0.00	0.00
Good-MV	0.09	0.00	**0.83**	0.05	0.01	0.02
Low-MV	0.16	0.00	0.13	**0.72**	0.00	0.00
Good-MA	0.00	0.00	0.03	0.02	**0.87**	0.08
Low-MA	0.00	0.00	0.03	0.03	0.2	**0.73**

in Figure 4. For example, the images in the first row are MV flow images that were misclassified as TR flow images. The images in the second row are MA (1^{st} column) and TR (2^{nd} column) flow images that were misclassified as TR and MV, respectively.

To improve the performance of MV class, we are planning to use other augmentation methods (e.g., GAN) and try penalization methods that bias the model to pay more attention to the minority class. We also plan to collect a larger dataset with a relatively balanced distribution. Although the results reported in this paper includes three flows, we believe our model can be easily extended to include other flows since it does not use a specific template or require prior knowledge about these flows.

D. Doppler Goodness Assessment

The goodness assessment model, which was applied to beats of different flows, achieved 88.9% overall accuracy. Table III and Table IV present the performance of the model. Table III shows the precision, recall, and F1-score of all six classes using VGG-16 and ResNet-50. As can be seen from the table, VGG-16 performs slightly better than ResNet-50 in most cases. The table also shows that TR has the highest performance for both VGG-16 and ResNet-50. Table IV presents the confusion matrix for VGG-16. As shown in the matrix, good-TR and low-TR have the highest true positive rate. We believe the performance of other classes would be improved by training the model using dataset with a relatively balanced distribution.

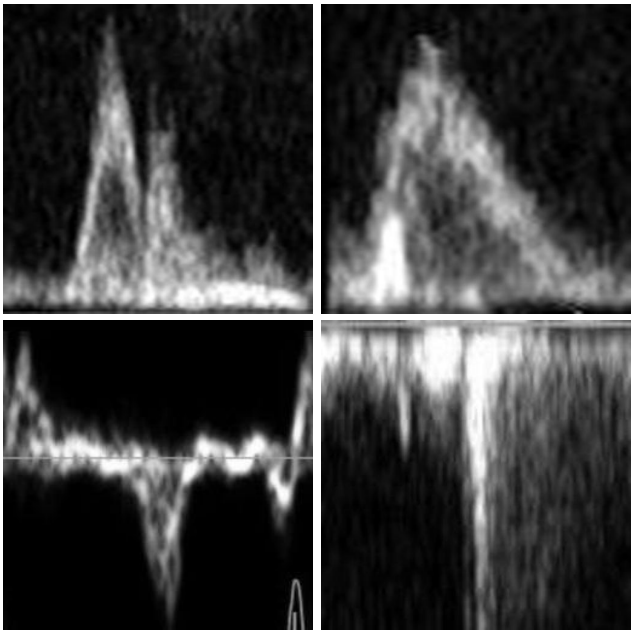

Fig. 4. Examples of misclassified cases. First row: MV flow images misclassified as TR flow images. Second row: MA (1^{st} column) flow image misclassified as TR and TR (2^{nd} column) flow image misclassified as MV.

These results are encouraging and prove the feasibility of using deep learning methods with echo Doppler for automatic blood flow classification and goodness assessment.

V. Conclusion and Future Directions

The current practice of interpreting echo Doppler requires echocardiographers to manually exclude low-quality images from further analysis. In addition to its subjectivity, this practice is time-consuming and requires expertise. Cardiological expertise is a heavily burdened resource and often unavailable in low-resource settings. To assist echocardiographers and enhance the utilization of POCUS (Point-of-Care Ultrasound) systems, we propose deep learning-based method for classifying echo Doppler flows and assessing their quality. To the best of our knowledge, this is the first study that proposes automated approach for classifying Doppler flows and assessing their quality. The preliminary results show that the trained classification model can distinguish different blood flows with 91.6% overall accuracy. Our assessment model achieves an overall accuracy of 88.9%. Since the proposed method does not use a specific template or require prior knowledge about the flows, we anticipate that it can be easily extended to include other Doppler flows. These preliminary results are encouraging and prove the feasibility of using fully automated methods for Doppler flow classification and goodness assessment.

Ongoing works include integrating other Doppler flows to the proposed method. We also plan to evaluate the proposed method using a larger dataset with a relatively balanced distribution. In addition, we plan to explore different augmentation techniques (e.g., elastic augmentation [15] or GAN [16]) and larger images (current image size is 224×224). Another important future direction would be to use quantitative scoring system, similar to [10], for assessing the acceptability or goodness of each flow instead of having two (low-quality or good-quality) labels. Finally, we plan to use several visualization techniques to show the important human-recognizable clinical features within images.

References

[1] E. J. Benjamin, P. Muntner, and M. S. Bittencourt, "Heart disease and stroke statistics-2019 update: A report from the american heart association," *Circulation*, vol. 139, no. 10, pp. e56–e528, 2019.

[2] B. F. Byrd III, T. P. Abraham, D. B. Buxton, A. V. Coletta, J. H. Cooper, P. S. Douglas, L. D. Gillam, S. A. Goldstein, T. R. Graf, K. D. Horton, *et al.*, "A summary of the american society of echocardiography foundation value-based healthcare: summit 2014: the role of cardiovascular ultrasound in the new paradigm," *Journal of the American Society of Echocardiography*, vol. 28, no. 7, pp. 755–769, 2015.

[3] C. Mitchell, P. S. Rahko, L. A. Blauwet, B. Canaday, J. A. Finstuen, M. C. Foster, K. Horton, K. O. Ogunyankin, R. A. Palma, and E. J. Velazquez, "Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: Recommendations from the american society of echocardiography," *Journal of the American Society of Echocardiography*, vol. 32, no. 1, pp. 1–64, 2019.

[4] D. Caivano, M. Rishniw, V. Patata, M. Giorgi, F. Birettoni, and F. Porciello, "Left atrial deformation and phasic function determined by 2-dimensional speckle tracking echocardiography in healthy dogs," *Journal of Veterinary Cardiology*, vol. 18, no. 2, pp. 146–155, 2016.

[5] A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, "Fast and accurate view classification of echocardiograms using deep learning," *NPJ digital medicine*, vol. 1, no. 1, p. 6, 2018.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] S. R. Snare, H. Torp, F. Orderud, and B. O. Haugen, "Real-time scan assistant for echocardiography," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, vol. 59, no. 3, pp. 583–589, 2012.

[8] S.-K. Pavani, N. Subramanian, M. D. Gupta, P. Annangi, S. C. Govind, and B. Young, "Quality metric for parasternal long axis b-mode echocardiograms," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 478–485, Springer, 2012.

[9] A. H. Abdi, C. Luong, T. Tsang, G. Allan, S. Nouranian, J. Jue, D. Hawley, S. Fleming, K. Gin, J. Swift, *et al.*, "Automatic quality assessment of apical four-chamber echocardiograms using deep convolutional neural networks," in *Medical Imaging 2017: Image Processing*, vol. 10133, p. 101330S, International Society for Optics and Photonics, 2017.

[10] A. H. Abdi, C. Luong, T. Tsang, J. Jue, K. Gin, D. Yeung, D. Hawley, R. Rohling, and P. Abolmaesumi, "Quality assessment of echocardiographic cine using recurrent neural networks: Feasibility on five standard view planes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 302–310, Springer, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[12] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472, IEEE, 2017.

[13] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[15] E. Castro, J. S. Cardoso, and J. C. Pereira, "Elastic deformations for data augmentation in breast cancer mass detection," in *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 230–234, IEEE, 2018.

[16] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.