Enhanced Quality Assessment of Echocardiographic Images for Pulmonary Hypertension Using Convolutional Neural Networks

Parnian Sattar[1], Constance Verdonk[2], Frida Hermansson[2], Xiu Tang [2], Alison Marsden [2], Francois Haddad[2], Seraina A Dual[1]

[1] KTH Royal Institute of Technology, Stockholm, Sweden
[2] Stanford University, Palo Alto, CA, USA

Abstract

Pulmonary Hypertension (PH) is associated with cardiopulmonary disease and carries strong prognostic information. Its accurate diagnosis is based on invasive procedures such as right-sided heart catheterization. Alternative non-invasive approaches rely on Doppler imaging, where signal quality may impede correct readings potentially limiting its clinical value. In this study, we propose an automated approach for the quality assessment of Doppler signals using a convolutional neural network (CNN). The CNN was trained on echocardiographic Doppler images and their quality was assessed by 2 independent expert readers. The dataset was subjected to preprocessing and augmentation techniques to enhance model resilience and generalization. Leveraging the VGG-16 architecture, the CNN demonstrated an accuracy of 86%, sensitivity of 86%, precision of 88%, and F1-Score of 86% on the test set. The CNN showed improved accuracy, recall, and F1-score as compared to an unseen clinical reader assessment. The results emphasize the variability in clinical reader assessment, such that automated assessment may prove highly clinically useful in the future. In this way, deep learning-driven image quality assessment could enhance diagnostic accuracy, reduce practitioner variability, and streamline patient care in PH management.

1. Introduction

Pulmonary Arterial Hypertension (PAH) is a condition identified by high pressure in the pulmonary artery, affecting approximately 1% of the global population across all age groups, underscoring PH as a significant global health concern [1, 2]. Today, PH is clinically managed to reduce symptoms and improve the quality of life of affected patients. If the cause is identified and treated early, permanent damage to the pulmonary vascular bed may potentially be avoided [3].

Accurate diagnosis and classification often rely on right-sided heart catheterization (RHC), which is considered the gold standard [3]. However, the invasive nature of RHC, which involves catheter insertion into the pulmonary artery, limits its practicality[1, 4]. Therefore, there is a pressing need for noninvasive methods to reliably assess pressure, enabling prompt diagnosis [4]. Doppler echocardiography serves this purpose by non-invasively indicating PH via measurement and analysis of maximal velocity (Vmax) in tricuspid regurgitation (TR) flow signals [5] (See Figure 1).

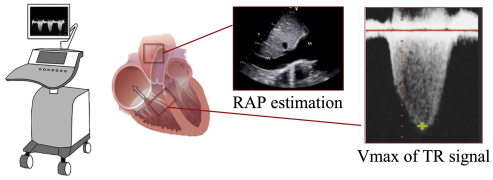

Figure 1. Doppler Echocardiography imaging. Acquired images can be used for Right atrial pressure (RAP) and maximum flow velocity (Vmax) estimation.

High user dependence of the signal quality across Doppler acquisitions, results in high variability of the reliability. In turn, this limits the clinical application of non-invasive assessment of PH using non-invasive Doppler echocardiography. Regardless of the reader's expertise, interpreting velocities from low-quality signals in Doppler imaging poses a greater risk of misdiagnosis. Hence, assessment of quality metrics before interpretation of the signal is likely to mitigate some of the user-dependency [6].

Concurrently, recent advancements in machine learning, particularly in deep learning, are making it easier to identify, classify, and quantify patterns in medical images [7]. Deep learning techniques such as Convolutional Neural Networks (CNNs) have been utilized to develop advanced models for image quality assessment [8]. Given that a meaningful interpretation of the TR flow has the potential to expedite an earlier diagnosis for PAH, machine

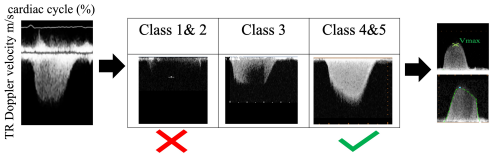

Figure 2. Detecting Maximum Velocity(Vmax). The quality of normalized TR Doppler signal (normalization based on the RR interval and maximal flow velocity) is classified across a 5-point Likert scale.

learning tools could provide meaningful support in rapid quality stratification. In this exploratory work, we propose a CNN as a quality screening tool for echocardiographic TR Doppler images and evaluate its performance against two independent expert readers.

2. Methods

2.1. Data

In this study, 371 Doppler echocardiography images from retrospectively enrolled patients from the Vera Moulton Wall Center for Pulmonary Vascular Disease obtained between 2003 and 2022 were used. The approval for this study was granted by the Institutional Review Board of Stanford University, and it was carried out under the Cardio Share protocol (IRB25673). The images were acquired by the Philips IE33 ultrasound system using Doppler continuous-wave echocardiography. For preprocessing, TR Doppler waveforms were segmented into individual waves using the ECG signal and standardized based on the RR interval based on an established pipeline [9]. The reshaped image was accompanied by the corresponding label of image quality employed for the deep learning model (see Figure 2).

According to the American College of Physicians, the grading scale for the quality of the TR signals is reported according to a 5-point Likert scale as follows: 5/4 (excellent quality with peak and transition phases well/not-optimally visualized), 3 (peak not clearly visualized but signal post-transition zones and interpolation possible clinically), 2/1 (peak not visualized and isovolumic signal seen/not seen). In our study, image classes were combined (see Figure 2) to balance the number of images per class without losing clinical relevance.

Two datasets were available. The first set was used for test and validation and includes 223 echocardiography images which classified by expert level III reader (Obs 1). We dedicate 190 images of the images (85%) for training and 33 images (15%) for validation. The test set comprises 148 images that have been classified by two readers (an independent reader(Obs 2) and the same reader as the train set(Obs 1)) with expert level III.

Acknowledging the presence of class imbalance, oversampling techniques were employed within the training dataset, focusing on increasing the representation of underrepresented classes using a weighted random sampler. Data augmentation techniques were also applied to train dataset such as random color jitter, which helps the model learn to recognize objects from different viewpoints and also increase the number of samples in the train set. Using these techniques enhances the model's resilience and generalization through exposure to a diverse and balanced training set.

Cross-validation with five folds was used to provide a more robust estimate of the model's performance compared to a single train-validation split.

2.2. Architecture

To construct our deep learning model, we employed the pre-trained VGG-16 architecture, a widely adopted convolutional neural network (CNN) renowned for its effectiveness in image classification tasks [10]. The architecture is depicted in Figure 3.

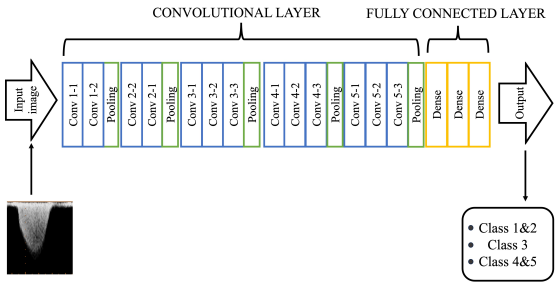

Figure 3. Architecture of the VGG-16. It takes the input image with the size of 224x224x3 pixels and then goes through 13 convolution layers and 3 fully connected layers, a ReLU and the Softmax activation function.

To train our model, we froze the convolutional layers of the pre-trained VGG-16 model which is trained on the ImageNet dataset. This decision was made to leverage the powerful feature extraction capabilities of the network while fine-tuning the model to suit the unique characteristics of our Doppler image dataset. Notably, despite freezing the convolutional layers, we left the batch normalization layers trainable because our dataset samples had distinct characteristics compared to the ImageNet dataset, on which VGG-16 was originally trained. The Cross-entropy loss function was used as the objective loss in the final layer to optimize the model for image classification. The deep learning model was implemented in PyTorch, using Python programming language.

The Adam optimizer was chosen as the optimizer for the model, with a learning rate of 0.001. This learning

rate was selected to optimize the model parameters and improve performance over time by comparing the model performance on the validation set.

2.3. Evaluation

We used the test set to evaluate our model performance on unseen data and compare against expert performance (Obs 1, Obs 2). Classification accuracy, precision, recall and F1-score of the model were evaluated in the unseen test set. The metrics are defined as follow:

$$Classification\ Accuracy = \frac{TP + TN}{\#AllSamples} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

Additionally, Cohen's kappa statistical analyses was performed. The results Cohen's kappa analyses was interpreted as described here [11] (Kappa = 1 being a perfect agreement).

3. Result

The progression of training, validation loss, and accuracy across epochs during the model training process sufficiently converged (Figure 4).

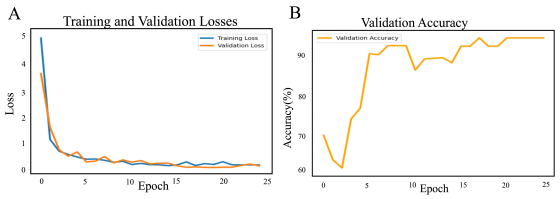

Figure 4. VGG-16 training. (**A**) changes in loss during training in both the training and validation loss. (**B**) changes in accuracy during training in validation set.

The performance of the CNN model against Obs 1 and Obs 2 was evaluated (Table 1).

The model performed well when compared to the second observer for all the metrics, when compared to the inter-reader results. This shows the accuracy 86% which was higher than the accuracy between two observers. The model performance against the second observer implies also a higher degree of reliability than the case of human observer as the precision was 88%. The area under the ROC curve (AUC) score of the model against the two observers was calculated and shows the model is quite effective at distinguishing between classes than observer 2(AUC Score = 89%) and also suggests reasonably well discrimination ability against observer 1(AUC Score = 78%).

Observer	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*Kappa*
CNN& Obs 1	0.70	0.71	0.70	0.65	0.39
CNN& Obs 2	0.86	0.88	0.86	0.86	0.57
Obs 1& Obs 2	0.71	0.87	0.71	0.76	0.42

Table 1. Accuracy, Precision, Recall, F1-Score, Kappa

The confusion matrix shows the high general performance and the variability in our two clinical observers, when assessed against our CNN model. In both cases the model predicted more images to be of high quality (class 4 and 5) than the expert readers. Observer 1 was even more critical and labeled more images to be of very low quality (class 1 and 2).

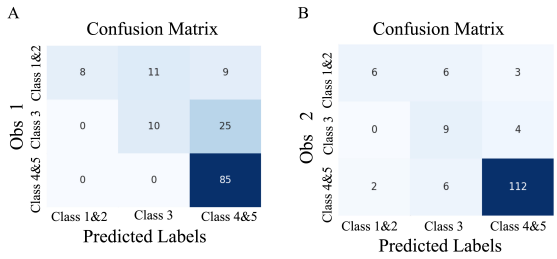

Figure 5. Confusion Matrix. (**A**) confusion matrix of observer 1 . (**B**) confusion matrix of observer 2

4. Discussion

Quality assessment in echocardiography plays a crucial role in improving the interpretation of TR images by enabling accuracy and reliable image readings. Clear and high-quality images can provide the clinician with better visualization of the TR Doppler signal, leading to more accurate interpretation with higher confidence of the pressure, indicating PH[6]. Interpolation approaches for pressure estimation have previously been attempted, but remain reliant on high quality images[9]. An automatic quality assessment could potentially improve confidence and reduce the likelihood of misinterpretations due to artifacts or technical limitations. This, in turn, enables clinicians to make more informed decisions regarding PH patient man-

agement and possible interventions.

At this stage of development, our CNN model accuracy is promising, despite low number of images. As the results show the model performed better or same for all the metrics against the second observer as compared to the inter-observer agreement. Our results suggests that we can provide clinicians with a meaningful and quick pre-assessment as to whether the signals should be discarded and the measurement repeated.

As study shows the interobserver agreement between developed model and human observer was better than comparison of both observers. This highlights the clinical need for standardized and automated pre-screening of images to maximize usage of non-invasive Doppler images for assessment of PH. The higher image quality ratings of our model with respect to signal quality (especially compared to Obs 1) may be of concern, as it may lead to usage of low quality images in clinical practice. Final clearance of the automatically labeled high quality images by a highly trained clinician will remain necessary to ensure valid assessment.

The quality assessment discussed in the study by Dong et al. [12] highlights the importance of domain specificity in evaluating image quality. Different fields may prioritize varying attributes based on their specific requirements and objectives. The study's quality assessment relied on a common 5-point Likert scale informed by expert clinical readers. Nevertheless, there's uncertainty regarding whether this criterion aligns precisely with what the network evaluates or if the network assesses other parameters. Further exploration and validation of the network's assessment criteria against established grading scales may be warranted to enhance the robustness and validity of future research in this area.

Despite the study's small number of images, the CNN model performed well which demonstrates the potential of automated quality assessment. However, further improvements are needed for clinical use. Increasing the number of images could enhance the consistency of the model and allow for increased understanding of parameters used by the model. It might be useful to consider including more observers in the training and validation data to ensure inclusion of a broad range of clinical views. Furthermore, expanding the training dataset would afford greater opportunity for fine-tuning various layers of the network and exploration of different architectures and hyperparameters.

In summary, this study shows the potential of deep learning to automatically predict quality assessment of TR Doppler echocardiography images.

References

[1] Hoeper MM, Humbert M, Souza R, Idrees M, Kawut SM, Sliwa-Hahnle K, Jing ZC, Gibbs JSR. A global view of pulmonary hypertension. The Lancet Respiratory Medicine 2016;4(4):306–322.

[2] Humbert M, Kovacs G, Hoeper MM, Badagliacca R, Berger RM, Brida M, et al. 2022 esc/ers guidelines for the diagnosis and treatment of pulmonary hypertension: Developed by the task force for the diagnosis and treatment of pulmonary hypertension of the European Society of Cardiology (esc) and the European Respiratory Society (ers). European Heart Journal 2022;43(38):3618–3731.

[3] Mandras SA, Mehta HS, Vaidya A. Pulmonary hypertension: a brief guide for clinicians. In Mayo Clinic Proceedings, volume 95. Elsevier, 2020; 1978–1988.

[4] Arkles JS, Opotowsky AR, Ojeda J, Rogers F, Liu T, Prassana V, et al. Shape of the right ventricular doppler envelope predicts hemodynamics and right heart function in pulmonary hypertension. American Journal of Respiratory and Critical Care Medicine 2011;183(2):268–276.

[5] Galiè N, Humbert M, Vachiery JL, Gibbs S, Lang I, Torbicki A, Simonneau G, et al. 2015 esc/ers guidelines for the diagnosis and treatment of pulmonary hypertension: the joint task force for the diagnosis and treatment of pulmonary hypertension of the European Society of Cardiology (esc) and the European Respiratory Society (ers). European Heart Journal 2016;37(1):67–119.

[6] Amsallem M, Sternbach JM, Adigopula S, Kobayashi Y, Vu TA, Zamanian R, et al. Addressing the controversy of estimating pulmonary arterial pressure by echocardiography. Journal of the American Society of Echocardiography 2016;29(2):93–102.

[7] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annual Review of Biomedical Engineering 2017; 19:221–248.

[8] Zhang W, Ma K, Yan J, Deng D, Wang Z. Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology 2018;30(1):36–47.

[9] Dual S, Verdonk C, Amsallem M, Pham J, Obasohan C, Nataf P, McElhinney D, Arunamata A, Kuznetsova T, Zamanian R, Feinstein J, Marsden A, Haddad F. Elucidating tricuspid doppler signal interpolation and its implication for assessing pulmonary hypertension. Pulmonary Circulation 08 2022;12.

[10] Kaur T, Gandhi TK. Automated brain image classification based on vgg-16 and transfer learning. In 2019 International Conference on Information Technology (ICIT). IEEE, 2019; 94–98.

[11] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159–174. ISSN 0006341X, 15410420.

[12] Dong J, Liu S, Liao Y, Wen H, Lei B, Li S, Wang T. A generic quality control framework for fetal ultrasound cardiac four-chamber planes. IEEE Journal of Biomedical and Health Informatics 2019;24(4):931–942.

Address for correspondence:

Seraina Dual, seraina@kth.se
Hälsovägen 10C, 14152 Huddinge, Sweden seraina@kth.se