Automated Assessment of Image Quality in 2D Echocardiography Using Deep Learning

R.B. Labs, A. Vrettos, N. Azarmehr, J. P. Howard, M. J. Shun-shin, G. D. Cole, D. P. Francis, M. Zolgharni

***Abstract*—**Echocardiography is the most used modality for assessing cardiac functions. The reliability of the echocardiographic measurements, however, depends on the quality of the images. Currently, the method of image quality assessment is a subjective process, where an echocardiography specialist visually inspects the images. An automated image quality assessment system is thus required. Here, we have reported on the feasibility of using deep learning for developing such automated quality scoring systems. A scoring system was proposed to include specific quality attributes for on-axis, contrast/gain and left ventricular (LV) foreshortening of the apical view. We prepared and used 1,039 echocardiographic patient datasets for model development and testing. Average accuracy of at least 86% was obtained with computation speed at 0.013ms per frame which indicated the feasibility for real-time deployment.

***Keywords*—**Deep learning, echocardiography, 2D image quality, quality assessment.

I. Introduction

CARDIACultrasound imaging or echocardiography is the day-to-day standard for assessing cardiac function [1]. This is mainly due to its low cost, safety, portability and non-invasive nature, and its capability to produce images to confidently detect heart abnormalities.

The impact of image quality on the reliability of echocardiographic measurements and diagnostic accuracy has been demonstrated [2], [3]. Currently, the method of image quality assessment is a subjective process, where an echocardiography specialist visually inspects and rates an image based on certain features such as wall definition and clarity of anatomical details in the image.

An objective and quantitative method for image quality assessment is a useful component for an operator guidance system, as well as a valuable tool for research and clinical practice. As part of an operator guidance system, it can provide quantitative information on the adequacy of the images obtained. It can also provide an independent measure.

A. Related Work

Image quality assessment is generally approached by defining a reference image and calculating the deviation of any given image to this reference [4]. However, in echocardiography, this method is not practical, since images vary significantly from patient to patient, and it is difficult to define an image with perfect quality. Therefore, it is necessary to develop a blind image quality assessment algorithm, which does not depend on a reference image. Studies have been carried out on blind image quality assessment [5]-[7]; largely focusing on the distortion of images due to compression, with some implementing machine learning algorithms using edge sharpness and random/structural noise level to evaluate image quality. This approach is difficult to apply to echocardiography because cardiac ultrasound does not present well defined edges due to two facts: 2D cardiac images are formed by interference pattern of scattering centres presenting an inherent poor resolution; and anatomical features do not present crisp edges because the endocardium is trabeculated-there are papillary muscles, the external purkinje network. Also, epicardium does not present as a crisp boundary, either because it is joined to the myocardium on one side, and to the pericardium in several layers, including pericardial fluid, on the other. So there exist a relatively subtle acoustic impedance transitions next to larger ones. Hence, new measures of image quality need to be developed and tested based on the global properties of the Echocardiographic images.

Following the recent success of deep Convolutional Neural Networks (CNN) in computer vision tasks, there have been a few reports on the application of deep learning for Echocardiographic image quality assessment.

Abdi et al. [8] investigated the feasibility of using CNN to assess the quality of apical four-chamber (A4C) echocardiographic images. Later, they expanded that work by proposing a framework for optimising the deep learning architecture to generate an automatic echo score [9]. Their model incorporated a regression model, based on hierarchical features extracted automatically from echocardiographic images. Abdi et al. [8] reported an average mean (absolute) error of 0.71±0.58 between the network score and expert's manual scores.

In a more recent study [10], the authors presented a deep learning model based on Recurrent Neural Networks (RNN) to realise quality assessment of five standard echocardiography views. A mean quality score accuracy of 85% compared to the manual score assigned by experienced Echosonographers was achieved.

Dong et al. [11] proposed a generic deep learning framework for quality control of fetal cardiac four-chamber views, consisting of three CNN-based networks, used to

R. B. Labs is with the School of Computing and Engineering, University of West London, London, UK (corresponding author, e-mail: robbie.labs@ uwl.ac.uk).

A. Vrettos is with Imperial College Healthcare NHS Trust, London, UK.

N. Azarmehr, is with the School of Computer Science, University of Lincoln, Lincoln, UK

J. P. Howard, M.J. Shun-shin, G.D. Cole, and D. P. Francis are with Imperial College, London, UK

M. Zolgharni is with School of Computing and Engineering, University of West London, London, UK

perform rough classification, classification refinement, and anatomical detection, respectively. They conducted experiments on a fetal ultrasound cardiac dataset, and reported a highest mean average precision of 93.52%.

B. Main Contributions

Interpreting the results of the proposed architectures in the literature is not straightforward. This is because a direct comparison of the models' performance would require access to the same patient dataset. At present, no echocardiography dataset and the corresponding annotations for the image quality assessment is publicly available. We, therefore, aimed at evaluating the performance of deep learning models for the automated image quality assessment using an independent echocardiography dataset which would be made available at website [13]

Although the inference time reported in the previous studies reviewed in Section I *A* was short enough the make it feasible for real-time applications, the utility of such systems in the clinical practice would be limited. This is because only an overall predicted image quality score is provided by the models. If employed as part of an operator guidance system, the operator is provided with no clues as to why the image is being tagged as low quality, and how to improve it to obtain optimal images. A practical quality control report should contain such information.

In the view of the above, the main contributions of this research can be summarized as follows:

- Preparation and annotation of an independent patient dataset of 2D Echocardiographic images for Fore-shortening, Contrast/Gain and Axial Target.
- Deep learning pipeline to compute quality measure from2D+t echocardiography sequences
- Custom-made program which closely replicated the interface of echo hardware
- Quality evaluation using 3 proposed quality attributes
- Quality scores representation using symbolic score to depict high quality, average quality and high-quality score values.

II. METHODS

In this section, a brief account of the patient dataset is provided, followed by the deception of expert annotation process. Details of the neural network model, training parameters, and evaluation metrics are then provided.

A. Dataset

The study population consisted of a random sample of 1,039 EchocardiographicA4C studies from patients with age ranges from 17 and 85 years, who were recruited from patients who had undergone echocardiography with Imperial College Healthcare NHS Trust.

The acquisition of the images had been completed by experienced Echocardiographers using ultrasound equipment from GE and Philips manufacturers according to the standard protocols.

Ethical approval was obtained from the Health Regulatory Agency (Integrated Research Application System identifier 243023). Patient automated anonymisation was performed to remove the patient-identifiable information.

DICOM-formatted videos were then split into constituent frames, and 25 sequence frames were extracted from each echo cine loop while each frame bearing the same clinical score as its respective original echo cine score from each video to represent arbitrary stages of the heart cycle, resulting in 25,975 frames. The dataset is randomly split into training (15,585 frames), validation (5,195 frames) and testing (5,195 frames) sub-datasets in a 60:20:20 ratio.

B. Quality Scoring Method

To obtain the gold-standard (ground-truth), the videos were manually annotated by two professionals i.e. one accredited and an experienced cardiology expert - i.e. Atrioventricular (AV), giving three quality scores for each quality specific model as depicted in Table I.

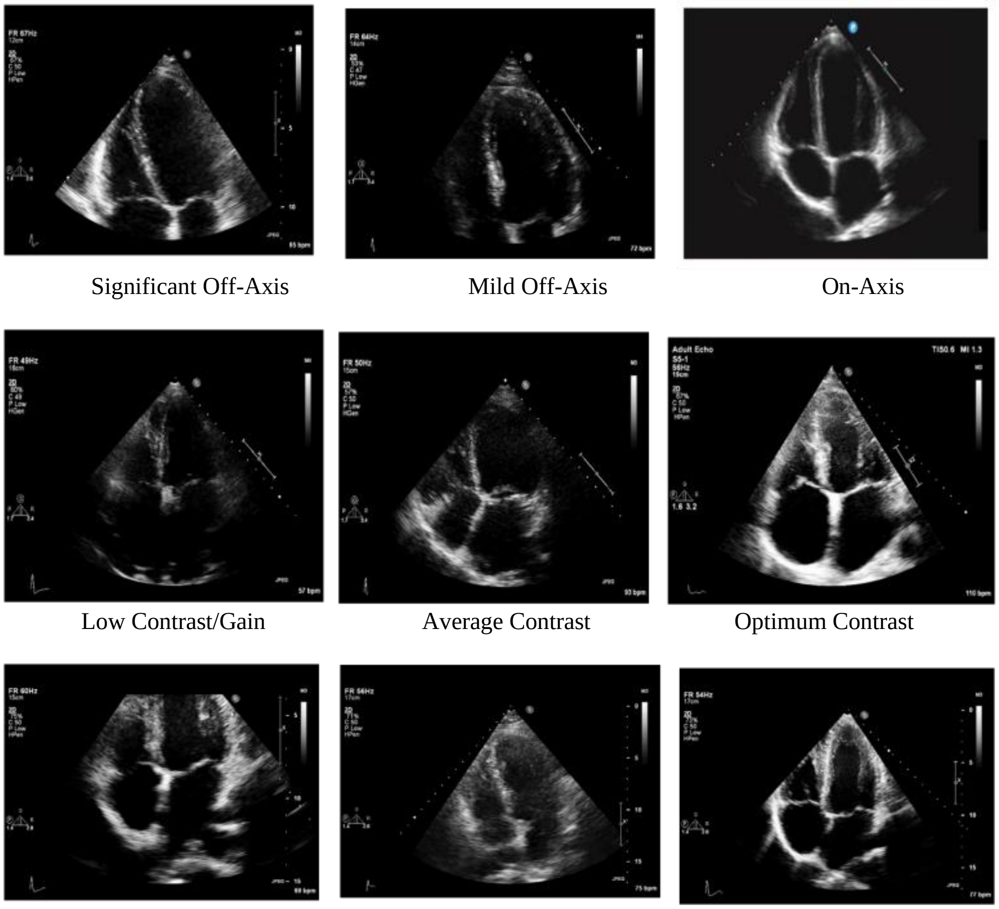
TABLE I
MANUAL SCORES CRITERIA FOR QUALITY-ATTRIBUTE MODELS

Manual Score/Model	**0**	**1**	**2**
On-Axis Target	Significant Off-Axis	Mildly Off-Axis	On-Axis
Contrast/Gain	Poor Contrast/Gain	Average Contrast/Gain	Optimum Contrast/Gain
Left Ventricle (LV) Fore-Shortening	Significant Fore-Shortening	Mild Fore-Shortening	No Fore-Shortening

We developed a custom-made program which closely replicated the interface of echo hardware. The expert visually inspected the cine loops by controlled animation of the loops using arrow keys. Fig. 1 shows examples of quality-specific samples used in the study.

Rather than obtaining an overall quality score for the image from a weighted average of these quality measures, we instead used each component separately hereby proposing an approach using 3 quality attributes stated in Table I.

Since quality components may have different maximum scores, for the sake of a fair comparison, all attributes measures were normalised to one. The cardiologist's annotations of the videos were used as the quality score for all constituent frames of that video and were used as the ground truth (Q_{GT}) for the model developments.



Significant Off-Axis

Mild Off-Axis

On-Axis

Low Contrast/Gain

Average Contrast

Optimum Contrast

LV Significant Shortening

LV Mild Shortening

LV No-Shortening

Fig 1. Examples of images for 3 proposed quality attributes.

2.*3 Model Architecture*

Inspired by Abdi et al. [10] proposed model, the architecture used is based on multivariate regression hybrid model with multiple branches, each branch models specific quality attribute proposed in section 2.2.

The architecture is logically divided into two parts; first part allows weight sharing across convolutional layers while extracting the hierarchical feature in the image sequence. This is a 3-layer CNN architecture which accepts fixed length sequence frames of spatial size 227 by 227. The frame sequence convolved with the shared layers and the resultant feature map is flattened to feed the multivariate hybrid network using Rectifier Linear Units (ReLUs) activation function.

The second part features multivariate, 3-branch hybrid architecture consist of 2-layer CNN and Long Short-Term Memory (LSTM) layer [12]. This hybrid model extracts temporal features for the specific quality attributes defined in Table I and makes predictions for quality scores on each attribute.

The model was trained simultaneously on three specific quality attributes using MAE (L1) cost function and stochastic gradient descent as optimizer. Since LSTM uses sigmoid for its internal gating, this was preserved to provide boundary for normalised output scores on each model output. Hence, specific quality attribute is therefore extracted, and network score is computed for each quality attributes per frame. The architecture is depicted in Fig. 2.

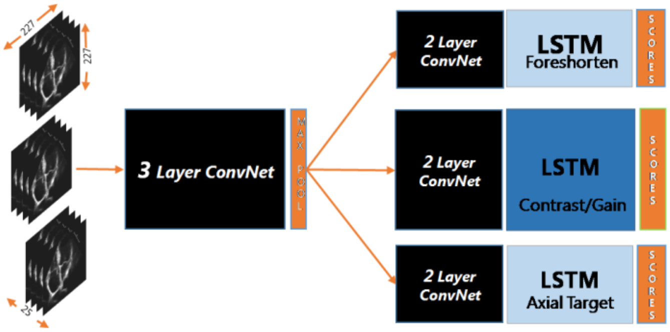
2.*4 Training*

Training hyper-parameters: The architecture consists of 3 regression models and was trained using 5-fold cross validation technique to ensure adequate learning on the dataset and performance was recorded for each model. The hyper parameters learning rate was set at 0.002 with high momentum 0.95 and decay rate of 0.1every 25 steps and were reproducibly initialized to minimise possible deviation in score performance. Training was initialised and completed as learning curves converged around 20 epochs.

Batch selection: The hardware computational cost during training phase ran high as batch selection of 3 and 6 were experimented, memory utilization becomes significantly apparent at batch selection of 6 at a fixed length sequence of 25 than running a batch size of 3 at the same fixed length sequence. Hardware performance difference of 0.025% in terms of computational speed was a negligible trade-off did not affect model's ability to properly generalize new test samples.

Data augmentation: Data augmentation was applied to allow optimum learning sequences for the models; a maximum translation of [-0.05, +0.05] pixels and maximum rotation of 10 degrees were applied randomly for horizontal, vertical and rotational angles respectively. To prevent over fitting in the training phase, we applied batch normalization at each convolution layer, early stopping and dropout (rate 0.30) for the training samples. Batch normalisation also helps stabilizes and speeds up convergence during the training phase.

Hardware and software resources: Model was implemented using PyTorch. The experiment was carried out on GPU GeForce GTX 970 chipset's Maxwell architecture and featuring 4GB RAM coupled to 1,664 CUDA cores.



Conv L1	Conv L2	Conv L3	Conv L4	Conv L5
8: 4x4	16:3x3	32:3x3	32:3x3	64:4x4
MP:2x2	MP:2x2	MP:2x2	MP:2x2	MP:2x2

Fig. 2 Multivariate Hybrid Network used in the research. Details of the kernels and filers sizes used in the multivariate hybrid network

2.5 *Evaluation Metric*

The Model's performance was evaluated in terms of difference between cardiologist's score (Q_{GT}) and model's predicted automatic score (Q_P):

$$Class_{err} = \sum_{i=0}^{n} |Q_{GTi} - Q_{pi}| \quad (1)$$

The average accuracy was computed as:

$$Model_{acc} = (1 - \sum_{i=0}^{n} |Q_{GTi} - Q_{pi}|) * 100 \qquad (2)$$

III.RESULT AND DISCUSSION

Given the complexity of varying pathological features in echo frames, our model could generalise on new echo frame with measured accuracy of 85.90 percent as shown in Table I. The error distribution per quality attribute is depicted in Fig.3 for axial target, contrast/gain and foreshortening properties, respectively.

The model prediction speed was found to be 0.013ms per frame for input pixel size of 227x 227 x 3, which is the assurance for real-time deployment and opportunity for enhancing clinical echo workflows.

TABLE II
COMPUTED ACCURACY FOR THE TEST SAMPLES

Models	**On-Axis**	**Contrast/ Gain**	**Fore-Shortening**	**Model Accuracy**
Accuracy	86.7%	84.6%	86.5%	86% (Avg.)

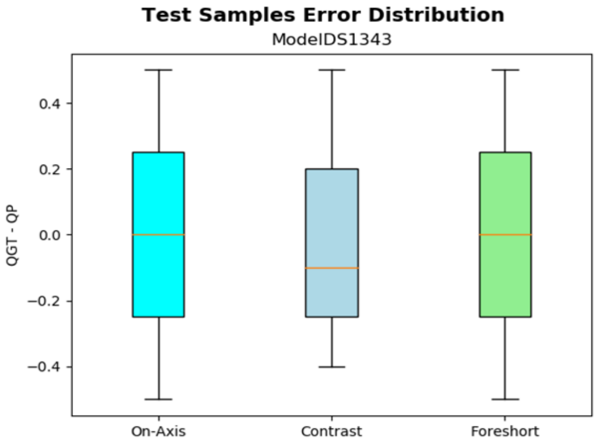


Fig. 3. Box Plot of the error distribution of test samples for each model. The error is computed as the difference between the prediction of the model and ground-truth. The x-axis shows three specific quality attributes of each model.

3.1 *Quality Scores Representation*

The scores for each quality properties are displayed in real-time on cardiac image during ultrasound image acquisition.

Since cardiologist must deal with multiple screen information, presenting two digits scores each for three quality properties is expected to add to screen complexity and may impair concentration or increase operator's capacity for additional performance. Therefore, we propose a symbolic representation of quality scores in defined to represent a range threshold as defined in Table II and a real time score output from three quality attribute model, showed in Fig. 4 as a coherent representation for operators' feedback. Fig. 5 shows some predicted samples from our experiment.

TABLE III.
SYMBOLIC REPRESENTATION OF QUALITY SCORE VALUES ON REAL-TIME CARDIAC IMAGING

Score Value	On-Axis Symbolic Score	Contrast/Gain Symbolic Score	Fore-Short Symbolic Score
0.63 - 0.90	A	A	A
0.41 - 0.62	B	B	B
0.00 - 0.40	C	C	C

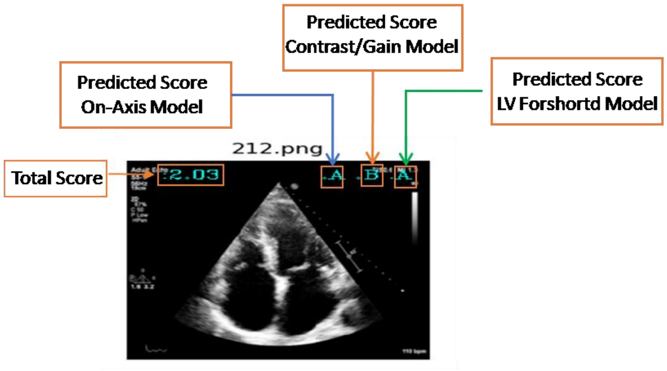


Fig. 4. Proposed feedback scores representation using symbolic letters to depict quality score ranges (depicted in Table III) from specific quality attributes models. This reduced complexity of flow of information to ultrasound operators during image acquisition phase.

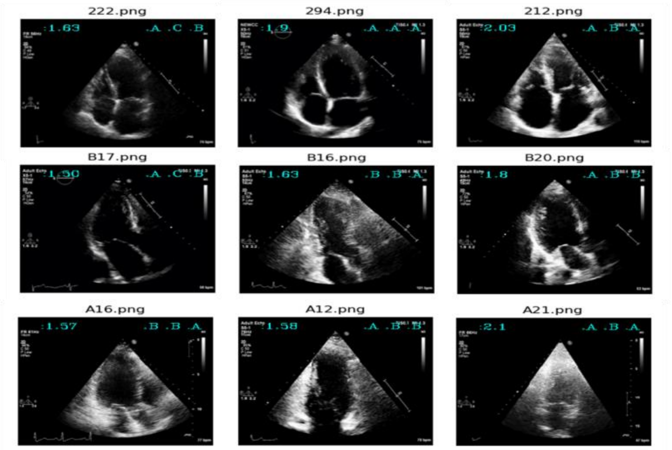


Fig. 5. Output samples showing predicted quality assessment scores for axial target, contrast/gain and foreshortening attributes of each image.

3.2 Study limitation and future work

In this study, we only considered A4C images. A future study can include different standard Echocardiographic views and more quality attributes that would satisfy wider requirements.

Several global characteristics can be used to distinguish between the different levels of quality and assign an image quality index. Here, we only considered 3attributes of image quality for feasibility studies. A more comprehensive study would include multiple criteria for 2D quality assessments.

Finally, we used the annotation provided by one expert cardiologist and once accredited annotator. Intra-observer variability can be examined by obtaining additional annotations from human experts and compared with the error in the predicted scores.

IV. CONCLUSION

We have presented the clinical significance and feasibility of developing an automated quality assessment in 2D Echocardiographic images.

An automated image quality assessment technique could be developed as part of a system that could accelerate the learning curve for those training in echocardiography, assisting as part of an automated quality control process (for both clinical and research purposes) and providing real-time guidance to less experienced operators to increase their chances of acquiring adequate images and enhance diagnostic accuracy of cardiac functions.

REFERENCES

[1] Mitchel et al. “Guidelines for performing a Comprehensive Transthoracic Echocardiographic Examination in Adults:” Recommendations from American Society of Echocardiography", 2018.

[2] Chirillo, F., Pedrocco, A., De Leo, A., Bruni, A., Totis, O., Meneghetti, P. and Stritoni, P.: Impact of harmonic imaging on transthoracic echocardiographic identification of infective endocarditis and its complications. Heart 91(3), 329{333 (2005)

[3] Nagata, Y., Kado, Y., Onoue, T., Otani, K., Nakazono, A.,Otsuji, Y. and Takeuchi, M.: Impact of image quality on reliability of the measurements of left ventricular systolic function and global longitudinal strain in 2D echocardiography. Echo research and practice, 5(1), 27{39 (2018)

[4] Eckert, M.P. and Bradley, A.P.: Perceptual quality metrics applied to still image compression. Signal processing, 70(3), 177{200 (1998)

[5] Wang, Z. and Bovik, A.C.: A universal image quality index. IEEE signal processing letters 9(3),81{84 (2002)

[6] Luo, H.: A training-based no-reference image quality assessment algorithm. International Conference on Image Processing, 2004. ICIP'04, Vol. 5, pp. 2973{2976. IEEE (2004). https://doi.org/10.1109/ICIP.2004.1421737

[7] Li, X.: Blind image quality assessment. In Proceedings. International Conference on Image Processing, Vol. 1, pp.I{I. IEEE (2002). https://doi.org/10.1109/ICIP.2002.1038057

[8] Abdi, A.H., Luong, C., Tsang, T., Allan, G., Nouranian, S., Jue, J., Hawley, D.,Fleming, S., Gin, K., Swift, J. and Rohling, R.: 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)

[9] Abdi, A.H., Luong, C., Tsang, T., Allan, G., Nouranian, S., Jue, J., Hawley, D.,Fleming, S., Gin, K., Swift, J. and Rohling, R.: Automatic quality assessment of echocardiograms using convolutional neural

networks: feasibility on the apical four chamber view. IEEE transactions on medical imaging, 36(6), pp.1221-1230 (2017)

[10] Abdi, A.H., Luong, C., Tsang, T., Jue, J., Gin, K., Yeung, D., Hawley, D., Rohling,R. and Abolmaesumi, P. 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 vol. vol. 10435, pp. 302{310. Springer, Cham (2017)

[11] Dong, J., Liu, S., Liao, Y., Wen, H., Lei, B., Li, S. and Wang, T.: A Generic Quality Control Framework for Fetal Ultrasound Cardiac Four-chamber Planeside journal of biomedical and health informatics (2019)

[12] Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(4), 677{691 (2017)

[13] IntSav Research Repository: http://echo-ids.com/intsav/dataset