Original Article

An automatic and real-time echocardiography quality scoring system based on deep learning to improve reproducible assessment of left ventricular ejection fraction

Xiaoshan Li[1#], Lisi Liao[2#], Kai Wu[3#], Alexander Thomas Meng[4], Yitao Jiang[5,6], Yuan Zhu[1,7], Chen Cui[8], Xiaowei Xu[1], Bobo Shi[2], Hongwen Fei[1^]

[1]Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China; [2]Department of Ultrasound, Shenzhen People's Hospital (the Second Clinical Medical College, Jinan University, the First Affiliated Hospital, Southern University of Science and Technology), Shenzhen, China; [3]Department of Ultrasound, The Third Affiliated Hospital (Shenzhen Luohu People's Hospital) of Shenzhen University, Shenzhen, China; [4]Shanghai American School, Shanghai, China; [5]School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Shenzhen, China; [6]Shenzhen Yingzhichuangsi Technology Co., Ltd., Shenzhen, China; [7]Guangdong Medical University, Zhanjiang, China; [8]Shenzhen MicroPort Xinsuanzi Medical Technology Co., Ltd. Shenzhen, China

Contributions: (I) Conception and design: H Fei, X Xu, B Shi; (II) Administrative support: H Fei; (III) Provision of study materials or patients: X Li, L Liao, B Shi, Y Zhu; (IV) Collection and assembly of data: C Cui, Y Jiang, AT Meng; (V) Data analysis and interpretation: K Wu, X Li; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Xiaowei Xu, MD, PhD. Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, No. 106 Zhongshan 2nd Road, Guangzhou 510080, China. Email: xuxiaowei@gdph.org.cn; Bobo Shi, MD, PhD. Department of Ultrasound, Shenzhen People's Hospital (the Second Clinical Medical College, Jinan University, the First Affiliated Hospital, Southern University of Science and Technology), 1017 Dongmen North Road, Luohu District, Shenzhen 518020, China. Email: sumi150117@163.com; Hongwen Fei, MD, PhD. Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, No. 106 Zhongshan 2nd Road, Guangzhou 510080, China. Email: feihongwen@gdph.org.cn.

Background: Echocardiography can conveniently, rapidly, and economically evaluate the structure and function of the heart, and has important value in the diagnosis and evaluation of cardiovascular diseases (CVDs). However, echocardiography still exhibits significant variability in image acquisition and diagnosis, with a heavy dependency on the operator's experience. Image quality affects disease diagnosis in the later stage, and even image quality assessment still has variability in human evaluation. This study aimed to develop an automated and real-time quality assessment system using deep learning (DL) techniques while decreasing the measurement error of left ventricular ejection fraction (LVEF).

Methods: This study involved over 5,000 echocardiography datasets from 2,461 participants across 10 medical centers in China to build the model. A 5-point quality scoring system was used to assess the integrity, clarity, and alignment of anatomical structures in each echocardiogram view. Additionally, an innovative DL model was developed to autonomously detect these essential cardiac anatomical structures in real-time, subsequently providing quality score estimations and LVEF. A total of 175 participants from two distinct external medical centers were enrolled for model validation. This dataset was employed to assess the consistency and repeatability of quality score and ejection fraction (EF) measurements, and the assessments made by human experts were compared with those of our model.

[^] ORCID: 0000-0003-3731-2668.

Results: The developed model demonstrated exceptional performance, achieving Intersection over Union (IoU) scores exceeding 0.8 for left ventricular (LV) segmentation, a mean average precision when IoU >0.5 (mAP50) of 0.91 for cardiac anatomical structures detection, and a 0.96±0.05 accuracy in view classification. The quality scores assessed by the model closely matched those of human experts, indicating strong agreement. The weighted average precision and weighted average recall scores fell within the range of 0.5 to 0.6. Notably, there was no statistically significant difference in LVEF assessments between human experts and our model (P=0.09), as demonstrated by an intraclass correlation coefficient (ICC) analysis of 0.821, reflecting high-level consistency. When assessing echocardiograms with high-quality scores, the model demonstrated a significantly closer alignment and a higher correlation coefficient with human experts (R=0.90±0.04).

Conclusions: This study demonstrates that artificial intelligence-assisted echocardiography scoring system aligns well with manual quality scoring. Through the supervision of real-time echocardiogram quality, the artificial intelligence model can assist doctors in providing more reproducible and consistent assessments of cardiac function.

Keywords: Echocardiography; quality control; cardiac function assessment; deep learning (DL)

Submitted Mar 14, 2024. Accepted for publication Dec 11, 2024. Published online Dec 30, 2024.
doi: 10.21037/qims-24-512
View this article at: https://dx.doi.org/10.21037/qims-24-512

Introduction

Cardiovascular diseases (CVDs) are a group of disorders that affect the heart and blood vessels. CVDs represent a significant global health and economic concern and remain a leading cause of morbidity and mortality worldwide, accountable for approximately 20 million deaths in 2021. In China, CVDs are the cause of over 40% of deaths, with an estimated 330 million individuals experiencing CVDs (1). Given their substantial impact on public health, addressing CVD stands as a top priority for healthcare systems and society as a whole. Compared to medical interventions and clinical management, early detection is a more effective and foundational strategy for reducing the prevalence and burden of CVDs.

Echocardiography, a non-invasive ultrasound imaging technique, serves as a cornerstone in the assessment of cardiovascular health. It furnishes critical information and various parameters that are indispensable in clinical diagnosis, medical management, and long-term monitoring. Consequently, echocardiography has emerged as the foremost imaging modality for most CVDs (2). However, echocardiography is operator-dependent, particularly in critical situations such as cardiac arrest, tamponade, or complications during procedures such as percutaneous cardiac interventions. In such cases, prompt ultrasound scanning by an experienced clinician is required. In remote or rural healthcare settings, the prompt availability of echocardiography may be hindered by a lack of equipment and training.

One of the critical parameters derived from echocardiography is the left ventricular ejection fraction (LVEF), which quantifies the heart's pumping efficiency and plays a pivotal role in diagnosing and monitoring cardiac conditions. The conventional pipeline for LVEF assessment entails manual frame selection at the systolic end (ES) and diastolic end (ED), as well as the left ventricular (LV) segmentation before any subsequent analysis (3). This process is both user-dependent and time-consuming, requiring extensive practice to master. Moreover, complications can arise when electrocardiography (ECG) data is missing during the echocardiography examination, leading to a potential decrease in the precision of the ES and ED frames selection and subsequently impacting the LVEF accuracy. Another challenge associated with the utilization of echocardiography is the inherent complexity and procedural intricacies, which demand extensive training and clinical experience to consistently produce high-quality echocardiograms that meet the requisite LVFF (4).

In response to these challenges, there is a growing demand for efficient tools capable of real-time echocardiography quality control and the automation of LVEF calculation. A quality control system can guide the operators, ensuring the attainment of standardized echocardiography views and thereby enhancing the consistency and accuracy of the cardiovascular function measurements, including LVEF.

Existing literature has introduced numerous solutions for assessing echocardiography quality and facilitating computer-assisted LVEF calculation. In many of these algorithms, echocardiography quality assessment is achieved by performing an image-level classification task encompassing factors such as clarity (5-7), on-axis attributes (6), depth grain (6,7), and view types (5,8,9). However, these studies still exhibit certain limitations. Firstly, many studies often lack well-defined criteria or quality control indicators and also have a deficiency in model interpretability. Secondly, these algorithms are often not tailored for specific clinical purposes, including LVEF estimation. Lastly, some models are excessively large and inefficient, impeding their ability to achieve real-time quality assessment in clinical scenarios.

This research presents a pioneering endeavor aimed at establishing an automatic, efficient, and real-time quality scoring system coupled with LVEF calculation using deep learning (DL) techniques. Leveraging the power of artificial intelligence and a vast dataset of thousands of participants, our study strives to revolutionize echocardiography examination by reducing the burden on healthcare professionals and enhancing diagnostic accuracy. Additionally, junior operators benefit from improved image acquisition efficiency and the automated cardiac function assessment, which serves as a valuable reference for triage and management decisions. We present this article in accordance with the TRIPOD+AI reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-24-512/rc).

Methods

Participant cohorts and data preparation

The participants in this retrospective study were drawn from 10 separate top-level hospitals located within mainland China: the Second Hospital of Jiaxing, Panjin Liaoyou Gemstone Flower Hospital, the First Affiliated Hospital of Guangxi University of Chinese Medicine, the First Affiliated Hospital of Chongqing Medical University, the Second Affiliated Hospital of Guangxi University of Chinese Medicine, Shanghai General Hospital, Guizhou Liupanshui Shougang Shuigang Hospital, Shunde Hospital of Southern Medical University, the First Hospital of Lanzhou University, and Peking University Shougang Hospital. We included consecutive cases that met our inclusion and exclusion criteria between November 2022 and July 2023. The patient inclusion criteria were as follows: (I) aged over 18 years; (II) had undergone or were planning to undergo echocardiographic examinations; and (III) there were no restrictions on the presence or type of diseases. The exclusion criteria were solely those who refused to participate in this study.

Firstly, the echocardiographic data of these participants spanning three consecutive cardiac cycles were collected and saved in the Digital Imaging and Communication in Medicine (DICOM) format. Secondly, we performed data filtering to select the apical 2-chamber (A2C) view and apical 4-chamber (A4C) view. Finally, ultrasound doctors selected images of the ES and ED stages of the cardiac cycle based on the electrocardiogram and manually labeled them. The regions of LV endocardium (LV-Endo) and LV epicardium (LV-Epi) were annotated through semantic segmentation. The LV, left atrium (LA), and mitral valve (MV) were delineated in both A2C and A4C views by bounding boxes. The right ventricle (RV), right atrium (RA), and tricuspid valve (TV) were only delineated in the A4C view. Each case was first annotated by two doctors (with 5 or more years of experience in cardiac ultrasound), and the inconsistencies were then arbitrated by a senior doctor (with 10 or more years of experience). A total of 60 cardiac ultrasound doctors participated in the data annotation step.

Echocardiographic quality scoring system

After extensive discussions among 10 highly experienced echocardiographers, we quantified the qualitative quality assurance grading scales based on the "Emergency Ultrasound Standard Reporting Guidelines" (10). As a result, we introduced a 5-point quality scoring system that is well-suited for computer-aided calculation in echocardiography examination. This quality scoring system primarily evaluates the integrity and clarity of the key cardiac anatomical structures in the echocardiogram. Specifically, this system is based on the following criteria for quality scoring: (I) whether the ventricles and atria can be recognized; (II) whether the LV-Endo and LV-Epi can be fully segmented; (III) whether the ratio of LV to LA length is between 1.5 and 2.5; and (IV) whether there is a clear view of the LV. If all these criteria are met correctly, the frame has a quality score of 5, and if all these criteria

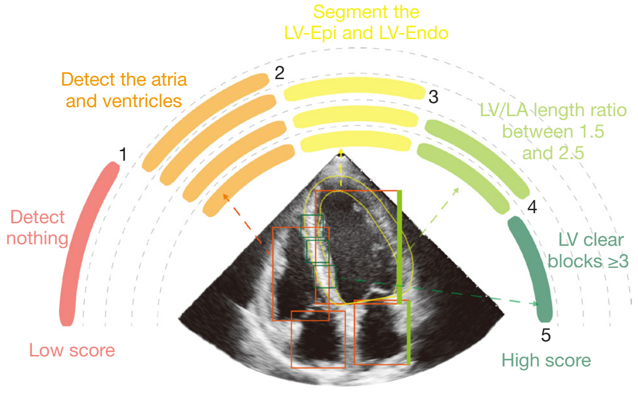

Segment the LV-Epi and LV-Endo
Detect the atria and ventricles
LV/LA length ratio between 1.5 and 2.5
Detect nothing
LV clear blocks ≥3
1
2
3
4
5
Low score
High score

Figure 1 The echocardiographic quality scoring system for A2C and A4C views. The indications of different colors in the diagram represent the criteria that each quality score needs to meet. If nothing can be detected, the echocardiographic quality score is 1. When meeting the criteria with orange, yellow, cyan, and green colors, the quality score is 5. LV, left ventricle; LA, left atrium; A2C, apical 2-chamber; A4C, apical 4-chamber.

are incorrect, the quality score for this frame is 1 (*Figure 1*). The average of all frame quality scores is used as the quality score for the entire echocardiogram.

To achieve automated and real-time quality assessment of echocardiograms, we constructed an efficient DL model to evaluate the above scoring criteria. This model consists of the following parts: (I) an encoder for extracting image features, derived from the backbone of YOLOX (11); (II) a decoder for providing the A2C and A4C view-class information, consisting of two layers of fully connected neural networks; (III) a decoder for detecting the regions of LV, LA, MV, RV, LV, and TV, derived from the detection head of YOLOX; (IV) a decoder for LV-Endo and LV-Epi segmentation, constructed with several deconvolution layers; and (V) two classification networks for recognizing the segmentation continuity and block clarity of LV-Endo, composed of convolutional neural networks (*Figure 2*). In practice, we divide the LV-Endo into five blocks based on the segmentation results. We then assess the clarity of the ultrasound images of these five blocks. If at least three out of the five blocks are clear, the image is considered to have a clear view of the LV.

The annotated echocardiographic data were randomly split 7:1:2 into training, validation, and test sets. During the training phase, we employed various data augmentation strategies, including blur, rotation, and translation, to enhance the model's generalization capabilities. These techniques have been shown to significantly expand the effective size of the dataset and improve the model's robustness. The outputs of these decoders and classification networks are used to measure whether the scoring criteria are met.

Ejection fraction (EF) calculation

We have referenced Simpson's biplane method to construct an automatic EF calculator (12). After obtaining the model outputs, we first compared the LV area of all echocardiogram frames. After applying the Kalman smoothing to the LV segmentation (13), we chose to label the frame with the smallest LV cavity area as the ES frame and the frame with the largest cavity area LV as the ED frame. Then, LV volume is calculated as:

$$Volume = \frac{\pi}{4}\sum_{i=1}^{n} a_i \cdot b_i \cdot \frac{L}{n} \qquad [1]$$

where L is the long-axis length of LV. When LV is divided into n equal parts, a_i and b_i are the short-axis lengths of LV in the A2C and A4C views, respectively. When obtaining the LV volume of ED ($Volume_{ED}$) and ES ($Volume_{ES}$), the LVEF is calculated as:

$$LVEF = \frac{Volume_{ED} - Volume_{ES}}{Volume_{ED}} \cdot 100\% \qquad [2]$$

When dealing with echocardiograms spanning multiple cardiac cycles, we consider the highest LVEF value within the sequence as the representative LVEF for that particular echocardiogram.

Comparing model performance to human experts

We conducted a two-center retrospective study to compare the performance of our model and human experts in echocardiographic quality score and LVEF calculation (*Figure 3*). The two centers were Guangdong Provincial People's Hospital and Shenzhen People's Hospital. Using the same inclusion and exclusion criteria as for model construction, External Dataset 1 comprised 65 participants who underwent echocardiogram examinations between February 2022 and June 2023. Similarly, External Dataset 2 encompassed 110 participants who also underwent echocardiogram examinations within the same time frame, from February 2023 to June 2024.

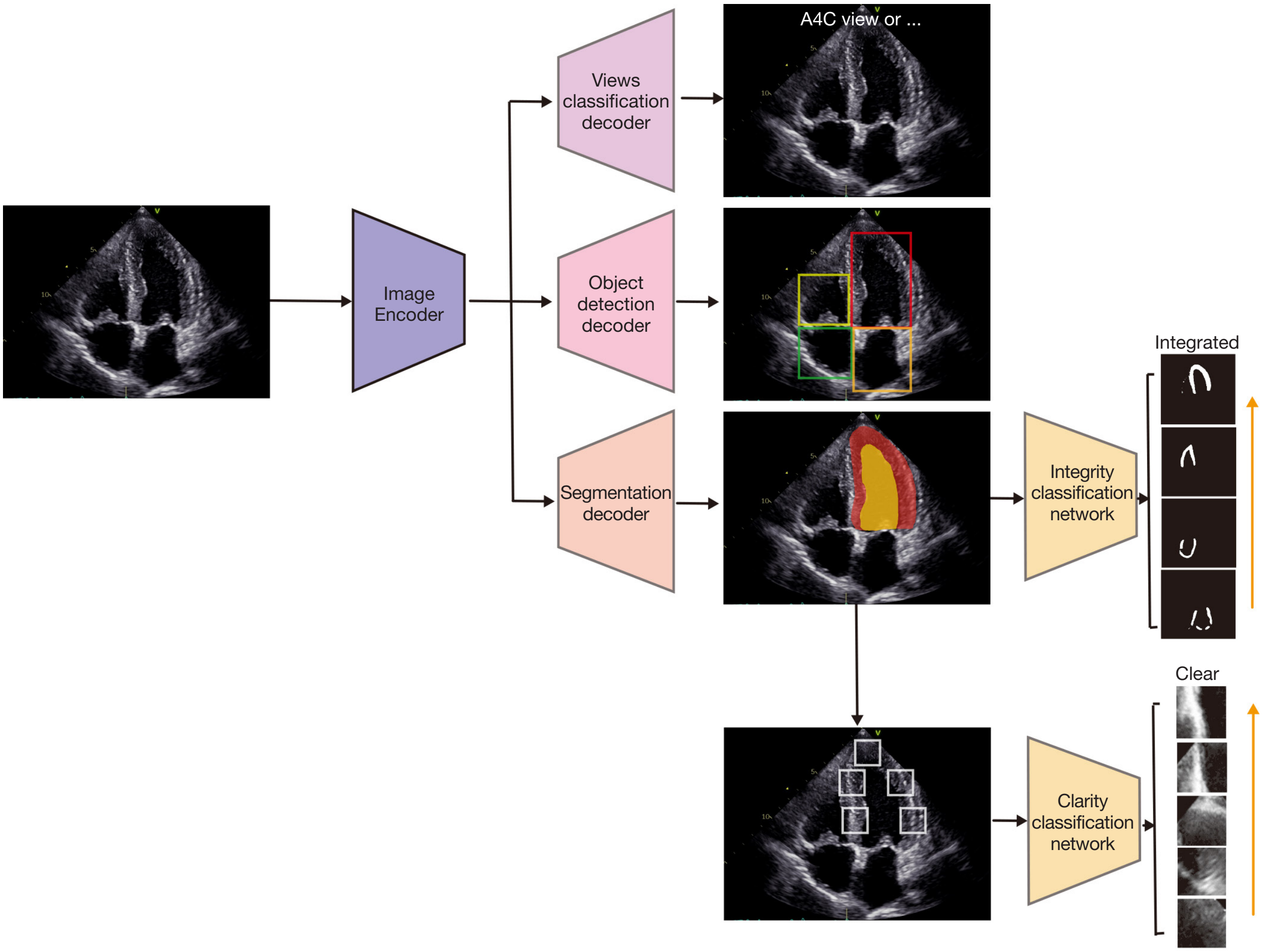

A4C view or ...
Views classification decoder
Image Encoder
Object detection decoder
Segmentation decoder
Integrity classification network
Integrated
Clarity classification network
Clear

Figure 2 The architecture of the DL framework. Our model comprises six modules (trapezoid): one for feature extraction, three for decoding, and two for classification using lightweight DL networks. All six modules are learnable and designed for efficient and real-time inference. A4C, apical 4-chamber; DL, deep learning.

We provided a set of paired echocardiograms, consisting of A2C and A3C views, covering two cardiac cycles, to two highly experienced doctors, each possessing over 10 years of clinical expertise in cardiac ultrasound. Their task was to conduct manual assessments of the echocardiogram quality scores based on our 5-point quality assessment system and perform LV segmentation. The manually segmented LV was used to calculate the LVEF by Simpson's biplane method. If there were discrepancies in the assessment between the two doctors, the echocardiogram was arbitrated by a doctor with over 15 years of clinical experience to establish the gold standard. At the same time, our model automatically calculated the quality scores and LVEF. External Dataset 2 was strategically utilized to assess the impact of our quality scoring system on the LVEF evaluation process conducted by human experts. In this dataset, LVEF was manually determined by two physicians, whereas the quality scores were concurrently evaluated by our model. Statistical analyses, including comparisons and consistency, were conducted among the two doctors, one arbitrator, and our model.

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This study was approved by the Research Ethics Committee of Guangdong Provincial People's Hospital (No. QX2023-041-02). The requirement for informed consent was waived since this was a retrospective and observational study. All participating hospitals were informed of and agreed to the study, and all

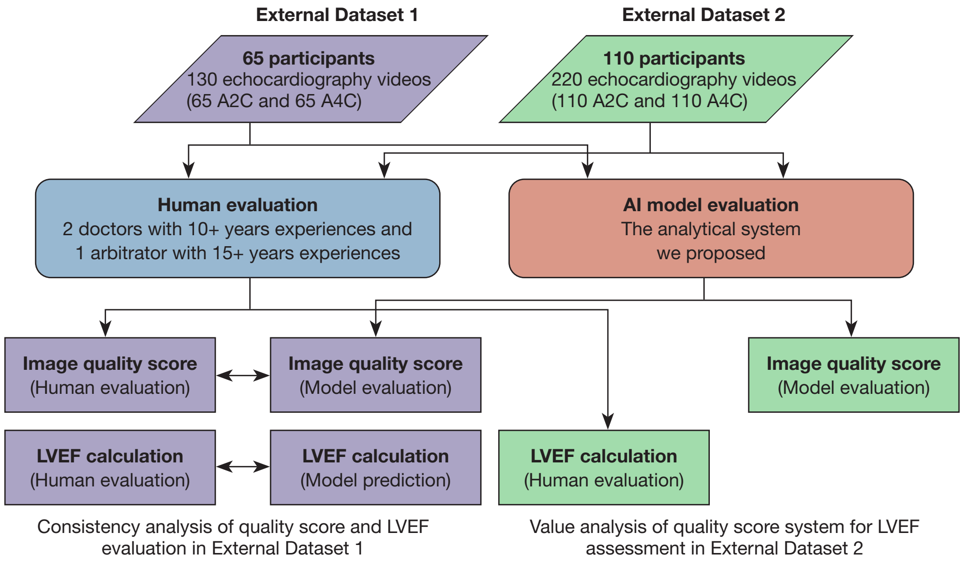


Figure 3 Diagram of the performance comparison experiment between our model and human experts in two external datasets. One hundred and seventy-five participants within two external datasets were involved in this study. The echocardiographic quality score and LVEF were assessed by two doctors, and one arbitrator and our model. External Dataset 1 was primarily utilized for conducting consistency analysis between the assessments of human experts and our model. In contrast, External Dataset 2 was specifically employed to evaluate the contribution of the quality scoring system to the LVEF evaluation process by human. A2C, apical 2-chamber; A4C, apical 4-chamber; AI, artificial intelligence; LVEF, left ventricular ejection fraction.

were subject to ethical approval through application.

Statistical analysis

One-way analysis of variance (ANOVA) was employed to evaluate whether there were significant differences in the LVEF measurements among multiple groups. Meanwhile, one-way intraclass correlation coefficient (ICC) was utilized to gauge the reliability and consistency of LVEF measurements conducted by different groups. The joint hypotheses test (*F*-test) and Welch's two-sample *t*-test were used to compare the variances and means of the two groups. Pearson's correlation coefficient was calculated to measure the correlation. The comparison of the two correlations was performed with the R-package "*cocor*" (v1.1-4) (14), which embedded the multiple comparison algorithms. All these DL experiments and statistical analyses were conducted in Python (v3.10) and R (v4.3.1). The results of statistical analysis were deemed significant when the P value <0.05. The uncertainty of the estimate such as accuracy and correlation coefficient were quantified at the 95% confidence interval (CI).

Results

Participant characteristics

Two-dimensional (2D) echocardiography data from a total of 2,461 participants (1,132 females and 1,329 males) were involved in the model construction process. The basic characteristics of the participants and 2D echocardiography data collected are shown in *Table 1*. In summary, the mean age was 52.4 years and the standard deviation (SD) was 15.9 years. Half of the participants (1,235, 50.2%) were from the eastern region of China, and the rest were from the northeast (285, 11.6%), western (255, 10.4%), and southern-central (686, 27.9%) regions. There were 4,736 echocardiograms with A2C and A4C views collected during this study. A total of 8,055 frames from 4,052 echocardiograms, at ED and ES, were annotated for LV-Endo and LV-Epi segmentation. In addition, 1,368 frames out of 684 echocardiograms were selected and annotated for LV, LA, MV, RV, RA, and TV using bounding boxes. The annotated echocardiographic frames were randomly divided into training, validation, and test sets at a 7:1:2 ratio for model construction.

Table 1 Basic characteristics of participants and dataset

Characteristics	Data
Basic characteristics	
Participants (cases)	2,461
Age (years)	52.4 (15.9)
Gender	
Female	1,132
Male	1,329
Regional distribution	
Eastern region of China	1,235
Northeast region of China	285
Western region of China	255
Southern-central region of China	686
Dataset characteristic	
Echocardiogram videos (A2C and A4C views)	4,736
Frames for LV segmentation	8,055
Frames for cardiac structure detection	1,368

Data are presented as number or mean (SD). A2C, apical 2-chamber; A4C, apical 4-chamber; LV, left ventricle; SD, standard deviation.

Real-time and efficient model identification of cardiac anatomical structures

The model we constructed demonstrated excellent and efficient recognition capabilities for cardiac anatomical structures. Using a video capture card, our model was shown to be capable of real-time prediction during echocardiographic examinations. We presented two video examples to illustrate the model's real-time prediction capabilities on A2C and A4C echocardiograms (Videos S1,S2). In *Figure 4A,4B*, two image examples, the A2C view and the A4C view, demonstrate the model's segmentation and object detection performance. On the entire test set, our model achieved high accuracy in LV-Endo and LV-Epi segmentation with Intersection over Union (IoU) scores of 0.81±0.08 and 0.83±0.07 at ES, and IoU scores of 0.86±0.05 and 0.87±0.05 at ES, respectively (*Figure 4C*). We calculated the offset rate for the myocardial end positions (the red points in *Figure 4A,4B*), which mark the end of LV segmentation and play a crucial role in LVEF calculations. In the test set, the average root mean square deviation (RMSD) between the predicted key points and the ground truth was 2.39 pixels at ES and 2.68 pixels at ED. Additionally, the mean and SD of the offset rate for the key points was 1.07%±0.75% at ES and 1.19%±0.70% at ED (*Figure 4D*). For the detection of cardiac anatomical structures, our model achieved IoU scores of 0.80±0.08, 0.83±0.07, and 0.66±0.12 for detecting LA, LV, and MV on A2C echocardiography (*Figure 4E*). Additionally, it achieves IoU scores of 0.79±0.13, 0.83±0.13, 0.80±0.10, 0.67±0.14, 0.70±0.13, and 0.62±0.17 for detecting LA, LV, MV, RA, RV, and TV on A4C echocardiography, respectively (*Figure 4F*). In summary, our model achieved a mean average precision when IoU >0.5 (mAP50) score of 0.91 and exhibited a 0.96±0.05 classification accuracy of A2C and A4C echocardiograms at a 95% CI.

Echocardiography quality assessment for the study participants

A total of 65 participants (20 females and 45 males) were involved in the External Dataset 1, comparing the model's performance to that of human experts. The basic and clinical characteristics of participants are displayed in *Table 2*. In summary, the mean age of the participants is 58.7 years, with an SD of 9.6 years. Among the participants, 29 had been diagnosed with coronary heart disease for whom the mean EF value was 56.80±9.87, 23 with LV hypertrophy (EF: 59.60±10.54), 27 with cardiac amyloidosis (EF: 53.36±10.71), 35 with preserved EF (EF: 58.23±10.09), 25 with pericardial effusion (EF: 57.23±10.48), 35 had been diagnosed with mitral regurgitation (EF: 55.37±10.11), and 7 with arrhythmia (EF: 57.81±10.00). In External Dataset 2, the mean age of the participants was 48.5 years, with an SD of 12.6 years. Among the participants, 25 had been diagnosed with coronary heart disease for whom the EF value was 66.67±8.65, 15 with LV hypertrophy (EF: 65.65±13.19), 9 with preserved EF (EF: 58.55±15.44), 8 with pericardial effusion (EF: 69.10±8.23), 35 had been diagnosed with mitral regurgitation (EF: 66.10±6.95), and 10 with arrhythmia (EF: 62.03±10.31). There were no patients with cardiac amyloidosis in the External Dataset 2.

The 65 paired A2C and A4C echocardiograms were manually scored for quality and segmented for LV segment by two doctors, with arbitration by one arbitrator. Based on our 5-point quality scoring system, the distribution of samples as assessed by the two doctors and one arbitrator is illustrated

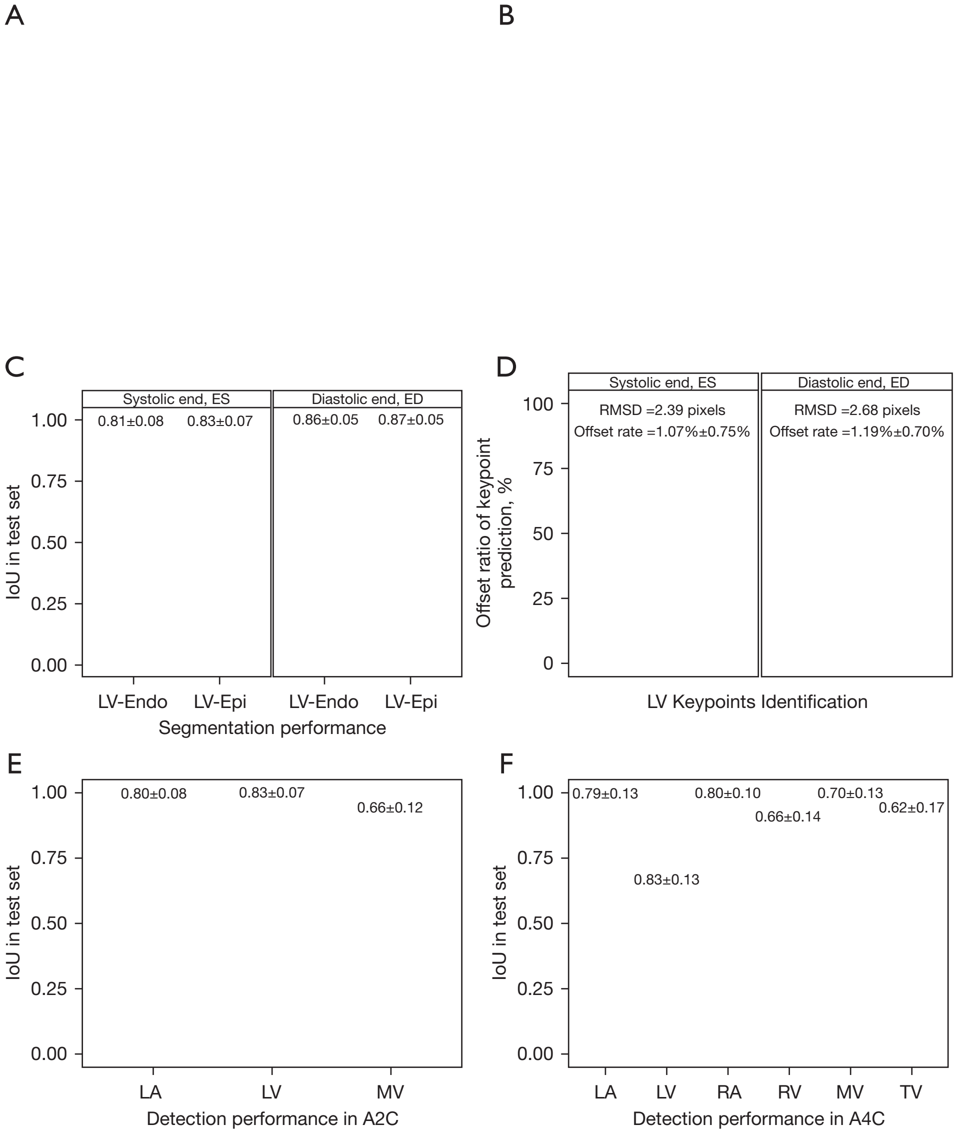


Figure 4 Model performance in cardiac anatomy recognition. (A) Model for LV-Endo (yellow region) and LV-Epi (red region) segmentation, and LV, LA, MV detection by bounding boxes at ED in echocardiographic A2C view. The numbers in the figure represent the confidence of the corresponding box. (B) Model for LV-Endo (yellow region) and LV-Epi (red region) segmentation, and LV, LA, MV, RV, RA, and TV detection by bounding boxes at the ES in echocardiographic A4C view. (C) The boxplot illustrates the IoU scores for LV-Endo and LV-Epi segmentation at both ES and ED in the test set. The higher the score, the more accurate the detection position is. (D) The boxplot depicts the offset rate (RMSD/image length) of the myocardial end positions at both ES and ED in the test set. (E) The boxplot shows the IoU scores for LA, LV, and MV detection in the A2C view of the test set. (F) The boxplot displays the IoU scores for LA, LV, MV, RA, RV, and TV detection in the A4C view of the test set. The numbers accompanied by "±" represent the mean ± SD. LA, left atrium; LV, left ventricle; MV, mitral valve; A2C, apical 2-chamber; RA, right atrium; RV, right ventricle; TV, tricuspid valve; A4C, apical 4-chamber; ES, systolic end; ED, diastolic end; LV-Endo, left ventricular endocardium; LV-Epi, left ventricular epicardium; IoU, Intersection over Union; RMSD, root mean square deviation; SD, standard deviation.

Table 2 Clinical characteristics of participants of a performance comparison study

Characteristics	External dataset 1	External dataset 2
Basic characteristics		
Participants (cases)	65	110
Age (years)	58.7 (9.6)	48.5 (12.6)
Female	20	64
Clinical characteristic		
Coronary heart disease	29	25
LV hypertrophy	23	15
Cardiac amyloidosis	27	0
Preserved EF	35	9
Pericardial effusion	25	8
Mitral regurgitation	35	35
Arrhythmia	7	10

Data are presented as number or mean (SD). LV, left ventricular; EF, ejection fraction; SD, standard deviation.

in *Figures 5A-5C*, respectively. Additionally, *Figure 5D* displays the results of our automatic, real-time quality scoring model. In all assessments, the distribution of quality scores for the samples exhibited remarkable similarity: more than 95% of the samples received quality scores of 3 or higher, with half of the samples receiving a score of 4.

Model quality assessment results consistent with human experts

To assess the consistency of echocardiographic quality scores between our model and human experts, we generated a multi-category confusion matrix comparing the results between the two doctors (*Figure 6A*), between the model and the two doctors (*Figure 6B,6C*), and between the model and the arbitrator (*Figure 6D*). In every confusion matrix, the weighted average precision and weighted average recall scores fell within the range of 0.5 to 0.6. The results indicate that the differences between our model and human experts are consistent with the difference between the two doctors.

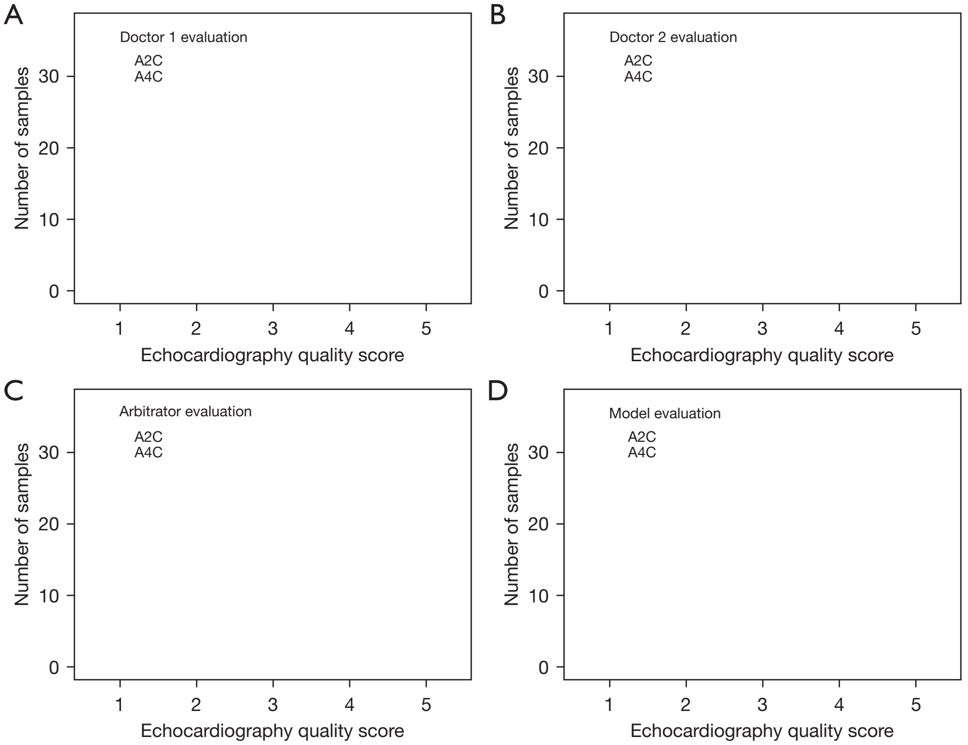
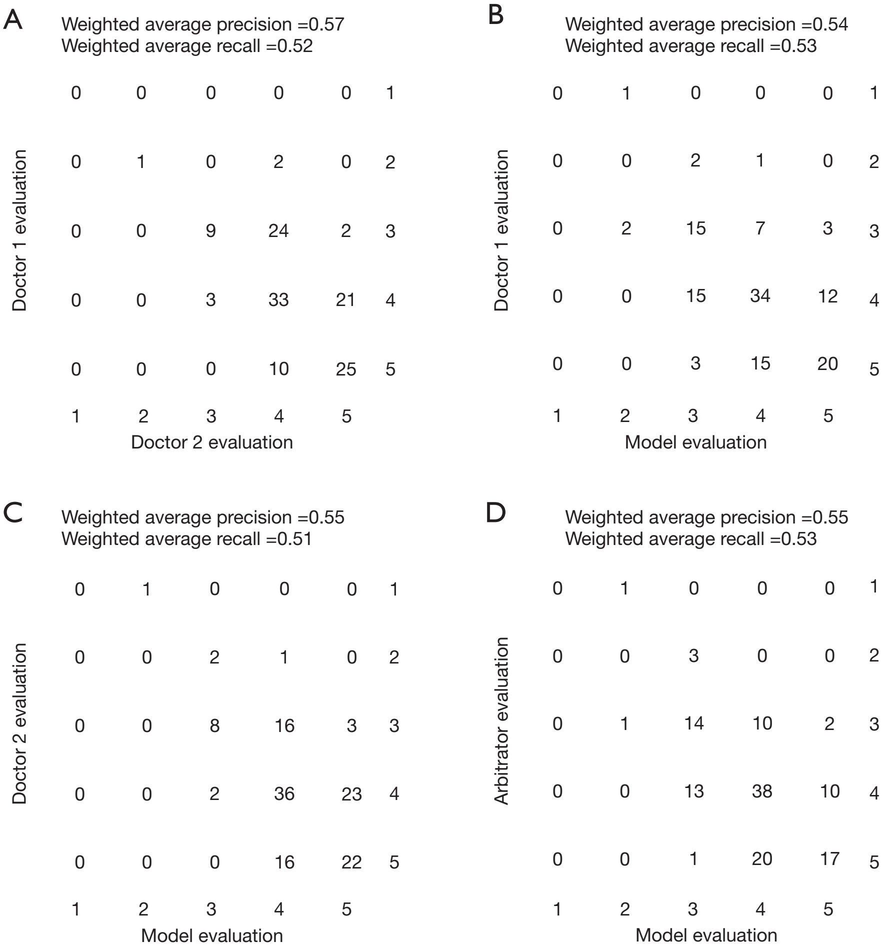


Figure 5 Distribution of quality scores in echocardiography across 65 participants. (A-D) The bar plots represent the distribution of echocardiographic quality scores for both the A2C view and the A4C view. These quality scores, which ranged from 1 to 5, were assessed by the two doctors, one arbitrator, and our model, respectively. A2C, apical 2-chamber; A4C, apical 4-chamber.



A

Weighted average precision =0.57
Weighted average recall =0.52

1	2	3	4	5	
0	0	0	0	0	1
0	1	0	2	0	2
0	0	9	24	2	3
0	0	3	33	21	4
0	0	0	10	25	5

B

Weighted average precision =0.54
Weighted average recall =0.53

1	2	3	4	5	
0	1	0	0	0	1
0	0	2	1	0	2
0	2	15	7	3	3
0	0	15	34	12	4
0	0	3	15	20	5

C

Weighted average precision =0.55
Weighted average recall =0.51

1	2	3	4	5	
0	1	0	0	0	1
0	0	2	1	0	2
0	0	8	16	3	3
0	0	2	36	23	4
0	0	0	16	22	5

D

Weighted average precision =0.55
Weighted average recall =0.53

1	2	3	4	5	
0	1	0	0	0	1
0	0	3	0	0	2
0	1	14	10	2	3
0	0	13	38	10	4
0	0	1	20	17	5

Figure 6 Multi-category confusion matrix comparing the evaluation of quality scores. The multi-category confusion matrix compares the quality scoring results between the two doctors (A), between the model and the two doctors (B,C), and between the model and the arbitrator (D). Consistency assessment was performed using the weighted average precision and weighted average recall score.

Efficient model computation of LVEF

After deriving the LV segmentation results, the LVEF for each participant was calculated using our automatic EF calculator. Subsequently, we performed one-way ANOVA and one-way ICC analysis to investigate the differences and consistency of LVEF measurements among the two doctors, one arbitrator, and our model. The results of the tests indicate that there is no statistically significant difference in the LVEF assessments conducted by the four groups, as evidenced by an ANOVA with a P value of 0.09. Moreover, there is a high level of consistency, as indicated by an ICC of 0.821 within a 95% CI, ranging from 0.752 to 0.877 (*Figure 7A*).

Then, we performed Pearson correlation analysis on the LVEF measurements between cardiac cycles 1 and 2 within each group. Remarkably, our model exhibited the highest correlation coefficient of 0.95±0.03, at a 95% CI. Next, we conducted a comparative analysis of the differences among these four sets of correlation coefficients utilizing the Fisher's z method from the R-package "*cocor*". The statistical findings revealed that our model exhibited a stronger correlation in LVEF between cardiac cycles 1 and 2 when compared to human experts. Specifically, the P value was 0.0002 for the statistical analysis between doctor 1 and the model, 0.055 for doctor 1 and the model, and 0.004 for the arbitrator and the model (*Figure 7B*). We computed the distance and correlation between the estimated LVEF of the

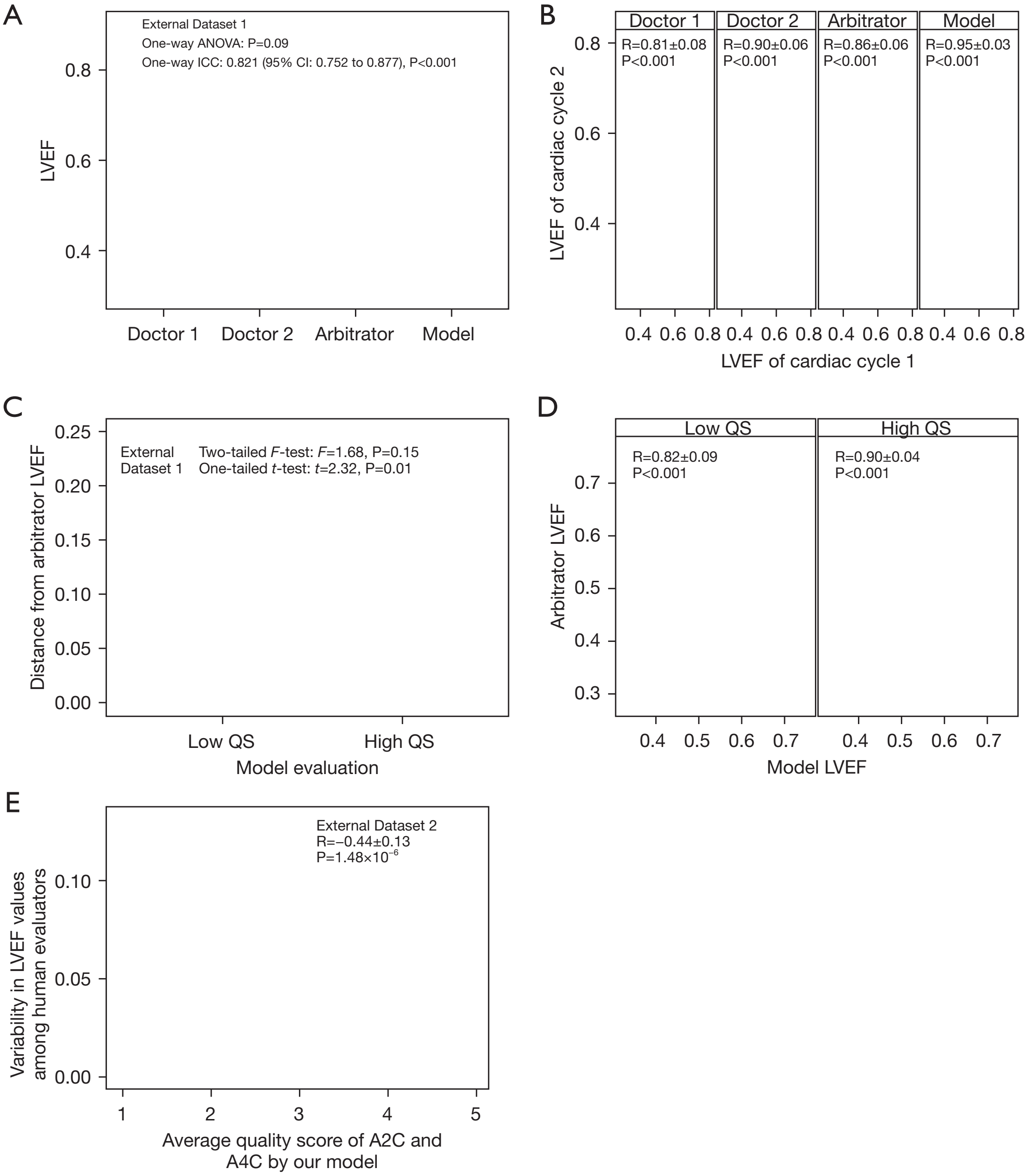


Figure 7 Model performance in LVEF calculation. (A) Boxplot of LVEF evaluated by the two doctors, one arbitrator, and our model. (B) The scatter plot and Pearson correlation of LVEF for cardiac cycles 1 and 2. The correlation coefficient (R) and statistical P values (P) are displayed at the top of the graph. (C) The boxplot indicates the distance between LVEF estimations made by our model and those by the arbitrator in echocardiogram with low- and high-quality scores. (D) The scatter plot and Pearson correlation of LVEF from our model and the arbitrator in echocardiogram with low- and high-quality scores. (E) The scatter plot and Pearson correlation between the variability in LVEF assessments by two doctors and the image quality scores assigned by our model within External Dataset 2. The numbers accompanied by "±" represent the correlation coefficient ± 95% CI. LVEF, left ventricular ejection fraction; ANOVA, analysis of variance; ICC, intraclass correlation coefficient; QS, quality score; A2C, apical 2-chamber; A4C, apical 4-chamber; CI, confidence interval.

model and that of the arbitrator. Our analysis revealed that when making LVEF assessments using echocardiograms with high-quality scores, the model exhibited a significantly closer distance (P value of 0.01 as determined by a *t*-test) (*Figure 7C*) and a higher correlation coefficient (0.90±0.04) (*Figure 7D*) with human experts.

Furthermore, an additional external test set of 110 participants was used to assess the impact of our quality scoring system on the LVEF evaluation process conducted by human experts (*Figure 7E*). The absolute difference in LVEF assessments between the two doctors was calculated and found to be significantly negatively correlated with the image quality scores determined by our model. This correlation was quantified with a Pearson's Correlation coefficient of −0.44, which falls within a 95% CI ranging from -0.58 to −0.28. Samples with higher quality scores exhibited a tendency towards smaller variations in LVEF assessments by the two doctors.

Discussion

The present study represents a significant stride in the realm of echocardiography by introducing an automated and efficient quality scoring system coupled with LVEF calculation, powered by DL techniques. Our proposed model has undergone extensive training and validation on a diverse dataset of 2,461 of participants, encompassing a wide spectrum of cardiac conditions and image qualities. Additionally, we conducted a performance comparison study involving 175 participants within External Dataset 2, pitting our model against human experts to assess its clinical utility and accuracy.

This study boasts 2 standout achievements: the development of a 5-point quality scoring system and the provision of an automated real-time scoring and LVEF calculation tool. Among the parameters derived from echocardiography, LVEF holds paramount significance as it provides quantitative information regarding the heart's pumping efficiency. Nevertheless, achieving a reliable estimation of LVEF poses a significant challenge owing to the considerable variability in echocardiography quality and cardiovascular anatomical structures. Historically, both the traditional quality assessment and LVEF calculations have been labor-intensive processes relying on the expertise of highly skilled human experts. The introduction of a DL-driven, automated quality scoring system, and LVEF calculation addresses the subjectivity and inefficiencies inherent in manual assessments.

There have been reports on automatic echocardiographic quality assessment and computer-assisted LVEF calculation. The majority of these studies employ image-level classification strategies for echocardiographic quality assessment. Labs *et al.* reported a multi-classification model designed to assess the echocardiographic frames for tasks involving depth gain, chamber clarity, on-axis orientation, and foreshortening (6). Huang *et al.* developed a qualified scoring model based on the classification of echocardiography views, leveraging the capabilities of the DenseNet-121 convolutional neural network (8). Luong *et al.* constructed a clarity classification model for a series of temporal echocardiographic frames, utilizing a long short-term memory (LSTM) module in conjunction with a DenseNet convolutional neural network as the feature extraction backbone (5). These solutions are not well suited for real-time clinical quality assessment for echocardiography due to their large model size and computationally intensive calculations. Furthermore, another significant challenge lies in the "black box" problem associated with image-level DL classification methods, which further constrains their practical clinical application (15). In the majority of studies involving computer-assisted LVEF estimation, the preferred approach is the biplane Simpson's method, which utilizes the cardiac chamber segmentation from both the A2C and A4C views. These studies have leveraged DL models to automate LV segmentation, subsequently applying Simpson's formulas to assess LVEF (16-24). Additionally, several studies have designed a 4-level EF classifier on three-dimensional (3D) echocardiography (25) and 2D echocardiographic videos (26). To enhance robustness, Zhang *et al.* introduced an analytical framework that integrates view classification, cardiac chamber segmentation, and LVEF calculation (16). Similarly, Smistad *et al.* incorporated an apical foreshortening detection module into the LVEF calculation pipeline (19). However, none of these approaches possess the capability to conduct quality control before estimating LVEF.

The 5-point quality scoring system we introduce primarily focuses on evaluating the integrity and clarity of cardiac anatomical structures within the echocardiogram (*Figure 1*). Considering the need for high-efficiency calculations in clinical applications, our automatic quality scoring model has been designed with lightweight and efficient DL modules. To illustrate, YOLOX represents an enhanced version of the You Only Look Once (YOLO) real-time object detection algorithm (11). We have integrated part of YOLOX modules into our automated quality scoring model, thereby achieving a more optimal trade-off between predictive accuracy and processing speed (*Figure 2*). To achieve real-time quality control, we have utilized the NVIDIA TensorRT module (NVIDIA, Santa Clara, CA, USA) to accelerate the algorithm's inference. TensorRT is a high-performance DL inference optimizer and runtime engine that enables our model to run efficiently on graphics processing units (GPUs). The model's processing speed was measured at an impressive 3 milliseconds per frame during testing on the NVIDIA Jetson Orin NX edge computing module. This translates to virtually no delay in our quality scoring model when deployed on most echocardiography equipment.

Our model was trained using a multicenter dataset that included data from commonly used echocardiogram devices such as Philips, GE, and Mindray. This ensures a broad applicability of our model across different clinical settings. Our model has demonstrated high performance across a wide range of tasks, boasting an impressive LV segmentation IoU score exceeding 0.8, an outstanding mAP50 score of 0.91, and an accuracy score of 0.96 on views classification. We have observed that the IoU score for MV and TV detection falls below 0.7 in both A2C and A4C views, as illustrated in *Figure 4E,4F*. IoU calculates the ratio of the area of overlap between the predicted and ground truth bounding boxes and is a critical metric in the field of object detection. Given that both the MV and TV are in continuous motion throughout the cardiac cycle, manual annotation as well as model-based detection pose significant challenges. Hence, we consider that a low IoU score in the MV and TV detection is acceptable, given that the quality score calculation is based on the presence of targets rather than their precise positional information. In the comparative performance analysis between our model and human experts, we found a high level of agreement in the quality assessments of echocardiograms from 65 participants. This agreement was reflected in both a similar distribution (*Figure 5*) and comparable discrepancies between the model and human experts, as well as among the human experts themselves (*Figure 6*). When it comes to LVEF estimation, although there were no statistically significant differences in the LVEF calculations between the human experts and our model (*Figure 7A*), it is worth noting that the model demonstrated a stronger correlation in LVEF estimation between cardiac cycle 1 and 2 (*Figure 7B*). Previous research has shown a roughly 10% mean absolute deviation (MAD) in the manual evaluation of LVEF (27). In contrast, our model demonstrates an even more minimal variance when compared to human experts, with a 5% MAD (*Figure 7C*). This observation suggests that utilizing our model for LVEF estimation can yield more objective and consistently stable results. Consistent with prior research (28,29), we also identified the persistent impact of echocardiography quality on LVEF estimation (*Figure 7C,7D*). This underscores the critical importance and essential need for quality control within the LVEF assessment procedure. In our 5-point quality scoring system, we recommend that the operator utilize automated LVEF calculation when the quality score exceeds 3.0.

However, this work still has several limitations that require further enhancement. Firstly, the transition from research findings to clinical implementation requires rigorous validation in diverse clinical settings and among varied patient populations. Although our findings are promising, further validation in real-world clinical scenarios remains essential. Secondly, the current model is equipped with quality scoring modules for only A2C and A4C views, and the development of quality scoring modules for other echocardiography views is currently in process. Lastly, the limited number of participants in a single-center performance comparison study may have introduced bias. Additional efforts should be made to include a larger and more diverse participant pool in future studies.

Conclusions

This research represents a substantial stride in the realm of echocardiography, offering an automated, efficient, and real-time quality scoring system coupled with LVEF estimation through DL. Leveraging a comprehensive dataset encompassing diverse cardiac conditions and echocardiography qualities, the model exhibited its proficiency and reproducibility in not only assessing quality but also estimating LVEF with precision. This breakthrough holds significant promise for augmenting clinical workflows and expediting the formulation of timely diagnoses and treatment strategies for CVDs.

Acknowledgments

Funding: This research was supported by the National Natural Science Foundation of China (No. 82371963 to

H.F.), the Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515011366 to H.F.), the NSFC launching fund of the Guangdong Provincial People's Hospital (Nos. 8207070477 and 8227070211 to H.F.), and the Commission of Science and Technology of Shenzhen (No. GJHZ20210705142204014 to B.S.).

Footnote

Reporting Checklist: The authors have completed the TRIPOD+AI reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-24-512/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-512/coif). Y.J. is from Shenzhen Yingzhichuangsi Technology Co., Ltd. C.C. is from Shenzhen MicroPort Xinsuanzi Medical Technology Co., Ltd. B.S. reports that this work was supported by the Commission of Science and Technology of Shenzhen (No. GJHZ20210705142204014). H.F. reports that this research was supported by the National Natural Science Foundation of China (No. 82371963), the Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515011366), and the NSFC launching fund of the Guangdong Provincial People's Hospital (Nos. 8207070477 and 8227070211). The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This study was approved by the Research Ethics Committee of Guangdong Provincial People's Hospital (No. QX2023-041-02). The requirement for informed consent was waived since this was a retrospective and observational study. All participating hospitals were informed of and agreed to the study, and all were subject to ethical approval through application.

References

1. Report on Cardiovascular Health and Diseases in China 2021: An Updated Summary. Biomed Environ Sci 2022;35:573-603.
2. Lee SH, Park JH. The Role of Echocardiography in Evaluating Cardiovascular Diseases in Patients with Diabetes Mellitus. Diabetes Metab J 2023;47:470-83.
3. Kosaraju A, Goyal A, Grigorova Y, Makaryus AN. Left Ventricular Ejection Fraction. 2024.
4. Dieden A, Carlson E, Gudmundsson P. Learning echocardiography- what are the challenges and what may favour learning? A qualitative study. BMC Med Educ 2019;19:212.
5. Luong C, Liao Z, Abdi A, Girgis H, Rohling R, Gin K, Jue J, Yeung D, Szefer E, Thompson D, Tsang MY, Lee PK, Nair P, Abolmaesumi P, Tsang TSM. Automated estimation of echocardiogram image quality in hospitalized patients. Int J Cardiovasc Imaging 2021;37:229-39.
6. Labs RB, Zolgharni M, Loo JP. Echocardiographic Image Quality Assessment Using Deep Neural Networks. In: Papież BW, Yaqub M, Jiao J, Namburete AIL, Noble JA, editors. Medical Image Understanding and Analysis. Cham: Springer International Publishing; 2021:488-502.
7. 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 J Biomed Health Inform 2020;24:931-42.
8. Huang KC, Huang CS, Su MY, Hung CL, Ethan Tu YC, Lin LC, Hwang JJ. Artificial Intelligence Aids Cardiac Image Quality Assessment for Improving Precision in Strain Measurements. JACC Cardiovasc Imaging 2021;14:335-45.
9. Liao Z, Girgis H, Abdi A, Vaseli H, Hetherington J, Rohling R, Gin K, Tsang T, Abolmaesumi P. On Modelling Label Uncertainty in Deep Neural Networks: Automatic Estimation of Intra- Observer Variability in 2D Echocardiography Quality Assessment. IEEE Trans Med Imaging 2020;39:1868-83.
10. American College of Emergency Physicians. Emergency Ultrasound Standard Reporting Guidelines. 2018. Available online: https://www.acep.org/siteassets/uploads/

uploaded-files/acep/clinical-and-practice-management/policy-statements/information-papers/emergency-ultrasound-standard-reporting-guidelines---2018.pdf

11. Ge Z, Liu S, Wang F, Li Z, Sun J. YOLOX: Exceeding YOLO Series in 2021. arXiv:2107.08430. 2021. Available online: https://arxiv.org/abs/2107.08430
12. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, Flachskampf FA, Foster E, Goldstein SA, Kuznetsova T, Lancellotti P, Muraru D, Picard MH, Rietzschel ER, Rudski L, Spencer KT, Tsang W, Voigt JU. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. Eur Heart J Cardiovasc Imaging 2015;16:233-70.
13. Pan J, Yang X, Cai H, Mu B. Image noise smoothing using a modified Kalman filter. Neurocomputing 2016;173:1625-9.
14. Diedenhofen B, Musch J. cocor: a comprehensive solution for the statistical comparison of correlations. PLoS One 2015;10:e0121945.
15. Petch J, Di S, Nelson W. Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. Can J Cardiol 2022;38:204-13.
16. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, Lassen MH, Fan E, Aras MA, Jordan C, Fleischmann KE, Melisko M, Qasim A, Shah SJ, Bajcsy R, Deo RC. Fully Automated Echocardiogram Interpretation in Clinical Practice. Circulation 2018;138:1623-35.
17. Jafari MH, Girgis H, Van Woudenberg N, Liao Z, Rohling R, Gin K, Abolmaesumi P, Tsang T. Automatic biplane left ventricular ejection fraction estimation with mobile point-of-care ultrasound using multi-task learning and adversarial training. Int J Comput Assist Radiol Surg 2019;14:1027-37.
18. Behnami D, Liao Z, Girgis H, Luong C, Rohling R, Gin K, Tsang T, Abolmaesumi P. Dual-view joint estimation of left ventricular ejection fraction with uncertainty modelling in echocardiograms. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap PT, Khan A, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Cham: Springer International Publishing; 2019:696-704.
19. Smistad E, Ostvik A, Salte IM, Melichova D, Nguyen TM, Haugaa K, Brunvand H, Edvardsen T, Leclerc S, Bernard O, Grenne B, Lovstakken L. Real-Time Automatic Ejection Fraction and Foreshortening Detection Using Deep Learning. IEEE Trans Ultrason Ferroelectr Freq Control 2020;67:2595-604.
20. Liu X, Fan Y, Li S, Chen M, Li M, Hau WK, Zhang H, Xu L, Lee AP. Deep learning-based automated left ventricular ejection fraction assessment using 2-D echocardiography. Am J Physiol Heart Circ Physiol 2021;321:H390-9.
21. Jafari MH, Van Woudenberg N, Luong C, Abolmaesumi P, Tsang T. Deep Bayesian image segmentation for a more robust ejection fraction estimation. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). IEEE; 2021:1264-8.
22. Tromp J, Seekings PJ, Hung CL, Iversen MB, Frost MJ, Ouwerkerk W, Jiang Z, Eisenhaber F, Goh RSM, Zhao H, Huang W, Ling LH, Sim D, Cozzone P, Richards AM, Lee HK, Solomon SD, Lam CSP, Ezekowitz JA. Automated interpretation of systolic and diastolic function on the echocardiogram: a multicohort study. Lancet Digit Health 2022;4:e46-54.
23. Dai W, Li X, Ding X, Cheng KT. Cyclical Self-Supervision for Semi-Supervised Ejection Fraction Prediction From Echocardiogram Videos. IEEE Trans Med Imaging 2023;42:1446-61.
24. Li H, Wang Y, Qu M, Cao P, Feng C, Yang J. EchoEFNet: Multi-task deep learning network for automatic calculation of left ventricular ejection fraction in 2D echocardiography. Comput Biol Med 2023;156:106705.
25. Silva JF, Silva JM, Guerra A, Matos S, Costa C. Ejection fraction classification in transthoracic echocardiography using a deep learning approach. In: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS). IEEE; 2018:123-8.
26. Kazemi Esfeh MM, Luong C, Behnami D, Tsang T, Abolmaesumi . A deep Bayesian video analysis framework: towards a more robust estimation of ejection fraction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer International Publishing; 2020:582-90.
27. Liu WY, Zhang PF, Gao YF, Chen X, Lin XX, Yang FF, Wang AA, He KL. Influence of echocardiographic image quality on the quality of left ventricular endocardial border delineation. Academic Journal of Chinese Pla Medical School 2022;43:855-61.
28. Sveric KM, Botan R, Dindane Z, Winkler A, Nowack T, Heitmann C, Schleußner L, Linke A. Single-Site Experience with an Automated Artificial Intelligence

Application for Left Ventricular Ejection Fraction Measurement in Echocardiography. Diagnostics (Basel) 2023;13:1298.

29. He B, Kwan AC, Cho JH, Yuan N, Pollick C, Shiota T, Ebinger J, Bello NA, Wei J, Josan K, Duffy G, Jujjavarapu M, Siegel R, Cheng S, Zou JY, Ouyang D. Blinded, randomized trial of sonographer versus AI cardiac function assessment. Nature 2023;616:520-4.

Cite this article as: Li X, Liao L, Wu K, Meng AT, Jiang Y, Zhu Y, Cui C, Xu X, Shi B, Fei H. An automatic and real-time echocardiography quality scoring system based on deep learning to improve reproducible assessment of left ventricular ejection fraction. Quant Imaging Med Surg 2025;15(1):770-785. doi: 10.21037/qims-24-512