Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 270 (2025) 4917–4926

Procedia Computer Science

www.elsevier.com/locate/procedia

29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025)

Deep Learning for Quality Assessment of Echocardiographic Images

Shuping Kang, Yulia Hicks, Rossitza Setchi

Research Centre in AI, Robotics and Human-Machine Systems (IROHMS), Cardiff University, Cardiff CF24 3AA, UK

Abstract

The growing need for standardised and automated cardiac ultrasound (US) acquisition has driven the integration of deep learning into echocardiographic workflows. While existing deep learning (DL) models have shown promising results in tasks such as view classification and image quality assessment, most of these approaches focus either on differentiating among standard views or grading image quality within a standard view. However, these methods lack the capacity to model the sequential spatial transitions that occur during the acquisition process, limiting their applicability to real-time probe guidance and robotic control. To address this gap, we propose a classification framework designed for the process of acquiring the parasternal long-axis (PLAX) view under a fixed scanning protocol. Based on extensive probe movement experiments across multiple patients, we identified four representative echocardiographic views that appear during the search for the optimal PLAX position. These views correspond to distinct probe positions and orientations and reflect varying levels of image completeness. A dataset of 7,200 annotated images was used to train a ResNet50-based deep network for multi-class classification. The model achieved robust performance with accuracy, sensitivity, specificity, and F1 scores above 89%, and AUC exceeding 97% in patient-level cross-validation. It effectively captures spatially relevant features, distinguishes subtle view differences, and generalizes well to unseen data. The outputs provide interpretable feedback correlating image quality with probe position, enabling real-time scanning assessment. Furthermore, this work introduces a novel problem formulation and multi-class view classification under a fixed acquisition protocol. It provides a foundation for developing the next generation of intelligent US systems. By linking image classification to probe position and orientation, the proposed framework enables real-time feedback that can ultimately support autonomous scanning agents in locating diagnostically optimal cardiac views.

Keywords: echocardiographic images; image classification; deep learning; quality assessment; PLAX; 3-fold cross-validation.

10.1016/j.procs.2025.09.618

1. Introduction

Cardiovascular diseases (CVDs), including heart disease and heart attack, are the leading cause of global mortality as well as a major contributor to disability [1]. According to data from *Our World in Data*, CVDs accounted for approximately 18.56 million deaths in 2019, which is nearly twice the number of fatalities attributed to cancer, the second most common cause of death globally [2]. The diagnosis and management of CVDs often require sophisticated imaging techniques, among which US has become one of the most widely utilised, owing to its numerous advantages including non-invasiveness, absence of ionising radiation, cost-effectiveness, and real-time imaging capabilities [3][4]. Beyond its diagnostic value, US also plays a crucial role in guiding treatment and monitoring disease progression over time. As a result, cardiac US, also known as echocardiography, has become a routine and essential tool in cardiac examinations for assessing cardiac structure and function in clinical practice.

A key component of echocardiographic examination is the classification of different cardiac views, which enables clinicians to interpret different anatomical planes and assess specific cardiac conditions. However, due to individual anatomical variations, complex cardiac structures, and differences in imaging angles, the classification of cardiac views remains a challenging task. Furthermore, the quality of US images is highly dependent on the operator's experience and skill, often resulting in suboptimal or inconsistent views that can increase the risk of misdiagnosis [5]. The process of training sonographers to consistently capture standard views is time-consuming, and many practitioners suffer from musculoskeletal strain due to prolonged scanning sessions [6]. These limitations underscore the need for operator-independent solutions, especially in emergency or primary care settings, where expertise may be limited [7].

In this context, artificial intelligence (AI), and in particular deep learning (DL), offers promising solutions for automating echocardiographic analysis. Recent advancements in DL have enabled systems to assist in view classification, image quality assessment, cardiac structure segmentation, functional measurements, and even disease diagnosis [8]. Among these, view classification and image quality assessment are considered foundational tasks that support further downstream analysis. Nonetheless, training a robust AI system for automatic cardiac view recognition is challenging. Several factors affect image quality—including probe position, rotation angle, tilt angle, and contact force, while the structural complexity and continuous motion of the heart exacerbate classification difficulty.

Typical echocardiographic protocols involve acquiring standard views from parasternal, apical, and subcostal windows. Each view reflects different diagnostic aspects of cardiac function. For example, the apical view is critical for evaluating hemodynamic parameters such as diastolic dysfunction, valvular regurgitation, and cardiac output. The PLAX view, typically acquired first, provides a comprehensive assessment of cardiac structure, including overall left and right ventricular sizes and ejection fraction. It is particularly useful in assessing cardiac cavity enlargement, valve dynamics, and myocardial condition, the parasternal long axis view facilitates the diagnosis of conditions such as myocardial infarction, cardiomyopathy, valvular disease, and arrhythmia [9]. Thus, ensuring the acquisition of a high-quality PLAX view is essential in echocardiographic examinations.

However, most existing DL studies focus on either distinguishing between different standard views, or evaluating image quality within the same standard view, such as differentiating between incomplete, speckled, and complete views [10]. To the best of our knowledge, no prior work has addressed the problem of classifying US images acquired under a fixed scanning protocol during the process of searching for the optimal PLAX view.

In this study, we aim to fill this gap by applying a DL–based image classification task that reflects the spatial and qualitative progression toward the optimal PLAX view. Specifically, we conduct multiple echocardiographic acquisition experiments and identify four representative cardiac views commonly observed during the search for the long-axis view. These views correspond to different probe positions and orientations. We collect data across different patients and train a DL model to classify these four view types. Since each view is closely tied to the probe's spatial configuration, the resulting model can provide meaningful feedback for controlling probe motion. Ultimately, this work lays the foundation for training intelligent agents capable of autonomously acquiring high-quality PLAX views in future robotic US systems.

2. Literature review

The rapid development of machine learning (ML) techniques has significantly benefited the medical field by enhancing patient care, quicker healthcare services, and supporting clinical decision-making [10][11][12]. As a

powerful subset of ML, DL simulates human cognition by stacking simple functions to make complex decisions in a deep structure, enabling automated pattern recognition and classification tasks with high accuracy, which has proved to work well in the medical field. Early studies utilized convolutional neural networks (CNNs) such as AlexNet and VGGNet for classifying Computed tomography (CT), Magnetic Resonance Imaging (MRI), and X-ray images, achieving strong performance compared to traditional feature-engineering methods [13][14].

Despite its diagnostic advantages, US imaging, particularly echocardiography, presents unique challenges due to speckle noise, anatomical variability, and operator dependency. Nonetheless, DL models have proven effective in view classification. For example, Sudharson and Kokil proposed an ensemble deep neural network (DNN) model using transfer learning (TL) for automatic classification of B-mode kidney US images, achieving superior performance in detecting multiple kidney abnormalities [15]. Lazo et al. used VGG-16 and InceptionV3 architectures with transfer learning for detecting lesions in breast US images [16], while another study [20] categorized grayscale abdominal US images into 11 categories based on technologist-provided annotations. In the cardiac domain, Zhang et al. trained a convolutional neural network (CNN) with multiple tasks including automated identification of 23 viewpoints [17]. Gao et al. [18] propose a novel automatic recognition method including three effective strategies based on CNN to identify nine standard cardiac views. Kusunose et al. developed a CNN trained on a dataset containing mislabelled images that were not checked by observers and demonstrated its feasibility in clinical classification tasks [19].

Beyond classification, image quality assessment (IQA) has become increasingly important in both clinical and autonomous echocardiography systems. Abdi et al. [20] introduced a DL–based quality assessment method for apical four-chamber (A4C) echocardiograms by training a regression-based convolutional neural network to predict expert-assigned quality scores ranging from 0 to 5. Their model used particle swarm optimization (PSO) to fine-tune hyperparameters and achieved a mean absolute error of 0.71, matching expert intra-rater reliability. Unlike traditional handcrafted or template-based methods, their approach learned interpretable features directly from the image, enabling real-time, view-independent quality scoring. In a more recent study, Elmekki et al. [10] proposed a comprehensive framework that simultaneously performs cardiac view classification and quality grading of US images using a transfer learning-based DL model. Their work introduces CACTUS, the first publicly available dataset of graded cardiac US images, annotated by clinical experts based on completeness and clarity. The framework employs a shared ResNet18 encoder with two heads: one for view classification and another for quality regression, reducing computational cost while achieving high performance (classification accuracy of 99.43% and grading loss of 0.3067). Unlike earlier work focusing solely on view classification, this approach integrates image quality scoring, making it a valuable tool for real-time clinical feedback and autonomous acquisition systems.

DL-based techniques for view classification and quality assessment of cardiac US images are essential enablers of intelligent imaging systems. However, most existing studies either focus on view classification across distinct standard planes such as PLAX, PSAX, and apical views or evaluate image quality within a specific view [10] [20]. These approaches typically assume independent static images and are not designed to model the progressive image variations encountered during probe navigation under a fixed acquisition protocol. To the best of authors' knowledge, no existing work attempts to classify intermediate cardiac US frames collected sequentially during the search for an optimal PLAX view, where each class reflects different probe positions and orientations. Addressing this gap is critical for enabling autonomous scanning agents that require fine-grained spatial awareness of view transitions to reach diagnostically optimal imaging positions.

3. Methodology

The proposed methodology consists of three main stages: (i) data acquisition and labelling, (ii) image preprocessing and augmentation, and (iii) final classification into four quality levels. Each stage is designed to ensure robust, spatially informed categorization of echocardiographic views across different patients.

3.1. Data acquisition protocol and data description

All the echocardiographic imaging experiments were conducted on an intelligent hybrid software and hardware US simulator BODYWORKS|Eve, which is a high-fidelity, AI-powered US simulator allowing for realistic 3D

simulations of the organs like heart, liver, kidney, and their surrounding structures. For each organ, there is a patient list with different physical conditions. The experiments presented in this paper have used the data from eight patients with varying cardiac pathologies and one healthy individual. These individuals exhibited differences in age, gender, and pathology including but not limited to conditions such as low left ventricular failure, mitral stenosis, pulmonary embolism, and mitral regurgitation.

For each participant a standardized rectangular scanning area was delineated on the chest surface to ensure the inclusion of PLAX view (Fig. 1). Within the predefined region, a grid of scanning points was established to provide dense spatial coverage. Usually, optimal PLAX view is shown when the orientation of the marker on it directly towards the right shoulder. Alignment of the US beam parallel to an imaginary line drawn from the patient's right shoulder to their left hip is crucial [21]. Therefore, at each scanning point, the US probe was initially positioned perpendicular to the surface of the mannequin, with the marker directed cranially, aligning the image plane with the sagittal plane. The probe was then gradually rotated clockwise through 90 degrees, while the marker passed the patient's right shoulder, aligning the imaging plane with the horizonal plane. During this rotation, controlled tilting and rocking manoeuvres were applied to simulate realistic probe manipulation. Image acquisition was automated, with screenshots captured at 0.2-second intervals. Concurrently, the probe's spatial information, including Cartesian coordinates, rotation, and tilt angle was recorded for each image.

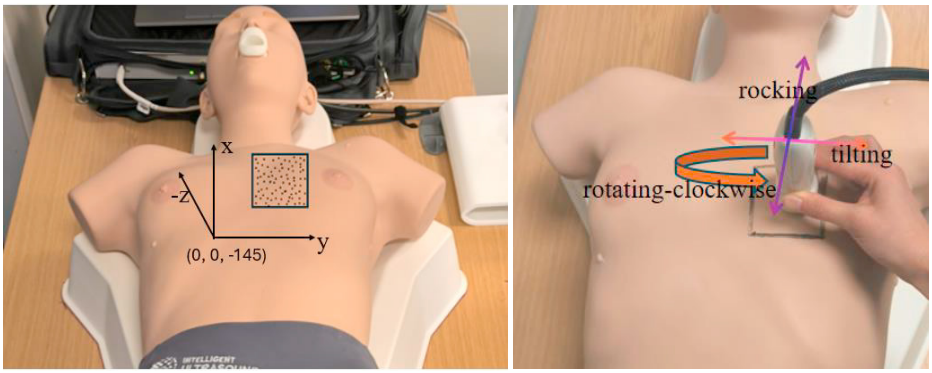

Fig. 1. The searching area and the directions of each axis (left) and data acquisition protocol (right).

The obtained dataset consists of 7,200 echocardiographic images collected from nine subjects. For each subject, 800 images were acquired, with 200 images assigned to each of four manually labelled categories: bad, not bad, good, and best. Fig. 2 illustrates representative examples from these four classes. In the figure, x, y, and z denote the Cartesian coordinates of the probe's position along the three spatial axes respectively. Meanwhile, A, B, and C represent the Euler angles of rotation around the x, y, and z axes, respectively, quantifying the probe's orientation in three-dimensional space. Figure 1 illustrates the coordinate system, where the x–y plane is parallel to the surface on which the patient lies (i.e., the horizontal plane of the body), and the z-axis is aligned with the direction of gravity. The classification into four categories was based on typical view types observed during the controlled probe movements described above. As the probe gradually rotated, tilted and rocked from the cranial to the horizontal orientation, distinct image patterns emerged naturally at different stages. These characteristic views were consistently identifiable across participants and were selected as representative categories.

The four categories, as illustrated in Fig. 2, represent distinct stages of image quality, each corresponding to specific characteristics of the probe's position and orientation. Image a shows an example of the "bad" category, where the cardiac shape appears smaller and more circular, with visible chambers that are misaligned, displaying an upper chamber smaller than the lower one. In contrast, image b exemplifies the "not bad" category, where the heart shape becomes larger and more elongated, and three chambers are visible in a tilted view, with the mitral valve more clearly discernible compared to the "bad" view. The "good" category in image c demonstrates a view closer to the "best" category but with noticeable differences: the aortic valve appears more blurred, and the heart in this image is more horizontally aligned, as opposed to the tilted view in the "not bad" category. Additionally, the chamber sizes are more varied, and the structural complexity increases. Finally, image d represents the "best" category, where both the aortic

valve (AV) and mitral valves (MV) are clearly visible, aligned slightly to the right of the centre of the display. In an ideal PLAX view, key anatomical structures, including the anterior and posterior mitral valve leaflets (AML, PML) and the aortic valve, are prominently discernible, providing an optimal and comprehensive view of the cardiac anatomy.

These four classes were not arbitrarily defined; rather, they were identified through extensive experiments and review as common view types associated with different spatial configurations of the US probe. Crucially, each image's quality level implicitly shows its spatial correlation from the optimal probe pose. Higher-learning-based are associated with probe poses that are closer to the optimal scanning configuration. Although expert annotations were not available, this labelling process was based on internal consistency and reproducible spatial cues observed across patients. Thus, the classification is not only clinically intuitive but also spatially informative, and the labelling strategy serves as a bridge between image content and the physical state of the probe.

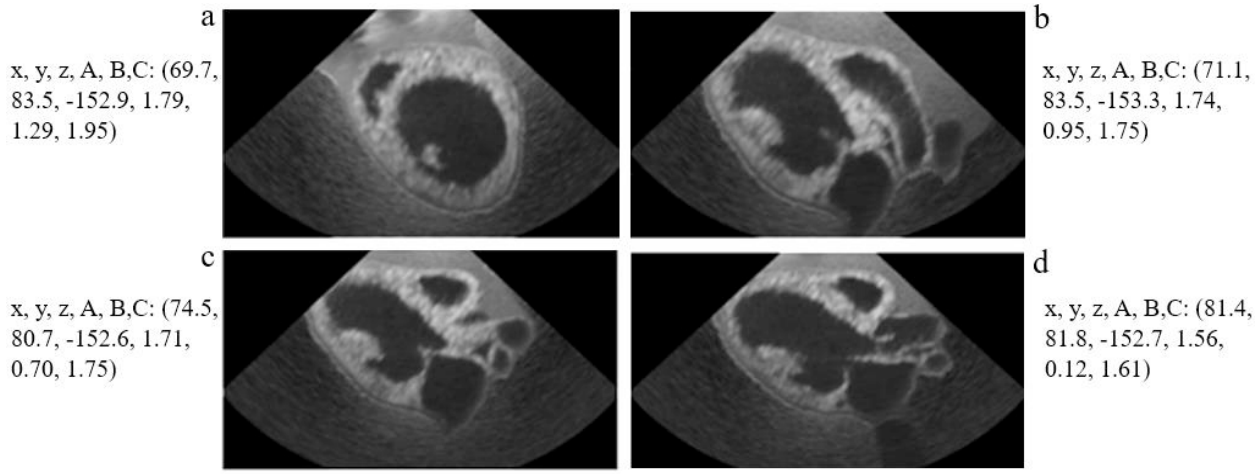

Fig. 2. Examples of the four categories: (a) bad; (b) not bad; (c) good; (d) best.

3.2. Data preprocessing and augmentation

To enhance the robustness and generalisation capability of the classification model, all echocardiographic images underwent a series of standardised preprocessing steps prior to model training. The raw images acquired during scanning sessions often include superimposed textual annotations (e.g., patient IDs, imaging settings) and are captured at high resolutions not optimized for deep neural networks. Initial cropping was applied to remove superimposed metadata, retaining only diagnostically relevant content. To preserve the original aspect ratio, each image was first resized with its longer side scaled to 256 pixels, followed by symmetric padding on the shorter side to obtain a uniform size of 256×256 pixels without geometric distortion. A centre crop was subsequently applied to produce input images of 224×224 pixels, aligning with the standard input size of the ResNet50 architecture. To promote training stability and accelerate convergence, all pixel intensities were normalized to have zero mean and unit variance based on the dataset's global statistics. Beyond normalization, data augmentation techniques were employed to improve the model's tolerance to acquisition variability and reduce overfitting. Specifically, augmentations included random brightness and contrast adjustments, and small-angle rotations (within ±10°). These transformations simulate the natural perturbations in probe positioning and orientation that may occur during manual or robotic scanning, thus helping the model learn more generalisable features. By integrating these preprocessing and augmentation strategies, the dataset was rendered more representative of real-world variability, which is critical for the eventual deployment of the trained system in autonomous and semi-autonomous US applications.

3.3. Network architecture

To explore the optimal network architecture for the proposed four-class echocardiographic image classification task, we evaluated a range of deep convolutional neural networks, including ResNet18, ResNet50, ResNet101, Inception_v3, and DenseNet121. These models were selected due to their proven effectiveness in medical image analysis and their architectural diversity, offering a comprehensive performance comparison across varying depths, connectivity patterns, and computational costs. Among them, ResNet50 was ultimately chosen as the backbone network. Compared with its shallower counterpart ResNet18, ResNet50 offers a greater capacity for feature representation due to its increased depth, while still maintaining manageable computational complexity. Although deeper models such as ResNet101 or denser architectures like DenseNet121 demonstrated competitive performance, ResNet50 achieved a superior balance between classification accuracy and model efficiency, without incurring significant overfitting or training instability.

ResNet50 architecture [22] consists of 50 convolutional layers organized with residual connections, which allow for efficient gradient propagation and improved convergence when training deep networks from scratch. A schematic overview of architecture is presented in Fig. 3. To apply the model to the specific classification task, the original fully connected layer was replaced with a new dense layer containing four output units, each corresponding to one of the predefined quality categories (bad, not bad, good, and best). A SoftMax activation function was employed at the output layer to produce normalised probability distributions across the four classes.

The model was trained end-to-end using the categorical cross-entropy loss function, which is well-suited for multi-class classification tasks. Optimization was performed using the Adam optimizer with an initial learning rate of 1e-4, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. A mini-batch size of 198 was used to balance GPU memory constraints and convergence speed. A learning rate scheduler was employed to reduce the learning rate by a factor of 0.1 upon plateauing of validation loss, which helped prevent overfitting in later stages of training.

Training was conducted for 50 epochs, as empirical observations indicated that model performance tended to plateau around this point; extending training beyond 50 epochs led to negligible improvements or even slight degradation in classification accuracy. On average, each fold in the three-fold cross-validation setting required approximately 1.5 hours to complete training. In addition, data augmentation strategies such as random rotation, and brightness adjustment were applied to reduce overfitting and enhance generalization and robustness.

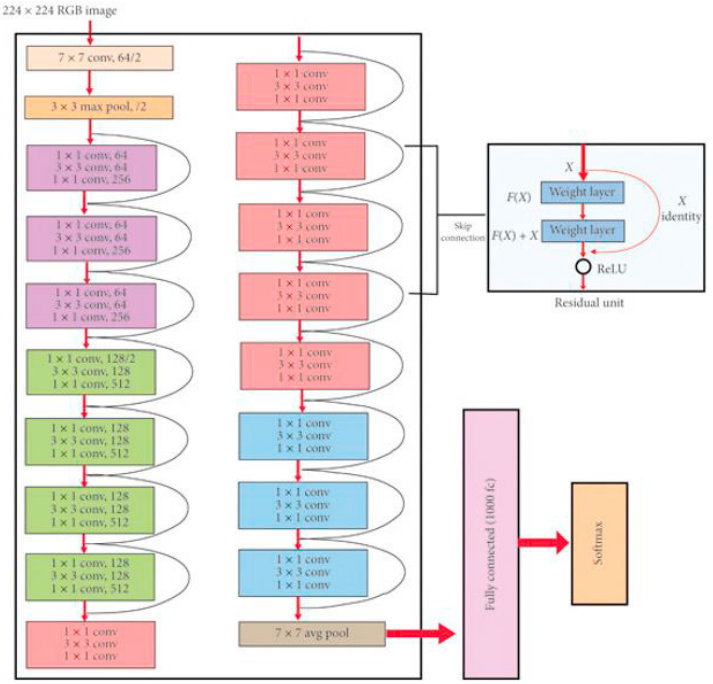

Fig. 3. ResNet50 architecture with internal layer wise detail [23].

4. Experimental Setup

In this section, the experiment is conducted on a high-performance computer running Ubuntu operating system, from where all results are obtained.

4.1. Hardware and software environment

The experiments were executed in Ubuntu 20.04.4 LTS Operating system on a machine with CPU Intel(R) Core(TM) i9-10900XCPU @ 3.70GHz*20, GPU NVIDA Corporation TU102[GeForce RTX 2080 Ti Rev.A] and RAM 125.5 GiB. The Anaconda3-x86_64-conda-linux-gnu and Pycharm-community-2024.2 versions are used as a software platform for simulations.

4.2. *Cross-patient data splitting strategy*

To rigorously evaluate the model's ability to generalise across different patients, a three-fold cross-validation strategy at the patient level is employed. In each fold, images from six patients were used for training, while the remaining three patients were held out as the test set. No separate validation set was used; instead, evaluation metrics were computed directly on the independent test patients in each fold.

To prevent information leakage and ensure a fair assessment of generalisation, the dataset was explicitly divided by patient identity. That is, all images belonging to a given patient appear exclusively in either the training or test set in any given fold, but no subject contributed data to both sets simultaneously. Specifically, the dataset comprises echocardiographic images collected from nine individuals, including eight patients with diverse cardiac conditions and one healthy individual. These individuals were randomly grouped into three folds, each containing three patients. The cross-validation procedure iteratively trained the model on six patients and evaluated it on the held-out three, rotating the fold assignment each time. This design enables evaluation across the full set of patient variations, including differing anatomies, probe handling patterns, and disease manifestations.

4.3. Evaluation metrics

Specific metrics were documented to evaluate the classification task performance, including accuracy (precision), sensitivity (recall), specificity, F1 score, and the area under the receiver operating characteristic (ROC) curve (AUC). TP, TN, FP, FN represent the number of predicted true positive, true negative, false positive, false negative samples respectively. The performance metrics are given by:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN). \tag{1}$$

$$\text{Sensitivity (Recall)} = TP / (TP + FN). \tag{2}$$

$$\text{Specificity} = TN / (TN + FP). \tag{3}$$

$$\text{F1 Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}), \text{ where Precision} = TP / (TP + FP). \tag{4}$$

To evaluate the model across the full dataset, these metrics were applied in each fold of the three-fold cross validation and reported the average performance. Furthermore, to provide a visual interpretation of class separability multi-class ROC curves were generated for each class to visualize class-wise discrimination, and the area under the ROC curve (AUC) was calculated to summarize model separability.

4.4. Experimental results and analysis

All quantitative results, including per-fold and averaged scores of ResNet50, are summarized in Table 1, with multi-class ROC curves of ResNet50 shown in Figure 4. Table 2 further compares average performance across different models, providing additional insight into model robustness and generalizability.

The proposed model achieved strong and consistent performance across all three folds of patient-level cross-validation. While Fold 1 yielded relatively lower scores compared to the other folds, all key metrics, including accuracy, sensitivity, specificity, F1 score, remained above 0.80, reflecting a reliable baseline level of performance. Fold 2, on the other hand, exhibited the best results, with all metrics exceeding 0.90, highlighting the model's potential to achieve excellent classification under certain patient distributions. The variation in performance across folds, most notably between Fold 1 and Fold 2, which may be attributed to inter-subject anatomical differences and the subjective nature of image quality labelling. Additionally, the limited dataset size means that the data from individual patients may disproportionately influence fold-specific results. Therefore, further validation on larger and more diverse datasets, including external validation cohorts, is necessary to comprehensively assess the model's robustness and generalizability. Nonetheless, the model exhibited stable accuracy and balanced F1 scores across folds, suggesting good generalization ability and low sensitivity to differences in patient anatomy or labelling bias. High sensitivity and specificity further indicate that the model can effectively identify target views while minimizing false detections, a crucial requirement for clinical deployment.

ROC curves exhibited strong separability among classes, with AUC values exceeding 0.90 in all cases. This reflects the model's high discriminative power in distinguishing between the four defined quality categories. Such high AUC scores are particularly encouraging given the fine-grained and subjective nature of the classification task, which inherently involves subtle visual distinctions.

To better understand model behavior, we examined the confusion matrix of the best-performing fold. Results showed that most misclassifications occurred between adjacent quality levels, especially between Class 1 (not bad) and Class 2 (good), where 43 and 36 samples were confused, respectively. This reflects the gradual visual transition between these classes and supports the spatial continuity assumption of the acquisition process. In contrast, Class 0 (bad) and Class 3 (best) were more distinct, with minimal confusion, indicating that the model effectively captures the endpoints of the view quality spectrum.

To select the optimal backbone, we compared five widely used CNN architectures: ResNet18, ResNet50, ResNet101, DenseNet121, and Inception_v3. As shown in Table 2, ResNet50 achieved the highest accuracy (0.8989), F1 score (0.8990), and a strong AUC (0.9771), indicating superior discriminative ability. While DenseNet121 reported a slightly higher AUC (0.9814), its lower accuracy (0.8893) and sensitivity (0.8893) suggest less consistent performance. ResNet18 and ResNet101 were limited by underfitting and overfitting tendencies, respectively. Inception_v3, despite its architectural sophistication, yielded the weakest results across all metrics, likely due to over-complexity and insufficient data support. These findings justify the use of ResNet50, which offers a strong trade-off between depth, complexity, and generalization. Overall, the results validate the model's ability to generalize across unseen patients and its suitability as a perception module in autonomous US systems, where accurate view recognition is essential for closed-loop control of probe positioning.

Table 1. 3-fold cross-validation evaluation results (ResNet50).

Fold	Accuracy	Sensitivity	Specificity	F1 score
1	0.8424	0.8400	0.9474	0.8416
2	0.9403	0.9426	0.9802	0.9415
3	0.9147	0.9139	0.9714	0.9138
Average	0.8989	0.8988	0.9663	0.8990

Table 2. Performance comparison of different CNN architectures on the validation set.

Model	Accuracy	Sensitivity	Specificity	F1 score	AUC
Resnet18	0.8751	0.8762	0.9550	0.8801	0.9762
Resnet50	0.8989	0.8988	0.9663	0.8990	0.9771
Resnet101	0.8784	0.8790	0.9595	0.8805	0.9789
Inception_v3	0.8073	0.8029	0.9345	0.8003	0.9351
Densenet121	0.8893	0.8893	0.9631	0.8912	0.9814

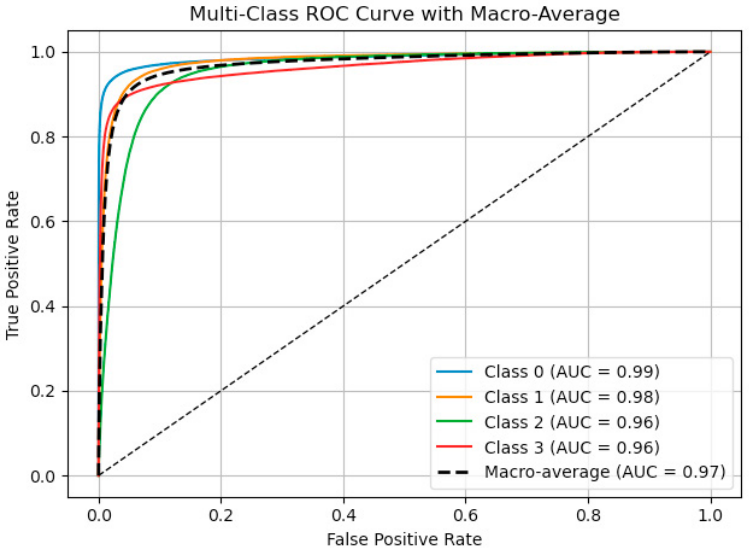

Fig. 4. Multi-class ROC curves (ResNet50).

5. Conclusions and future work

This study introduces a DL framework for automated quality assessment of echocardiographic images, grounded in a carefully designed acquisition protocol and a clinically relevant view taxonomy. By leveraging ResNet50, the system achieves strong performance in distinguishing view quality levels that correspond to varying probe poses. One of the central contributions of this work lies in the identification and formalization of four representative view types during the search for the PLAX view. These categories not only reflect diagnostic image quality but also reflect implicit information about probe pose relative to the ideal scanning position. This labelling strategy bridges the gap between image perception and physical probe configuration, enabling meaningful guidance for robotic navigation.

Despite these promising results, some limitations should be acknowledged. The dataset comprises images from nine subjects, which are collected under controlled conditions using a simulator. While the class distribution was balanced within subjects, the limited sample size and lack of clinical diversity may reduce the generalizability of the findings to real-world settings. Future work will therefore focus on validating the framework on larger and more heterogeneous datasets, including data acquired from human subjects across diverse clinical environments and imaging systems. To further improve label consistency, we also plan to involve expert annotators and assess inter-rater reliability to refine category definitions.

In parallel, we aim to extend the proposed framework to closed-loop robotic US systems. Specifically, the trained classifier will be incorporated as part of the reward function in reinforcement learning algorithms to enable autonomous probe navigation. By integrating real-time feedback, the system can iteratively adjust probe orientation and pressure to acquire diagnostic optimal PLAX views. Furthermore, we will explore domain adaptation techniques to enhance model robustness across varying patient anatomies and US devices.

Acknowledgements

The authors gratefully acknowledge Zebang Liu for his valuable assistance in code debugging and methodological suggestions. Financial support from the China Scholarship Council (CSC) during the PhD study is also sincerely appreciated.

References

[1] Roth, Gregory A., George A. Mensah, Catherine O. Johnson, Giovanni Addolorato, Enrico Ammirati, Larry M. Baddour, Noël C. Barengo et al. (2020) "Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study." Journal of the American college of cardiology 76, no. 25: 2982-3021.

[2] Ritchie, H., Spooner, F. and Roser, M. (2018). Causes of death. Our World in Data. Available at: https://ourworldindata.org/causes-of-death.

[3] Shung, K. Kirk. (2011) "Diagnostic ultrasound: Past, present, and future." J Med Biol Eng 31, no. 6: 371-4.

[4] Li, Keyu, Yangxin Xu, and Max Q-H. Meng. (2021) "An overview of systems and techniques for autonomous robotic ultrasound acquisitions." IEEE Transactions on Medical Robotics and Bionics 3, no. 2: 510-524.

[5] Berg, Wendie A., Jeffrey D. Blume, Jean B. Cormack, and Ellen B. Mendelson. (2006) "Operator dependence of physician-performed whole-breast US: lesion detection and characterization." Radiology 241, no. 2: 355-365.

[6] Brown, Grahame. (2003) "Work related musculoskeletal disorders in sonographers." BMUS Bulletin 11, no. 3: 6-13.

[7] Schneider, Matthias, Philipp Bartko, Welf Geller, Varius Dannenberg, Andreas König, Christina Binder, Georg Goliasch, Christian Hengstenberg, and Thomas Binder. (2021) "A machine learning algorithm supports ultrasound-naïve novices in the acquisition of diagnostic echocardiography loops and provides accurate estimation of LVEF." The International Journal of Cardiovascular Imaging 37: 577-586.

[8] Gao, Yanhua, Yuan Zhu, Bo Liu, Yue Hu, Gang Yu, and Youmin Guo. (2021) "Automated recognition of ultrasound cardiac views based on deep learning with graph constraint." Diagnostics 11, no. 7: 1177.

[9] Shida, Yuuki, Souto Kumagai, Ryosuke Tsumura, and Hiroyasu Iwata. (2023) "Automated image acquisition of parasternal long-axis view with robotic echocardiography." IEEE Robotics and Automation Letters 8, no. 8: 5228-5235.

[10] Elmekki, H., Alagha, A., Sami, H., Spilkin, A., Zanuttini, A. M., Zakeri, E., ... & Mourad, A. (2025). "CACTUS: An open dataset and framework for automated Cardiac Assessment and Classification of Ultrasound images using deep transfer learning." Computers in Biology and Medicine, 190, 110003.

[11] Javaid, Mohd, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, and Shanay Rab. (2022) "Significance of machine learning in healthcare: Features, pillars and applications." International Journal of Intelligent Networks 3: 58-73.

[12] May, Mike. (2021) "Eight ways machine learning is assisting medicine." Nat Med 27: 2-3.

[13] Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I. Sánchez. (2017) "A survey on deep learning in medical image analysis." Medical image analysis 42: 60-88.

[14] Shen, Dinggang, Guorong Wu, and Heung-Il Suk. (2017) "Deep learning in medical image analysis." Annual review of biomedical engineering 19, no. 1: 221-248.

[15] Sudharson, S., and Priyanka Kokil. (2020) "An ensemble of deep neural networks for kidney ultrasound image classification." Computer Methods and Programs in Biomedicine 197: 105709.

[16] Lazo, Jorge F., Sara Moccia, Emanuele Frontoni, and Elena De Momi. (2020) "Comparison of different CNNs for breast tumor classification from ultrasound images." arXiv preprint arXiv:2012.14517.

[17] Zhang, Jeffrey, Sravani Gajjala, Pulkit Agrawal, Geoffrey H. Tison, Laura A. Hallock, Lauren Beussink-Nelson, Mats H. Lassen et al. (2018) "Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy." Circulation 138, no. 16: 1623-1635.

[18] Gao, Yanhua, Yuan Zhu, Bo Liu, Yue Hu, Gang Yu, and Youmin Guo. (2021) "Automated recognition of ultrasound cardiac views based on deep learning with graph constraint." Diagnostics 11, no. 7: 1177.

[19] Kusunose, Kenya, Akihiro Haga, Mizuki Inoue, Daiju Fukuda, Hirotsugu Yamada, and Masataka Sata. (2020) "Clinically feasible and accurate view classification of echocardiographic images using deep learning." Biomolecules 10, no. 5: 665.

[20] Abdi, Amir H., Christina Luong, Teresa Tsang, Gregory Allan, Saman Nouranian, John Jue, Dale Hawley et al. (2017) "Automatic quality assessment of echocardiograms using convolutional neural networks: feasibility on the apical four-chamber view." IEEE transactions on medical imaging 36, no. 6: 1221-1230.

[21] U. F. O. Themes. (2019), ‘‘Cardiac Ultrasound Technique.’’ Radiology Key. [Online]. Available: https://radiologykey.com/cardiac-ultrasound-technique/.

[22] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition. (2016)" In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.

[23] Shabbir, Amsa, Nouman Ali, Jameel Ahmed, Bushra Zafar, Aqsa Rasheed, Muhammad Sajid, Afzal Ahmed, and Saadat Hanif Dar. (2021) "Satellite and scene image classification based on transfer learning and fine tuning of ResNet50." Mathematical Problems in Engineering 2021, no. 1: 5843816.