Malarkodi G, 2026, 14:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

EchoGuide AI: An Uncertainty-Aware Deep Learning Pipeline for Echocardiographic View Classification and Cardiac Structure Segmentation, with an Evaluation of Label-Free Quality Scoring and of the Data Barrier to Probe Guidance

Malarkodi G[1], R.K. Soundarayaa[2], and S.M. Madhumathi[3]
P.S.R. Engineering College, Sivakasi, Tamil Nadu 626140, India

Abstract- Background. Transthoracic echocardiography depends heavily on the operator. A recent multi-task framework proposed deriving image quality from the predictive uncertainty of view-classification and segmentation networks, and recommending corrective probe maneuvers [1]. Its authors reported training and validating the system on a 43-case maneuver-tagged dataset that they constructed; we did not find a public release of these data, which prevents independent reproduction of the maneuver component. Objective. To build an open, reproducible counterpart on public data and to establish which parts of such a system public data can actually support. Methods. We used the public CAMUS dataset [2]: 500 patients and 1,998 expert-annotated apical two- and four-chamber (A2C, A4C) frames at end-diastole and end-systole. We partitioned it at subject level into 300/100/100 patients with a fixed seed and an automated leakage check. View classifier: an ImageNet-initialised ResNet-18 separating A2C from A4C. Segmentation network: a U-Net trained from scratch to segment the left-ventricular (LV) cavity, myocardium (MYO) and left atrium (LA). Quality score: a score Q built from normalised softmax entropy, with no quality labels used in training. We compared Q with independent expert quality ratings. Statistics: all test-set statistics use 95% confidence intervals (CIs) from a subject-level cluster bootstrap. Results. On 400 frames from 100 held-out patients: View classification: accuracy 0.985 (95% CI 0.9675–0.9975), macro-F1 0.985, ROC-AUC 0.996 (0.989–1.000). The ex-pected calibration error (ECE) was 0.016. Five of the six errors were A2C frames predicted as A4C, and five of the six erroneous frames had been rated Good by the expert. Segmentation (Dice): LV 0.934 (0.926–0.940), MYO 0.873 (0.865–0.879), LA 0.892 (0.880–0.904), macro 0.900 (0.893–0.906). Segmentation (95th-percentile Hausdorff distance, HD95): 4.01, 4.23 and 5.42 mm for LV, MYO and LA. Quality score: Q correlated moderately with the model's own per-frame Dice (Spearman ρ = 0.554) but only weakly with expert ratings (ρ = 0.281). ROC-AUC for detecting Poor frames was 0.683, and 94.8% of frames scored ≥ 90 even though experts rated 54.5% as Good. Probe maneuvers: none of the four public datasets we reviewed (CAMUS, EchoNet-Dynamic, TMED-2, HMC-QU) provided the probe-maneuver annotations required for supervised maneuver prediction; CAMUS contains none (0 of 1,998 frames). The maneuver component was therefore not trained, and no maneuver result is reported. Conclusions. On public data, view classification and multi-structure segmentation can be reproduced with a transparent, leakage-controlled protocol. Single-pass entropy, however, tracks model error far better than expert-judged image quality. For probe guidance, the absence of maneuver annotations in the reviewed public datasets, rather than model design, was the limiting factor. The models are served by a full-stack research application that exposes its supported scope explicitly and never fills unsupported outputs.

Keywords- Echocardiography, deep learning, view classification, cardiac segmentation, predictive uncertainty, image quality assessment, probe guidance, reproducibility, CAMUS.

© 2026 Malarkodi G. his is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

I. INTRODUCTION

Transthoracic echocardiography (TTE) is a widely used, non-invasive cardiac imaging examination [1]. Its diagnostic value depends on acquiring standardised views: the sonog-rapher positions and angles a hand-held transducer so that the imaging plane intersects the heart as professional guidelines specify. Foreshortened or off-axis acquisitions compromise chamber quantification, and real-time deep-learning tools that warn against foreshortening during scanning have been shown to improve measurement workflow and reproducibility [3, 4]. Because acquisition is a learned skill, assisting less experi-enced operators is a practical target for machine learning [5, 6].

Deep learning is now established for several stages of echo-cardiographic analysis:

- View recognition: [7–9]
- Interpretation and phenotyping: [9–12]
- Chamber segmentation: [2, 13–15]
- Functional quantification: [3, 4, 16]
- Image-quality estimation: [7, 8, 17]

Deep-learning acquisition guidance for novice operators has also been evaluated in clinical studies [5, 6].

Jeong et al. recently proposed a multi-task framework that jointly performs view classification and cardiac structure seg-mentation, derives structure-specific image-quality scores from entropy, and passes these features to a lightweight pre-dictor that suggests one of seven corrective probe maneuvers [1]. According to the published abstract, the authors construct-ed a 43-case maneuver-tagged dataset, capturing intentional transitions from standard to non-standard views, to train and validate the system. They reported:

- quality-metric areas under the curve (AUCs) of 0.901–0.987 for distinguishing standard from non-standard views across multiple cardiac structures;
- a top-1 maneuver-prediction accuracy of 85.1% (mAP 0.904);
- an inference time of 17 ms per frame, which the authors describe as supporting feasibility for real-time use.

These values are results of the base study on its own data. None of them was measured, reproduced or compared in the present work.

The central idea of that framework is attractive because it removes the need for dedicated quality annotations. We did not, however, find a public release of the maneuver-tagged dataset, which leaves two questions open.

1**) Does uncertainty-derived quality agree with expert judgement when tested against ratings the model never saw?** Modern neural networks are often over-confident [18, 19], so low entropy need not imply a good acquisition.

2**) Which components can be trained and verified on open data?** Maneuver prediction requires labels that de-scribe the physical motion between frames. Probe-guidance studies derive such labels from acquisitions in which probe motion was recorded [20–22]. The public echocardiography datasets we reviewed do not provide such labels (Section II-D).

This paper addresses both questions with an openly speci-fied pipeline built on the public CAMUS dataset [2]. We do not attempt, and do not claim, a reproduction of the base framework [1]: the dataset, view set, structure set and task definitions differ. Our contributions are:

- **A leakage-controlled CAMUS baseline** for two-view classification and three-structure segmentation. It is report-ed with subject-level bootstrap CIs, calibration, boundary-distance metrics and a documented qualitative sampling rule.
- **An independent evaluation of an entropy-derived quality score** against held-out expert ratings, including its negative findings. The score is fixed before evaluation and not fitted to the ratings.
- **A documented data-verification procedure for CAMUS.** It identifies an array-orientation convention, an intensity-range convention and a duplicated image with conflicting annotations, each of which would silently cor-rupt training if missed.
- **A deployment-oriented research software architec-ture** (not clinically deployed) that

serves the models through REST and WebSocket interfaces with an explicit capability descriptor. Outputs the models were not trained to produce (probe maneuvers, right ventricle, aortic valve, unsupported views) are declared unavailable rather than filled with placeholder values.

- **An analysis of the data barrier to probe guidance**, specifying the data a reproducible maneuver-prediction study would require.

We do not present the user interface itself as a scientific contribution, and we report no user study.
The rest of the paper is organised as follows. Section II re-views related work. Section III describes the data, methods and evaluation protocol. Section IV reports the results. Sec-tions V and VI discuss the findings and limitations, and Sec-tion VII concludes.

II. RELATED WORK

1. Echocardiography view classification

Li et al. trained a single network that classifies six standard echocardiographic views and scores image quality in real time [7]. Jansen et al. combined view classification with the recog-nition of views the network was not trained on [8]. More re-cent systems treat view recognition as one component of broad interpretation models, including multitask video models and vision–language foundation models [9–11].

Two-class discrimination between A2C and A4C, as studied here, is a much narrower task than the multi-view problems addressed in that work. The results below should be read in that light.

2. Cardiac structure segmentation

Encoder–decoder networks derived from U-Net dominate medical image segmentation [23], and properly configured U-Net baselines such as nnU-Net remain highly competitive un-der rigorous validation [24]. For echocardiography, the CAMUS release established a public benchmark for LV cavi-ty, myocardium and LA segmentation and compared several encoder–decoder designs [2]. Subsequent work adapted seg-mentation foundation models to echocardiography video [13], segmented multiple echocardiographic planes with a single model [14], and enforced temporal consistency across the cardiac cycle [15]. Segmentation also underpins automated functional measurement [3, 4, 16].

3. Image quality assessment

Supervised quality estimation regresses or classifies expert quality scores. Li et al. predicted a quality score jointly with the view [7]. Jansen et al. regressed view quality from classi-fier embeddings and reported agreement with expert ratings comparable to that of a second observer [8]. Du et al. assessed several interpretable quality criteria, including cardiac-cycle integrity, depth and gain, with a multitask network [17].

These approaches share a measurable limitation: they need quality annotations, which are costly and variable between raters. The base framework avoids this by deriving quality from the uncertainty of supervised heads [1]. That strategy rests on the broader uncertainty-quantification literature, which presents uncertainty estimates as a means of increasing clinicians' trust in model outputs [25]. The same literature also documents over-confidence and miscalibration [18, 19] and evaluates alternatives to single-pass softmax confidence, such as ensembles and test-time augmentation for segmentation [26, 27], including in echocardiography [28].

4. Probe guidance

Recent clinical studies showed that AI guidance allowed nurses and other novices without ultrasound experience to ac-quire echocardiograms of diagnostic quality [5, 6]. Probe nav-igation has also been learned from expert demonstrations in robotic ultrasound [20]. Recent echocardiography work learns world models of probe motion [21, 22].
In each case, the training signal comes from acquisitions in which probe motion was measured or annotated at acquisition time.
Review of public datasets for maneuver annotations. We reviewed four public echocardiography datasets:

- CAMUS [2, 29]: structure masks, view identity and per-view quality ratings.
- EchoNet-Dynamic [30]: apical four-chamber videos with LV tracings and functional measurements.

- TMED-2 [31]: view labels (PLAX, PSAX, A2C, A4C, other) and aortic-stenosis severity labels.
- HMC-QU [32]: apical two- and four-chamber recordings used for myocardial-infarction detection, with LV-wall segmentation for a subset.

According to their documentation, none of these datasets provided the probe-maneuver annotations required for super-vised maneuver prediction. For CAMUS we verified this di-rectly: 0 of 1,998 manifest records carry a maneuver label (Section III-B). The data of the EchoWorld study, which rec-orded probe motion, cannot be publicly released according to its authors' code repository [22, 33].

This review covers only the datasets listed and does not es-tablish that no maneuver-annotated echocardiography data exist elsewhere. It does mean that maneuver prediction could not be trained or independently benchmarked with the data available to this study.

5. Multi-task and explainable learning

Multi-task learning shares representations across related ob-jectives [34]. The base framework uses this principle to obtain view, structure and quality information from one network [1]. Explainability methods in medical imaging range from attribution maps such as Grad-CAM to uncertainty visualisation, and have been reviewed extensively [35].

The present system offers uncertainty visualisation (a per-pixel entropy map and a decomposition of the derived score) but no attribution method. We do not evaluate whether these visualisations help users.

5. Research gap

The literature above supports five specific gaps.

1. Quality supervision cost. Supervised quality models need expert quality labels [7, 8, 17]. The label-free alterna-tive [1] has not, to our knowledge, been evaluated on pub-lic data against independent expert ratings.
2. Calibration of the uncertainty signal. Entropy-based quality assumes that confidence tracks acquisition quality, but modern networks are frequently miscalibrated [18, 19].
3. Reproducibility of acquisition guidance. Guidance models are trained on acquisitions with recorded or anno-tated probe motion [1, 20, 22]. None of the public datasets we reviewed (Section II-D) allows maneuver prediction to be retrained or compared.
4. Evaluation rigour on small public datasets. Data leakage, including leakage between training and test data, inflates reported performance in machine-learning-based science [36, 37]. Reporting and validation guidance em-phasises clear data partitioning, uncertainty in performance estimates and appropriate metric choice [38–40].
5. Deployment transparency. Reporting guidance asks authors to state a model's intended use and limitations ex-plicitly [38, 39]. When a model covers only part of an in-tended task, as any CAMUS-trained model must (item 3), this requirement extends to the deployed software, which should not imply capabilities the model lacks.

III. MATERIALS AND METHODS

1. System overview

Fig. 1 shows the pipeline.

1. A B-mode frame is converted by a canonical prepro-cessing function into a single-channel 256 × 256 tensor.
2. Two independently trained networks process the tensor: a view classifier and a segmentation network. The two share preprocessing and the serving path but not weights.
3. Softmax outputs yield the predicted view, the label map and per-pixel class probabilities.
4. Normalised entropy is computed for the view posterior and for every pixel, and combined into the derived score Q (Section III-H).
5. The response carries a capability descriptor that lists supported and unsupported outputs.
6. The maneuver component exists only as an interface and returns an explicit "model unavailable" status.

Although the system reports several outputs, it is therefore not a single shared-backbone multi-task network in the sense of [1].

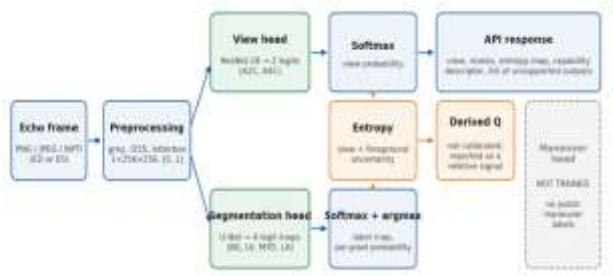


Fig. 1. Overview of the implemented system. Green blocks are trained networks. Orange blocks are the label-free uncertainty computation; its evaluation (Section IV-F) showed weak agreement with expert ratings and no calibration. The dashed block is the probe-maneuver component, which was not trained because the reviewed public datasets, including CAMUS, provide no maneuver labels.

B. Dataset

CAMUS [2] provides 2D apical two-chamber (A2C) and four-chamber (A4C) echocardiograms from 500 patients ac-quired at the University Hospital of Saint-Étienne, France [29]. The dataset is available free of charge from its organisers [29] for non-commercial scientific research only, under a Cre-ative Commons BY-NC-SA 4.0 licence and dataset-specific terms that each user must accept.

The organisers' project page describes an earlier distribution in MHD format. The release used here is the NIfTI version (archive CAMUS_public.zip), in which all 500 patients carry ED and ES annotations (Table I).

• **Annotations:** for each patient and view, the end-diastolic (ED) and end-systolic (ES) frames carry expert masks of the LV cavity (label 1), myocardium (label 2) and left atri-um (label 3).

• **Quality ratings**: each patient–view pair carries an im-age-quality rating (Good, Medium or Poor).

• **Sequences:** the release also contains the half-cardiac-cycle sequences from which the ED and ES frames come. We used them only for the replay function of the applica-tion (Section III-J), never for training or evaluation.

Our inspection of the release found:

- 7,508 files, including 6,000 image volumes, with no un-readable files;
- 74 distinct native frame sizes;
- isotropic pixel spacing of 0.308 mm in every file checked in this release: all 400 test frames and 40 sampled A4C frames (reports/dataset_report.json, re-ports/camus_conventions_check.json, re-ports/phase5/test_per_frame_full.json).
- What CAMUS does not contain:
- views other than A2C and A4C;
- annotations of the right ventricle (RV) or aortic valve (AV);
- any probe-position, probe-motion or maneuver infor-mation (0 of 1,998 manifest records carry a maneuver la-bel).

The quality ratings are coarse: one rating per patient–view pair, shared by its ED and ES frames. The release ships an of-ficial 400/50/50 partition. We did not use it, for the reasons given in Section III-D.

C. Dataset quality control

We checked three data conventions before any training. Each would have corrupted the models without raising an er-ror.

Array orientation. The NIfTI reader returns arrays in (width, height) order, whereas the conventional image order is (height, width). An acoustic sector is narrow at the transducer and wide at depth. Across 40 A4C ED frames, the width of the non-zero region in the bottom 10% of rows divided by that in the top 10% was:

- 0.98 (range 0.96–1.00) for arrays as loaded;
- 7.79 (5.96–10.57) after transposition.
- All frames and masks are therefore transposed on load.

Intensity range. Frames are stored as 32-bit floating point but hold 8-bit values (per-frame maxima 221–255). A normaliser that clips floating-point input to [0, 1] would saturate nearly every pixel. The normalisation routine therefore rejects floating-point input above 1 unless the source range is stated explicitly, and frames are cast to 8-bit only after a lossless-cast check.

Duplicated image. The A4C ED frames of patients 0033 and 0068 have identical SHA-256 digests and identical pixel arrays, yet their expert masks differ:

Dice between the two an-notations is 0.958 (LV), 0.903 (MYO) and 0.936 (LA). Keep-ing both would give contradictory supervision and, if the two patients fell into different partitions, would place the same im-age in training and test data. Both copies were excluded, leav-ing 1,998 of 2,000 annotated frames.

Label mapping. CAMUS labels were kept in their native label space. In particular, the myocardium was not re-mapped onto any other structure, and no RV or AV channel was creat-ed.

D. Subject-level data partitioning

The ED and ES frames of one patient are strongly correlat-ed, and the A2C and A4C views of a patient share anatomy. Frame-level random partitioning would therefore leak near-duplicates across splits [36, 37].

We partitioned the 500 patients 60/20/20 using a seeded shuffle (seed 20260908) of the sorted subject identifiers (Ta-ble I). The training code compares the subject sets of the three partitions at start-up and refuses to run if any subject appears in more than one; a unit test asserts this refusal on a deliber-ately leaked manifest.

We preferred this partition to the official 400/50/50 split be-cause a 100-patient test set gives narrower CIs than a 50-patient one. The consequence is that our numbers are not di-rectly comparable with results reported on the official split.

The official test split also cannot serve as an independent test set for the models reported here. Of its 50 patients, 31 are in our training partition and 7 in our validation partition. A comparison on the official split would require retraining on the official training and validation partitions. That experiment was not performed, and no official-split result is reported.

Model selection used only the validation partition. The test partition was read for the first time by the evaluation script, after both models were frozen. It was read a second time, with the same frozen weights, only to compute the statistics in Sec-tion IV; that re-run reproduced the first evaluation exactly (identical accuracy, confusion matrix and per-structure Dice to six decimal places).

TABLE I
Dataset statistics after quality control

Partition	Patients	Frames	A2C / A4C	ED / ES	Expert quality (Good / Medium / Poor)
Training	300	1,199	600 / 599	599 / 600	584 / 483 / 132
Validation	100	399	200 / 199	199 / 200	208 / 130 / 61
Test	100	400	200 / 200	200 / 200	218 / 144 / 38
Total	**500**	**1,998**	**1,000 / 998**	**998 / 1,000**	**1,010 / 757 / 231**
Frames with maneuver labels	—	**0**	—	—	—

E. Image preprocessing

A single function preprocesses frames for training, evaluation and serving; an automated test asserts that the serving path and the training pipeline produce byte-identical tensors. The steps are:

- Each frame is converted to one grey channel.
- Intensities are divided by the maximum of the data type (255), not by the per-image maximum, so that differences in acquisition gain remain in the input.
- A frame of height H and width W is resized with the aspect-preserving factor

$$s = \mathrm{Min}\left(\frac{256}{H}, \frac{256}{W}\right) \qquad (1)$$

using bilinear interpolation, and centred on a 256 × 256 canvas padded with zeros. Zero corresponds to the absence of acoustic return, so padding adds no artificial signal. 4. Masks undergo the same geometric transform with nearest-neighbour interpolation, which cannot create labels at boundaries.

The model input is a tensor of shape (1, 1, 256, 256) with values in [0, 1] (Fig. 2).

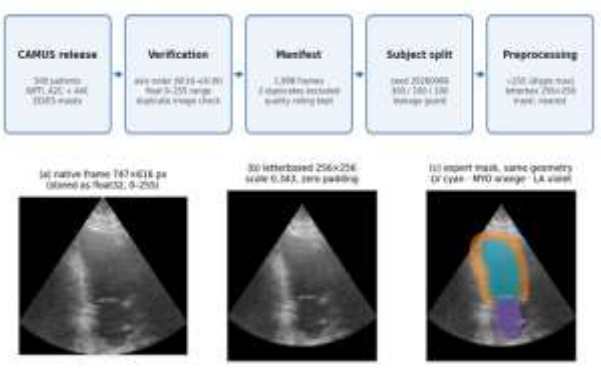


Fig. 2. Data pipeline (top) and preprocessing of a held-out test frame, patient0441 A2C ED (bottom): (a) native frame; (b) letterboxed network input; (c) expert mask transformed with the same geometry.

Augmentation. Echocardiographic views are chiral: a mirrored A4C image shows an anatomy that does not occur in a standard acquisition and would encourage confusion between left- and right-heart structures. Horizontal flips, vertical flips and rotations were therefore not implemented; a unit test asserts that the augmentation module offers no flip or mirror transform. The augmentations used were:

- gamma adjustment ($\gamma \in [0.7, 1.4]$, $p = 0.5$);
- brightness and contrast perturbation (±0.10 and ±0.15, $p = 0.5$);
- multiplicative speckle noise ($\sigma = 0.08$, $p = 0.3$);
- translation of up to 5% of the frame, applied jointly to image and mask ($p = 0.5$).

Random sector occlusion (up to 25% of the frame, attenuation 0.25, $p = 0.2$) was applied only to the classifier. For segmentation it would darken pixels without changing their labels, rewarding the network for segmenting structures that are not visible.

F. View classification network

The classifier (Fig. 3) is a ResNet-18 initialised with ImageNet weights, a common transfer-learning starting point whose benefit varies across medical imaging tasks [41]. We chose it as the smallest member of the standard residual family, because the training set is small (1,199 frames) and training was CPU-only. Architectures designed for constrained hardware, such as MobileNetV2, were not evaluated.

- **Input stem:** the first convolution (7 × 7, stride 2, 64 filters) is replaced by a single-channel convolution whose weights are the sum of the pretrained RGB filters. For a grey input, this preserves the activation magnitude that the pretrained batch-normalisation statistics expect.
- **Feature path:** unchanged. After max pooling, four residual stages of two basic blocks each produce 64 × 64 × 64, 128 × 32 × 32, 256 × 16 × 16 and 512 × 8 × 8 feature maps, followed by global average pooling to a 512-dimensional vector.
- **Head:** the original classification layer is replaced by dropout ($p = 0.2$) and a linear layer with two outputs (A2C, A4C). A softmax is applied at inference.
- **Frozen layers:** the stem convolution, its batch normalisation and the first residual stage (layer1) are frozen, leaving 11,020,034 of 11,171,266 parameters trainable. Freezing stops these parameters from updating but does not freeze the batch-normalisation running statistics in those layers, which continue to adapt in training mode.
- **Loss:** cross-entropy,

$$\mathcal{L}_{\text{cls}} = -\sum_{k} y_k \log p_{v,k}, \qquad \boldsymbol{p}_v = \text{softmax}(f_v(\boldsymbol{x})) \qquad (2)$$

The two classes are balanced (600 vs 599 training frames), so no class weighting was used.



Fig. 3. View classification network. Grey blocks are frozen during fine-tuning; feature-map sizes are channels × height × width.

G. Cardiac segmentation network

The segmentation network (Fig. 4) is a U-Net [23] with batch normalisation, trained from scratch (1,963,860 parameters).

- **Encoder:** an input block and four down-sampling stages, each a 2 × 2 max-pool followed by two 3 × 3 convolution–batch-normalisation–ReLU layers and spatial dropout ($p = 0.1$).

Channel widths are 16, 32, 64, 128 and 256 at resolutions 256, 128, 64, 32 and 16.

- **Decoder:** four stages, each performing ×2 bilinear up-sampling, concatenation with the corresponding encoder features, and the same double-convolution block, returning to 16 channels at 256 × 256.
- **Output:** a 1 × 1 convolution emits four logit maps: background, LV cavity, myocardium and LA.

The per-pixel posterior and label map are

$$p_s(c \mid i,j) = \mathrm{softmax}_c\big(f_s(\boldsymbol{x})_{c,i,j}\big), \qquad \hat{y}(i,j) = \underset{c}{\mathrm{argmax}}\, p_s(c \mid i,j). \qquad (3$$

The network has no RV or AV output.

Loss. Training minimises an equally weighted sum of pixel-wise cross-entropy and soft Dice loss. For each foreground class c, with sums taken over all pixels of the mini-batch and ε = 1,

$$D_c = \frac{2\sum p_c\, y_c + \varepsilon}{\sum p_c + \sum y_c + \varepsilon}, \qquad \mathcal{L}_{\mathrm{seg}} = 0.5\,\mathcal{L}_{\mathrm{CE}} + 0.5\left(1 - 1/3 \sum_{c\in\{\mathrm{LV,MYO,LA}\}} D_c\right). \qquad (4)$$

Background is included in the cross-entropy term but excluded from the Dice term. Because background covers most of each frame, its Dice is close to one regardless of segmentation quality.

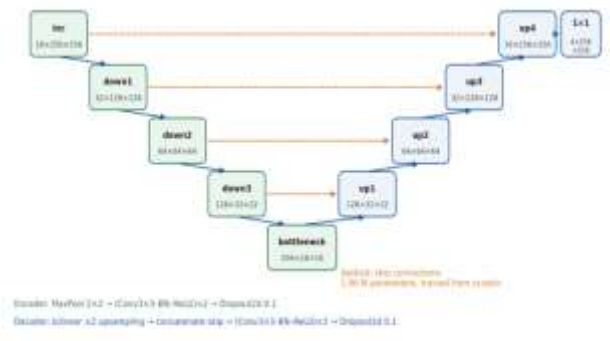


Fig. 4. Segmentation network. Feature sizes are channels × height × width; dashed arrows are skip connections.

Training configuration. Both networks were trained with AdamW, a cosine learning-rate schedule, gradient-norm clip-ping at 1.0 and early stopping on a validation metric (Table II). All random number generators used seed 20260908. Each checkpoint stores the resolved configuration, its hash, the SHA-256 digests of the three partition manifests, the training and validation subject lists, and the software versions.

Checkpoints are written atomically on every improvement of the monitored metric. This rule was introduced after a first segmentation run was terminated by an external process inter-ruption at epoch 24, before its only end-of-training checkpoint was saved. The repeated run reproduced the interrupted run's validation macro Dice at every epoch compared (0.7712, 0.8588, 0.8819, 0.8890, 0.8946 and 0.8963 at epochs 1, 5, 10, 15, 20 and 21), which is empirical evidence that training is de-terministic in this environment.

TABLE II
Training configuration

Setting	View classifier	Segmentation network
Architecture	ResNet-18, ImageNet-initialised	U-Net, depth 4, base width 16
Parameters (trainable / total)	11,020,034 / 11,171,266	1,963,860 / 1,963,860
Input	1 × 256 × 256, [0, 1]	1 × 256 × 256, [0, 1]
Output	2 logits	4 logit maps
Loss	Cross-entropy	0.5 CE + 0.5 soft Dice (background excluded from Dice)
Optimiser, learning rate	AdamW, 3×10^{-4}	AdamW, 1×10^{-3}
Weight decay	1×10^{-4}	1×10^{-4}

Setting	View classifier	Segmentation network
Schedule	Cosine to 1×10^{-6}	Cosine to 1×10^{-6}
Batch size / maximum epochs	16 / 20	8 / 25
Early stopping	Validation accuracy, patience 5, min Δ 1×10^{-4}	Validation macro Dice, patience 6, min Δ 5×10^{-4}
Dropout	0.2 (head)	0.1 (spatial)
Augmentation	Gamma, brightness/contrast, speckle, translation, sector occlusion	Gamma, brightness/contrast, speckle, translation
Epochs run / selected epoch	13 (early stopped) / 8	25 / 21
Training time	36 min 55 s	6 h 14 min wall clock*
Configuration hash	b10c2a700597c1c4	6bfae38b8ee1a9db
Seed	20260908	20260908

H. Uncertainty and derived quality score

For a distribution p over C classes, the normalised predictive entropy is

$$H(\boldsymbol{p}) = -\frac{1}{\ln C}\sum_{c=1}^{C} p_c \ln p_c \in [0,1]. \qquad (5)$$

It is applied to:

- the view posterior, giving H_v=H(p_v);
- every pixel of the segmentation posterior, giving an entropy map.

Background is the most frequent and most confidently predicted class, so a whole-image average makes almost every frame look certain. Segmentation uncertainty is therefore averaged over the predicted foreground Ω={(i,j):y^(i,j)≠"background"}:

$$H_{fg} = \frac{1}{|\Omega|}\sum_{(i,j)\in\Omega} H\left(p_s(\cdot \mid i,j)\right). \qquad (6)$$

The confidence S_c of structure c is the mean posterior p_s (c|i,j) over the pixels predicted as c, and is set to zero if the structure is not detected. With $\bar{S}$ the mean of S_c over the three structures, the derived score is

$$Q = 100\left[1/3\,(1-H_v) + 1/3\,(1-H_{fg}) + 1/3\,\bar{S}\right]. \qquad (7)$$

The equal weights are a fixed default, not fitted to data. Fitting them to expert ratings would turn the method into a supervised quality model, which the approach is meant to avoid. The expert ratings were used only to evaluate Q (Section IV-F).

The deployed service returns Q together with its three components and the flag calibrated: false. Its capability descriptor reports quality assessment as not supported.

I. Probe guidance: scope and data requirements

No probe-maneuver model was trained in this study. The reasons are structural rather than technical.

- A maneuver label describes an action, not an image. It records the physical transition between two probe poses, such as a tilt, rock or rotation that would move the imaging plane from its current position towards a target view. That information is a property of the acquisition process.
- Archived still frames cannot supply it. A still frame, even with a view label and a quality rating, does not encode the pose from which it was acquired or the motion needed to reach a better one. Nor can a trajectory be reconstructed from ED/ES frames of different views, because these are separate acquisitions with no recorded spatial relationship.
- Assigning maneuver labels to CAMUS images by rule would create unsupported ground truth. We did not do so.

A reproducible maneuver-prediction study would need:

1. continuous acquisitions with probe motion or pose measured by a tracking system; recent guidance studies relied on such recorded probe motion [20, 22];
2. an operational definition of each corrective maneuver relative to the pose of an expert-confirmed target view;

3. annotations by experienced sonographers, with inter-rater agreement reported;
4. subject-level partitioning and ethical approval for prospective collection;
5. public release, or a controlled-access protocol, so that results can be independently verified.

The software is prepared for such data but makes no use of it. The inference service defines a seven-class maneuver interface and a model-loading slot. When no maneuver model is present, the service returns an explicit unavailability status, and the user interface displays an empty guidance panel with the reason.

J. Software, hardware and deployment

Environment. The models were implemented in PyTorch 2.8.0 (CPU build) with torchvision 0.23.0 [42], Python 3.11.5 and NumPy 2.4.6. The analysis used SciPy 1.16.2, scikit-learn 1.7.2 and nibabel 5.4.2. All training and inference ran on the CPU of a single workstation without a GPU. The CPU model (AMD Ryzen 5 5625U, 12 logical cores) was recorded on that workstation by the benchmarking script; the training checkpoints record only CPU-only execution.

Serving. A FastAPI service exposes:

• REST endpoints for capability discovery, single-frame analysis and evaluation reports;

• a WebSocket endpoint that streams frame-by-frame results (Fig. 5).

Every analysis response includes the capability descriptor and a list of outputs the models do not produce. The model loader checks each checkpoint's declared scope, so a two-class baseline cannot be loaded into the slot reserved for a nine-view model.

Client. A React/TypeScript client provides upload analysis, result views with switchable mask and entropy layers, a page of measured test metrics, and a "live scan" page. Because no ultrasound hardware was connected, the live-scan page replays recorded half-cycle sequences of test-partition patients through the models, and it is labelled as a replay.

Verification. The backend is covered by 253 automated tests (all passing at the time of writing). The client passes static type checking and a production build.

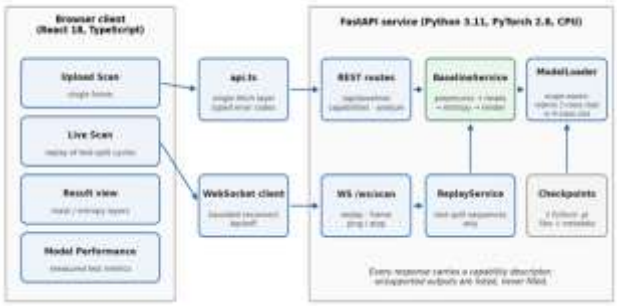


Fig. 5. Deployment architecture.

K. Evaluation metrics and statistical analysis

View classification is evaluated with:

- accuracy, per-class precision, recall and F1, and their macro averages;
- ROC-AUC computed on the A4C probability;
- sensitivity and specificity with A4C as the positive class;
- the confusion matrix;
- calibration, by the expected calibration error (ECE) over ten equal-width confidence bins [18] and the Brier score.
- Segmentation is evaluated per frame with:
- the Dice coefficient and intersection over union [43];
- the 95th-percentile Hausdorff distance (HD95) and the average symmetric surface distance (ASSD), computed from the pooled distances between the one-pixel bounda-ries of prediction and ground truth [40, 43].

Boundary distances were computed on the 256 × 256 model grid and converted to millimetres using the native pixel spacing read from each file header (0.308 mm, isotropic, for all 400 test frames) divided by the letterbox factor s. They therefore carry a discretisation error of about one grid pixel (0.6–0.8 mm). A structure absent from both prediction and ground truth would give an undefined Dice and would be excluded from that structure's mean rather than scored as 1; in the test partition every structure was present in all 400 frames.

Confidence intervals. Because frames from one patient are correlated, all 95% CIs were obtained by a subject-level cluster bootstrap: 100 patients resampled with replacement, 10,000 replicates, percentile intervals, seed 20260908. We report frame-level mean ± standard deviation (SD) and

subject-level mean ± SD, the latter computed over per-patient means.

Post hoc comparisons. Two exploratory comparisons (A2C vs A4C, and ED vs ES) used per-patient mean Dice in each condition, two-sided Wilcoxon signed-rank tests, and Holm correction across the three structures within each comparison. These were not pre-registered.

We performed no significance tests between models, be-cause no alternative model was trained under the same proto-col (Section IV-G).

Quality score. Q was compared with the ordinal expert rat-ings using Spearman's rank correlation, the Kruskal–Wallis test across rating groups, and the ROC-AUC for identifying Poor frames. It was compared with segmentation accuracy using Spearman's correlation with per-frame mean Dice.

Reproducibility commands.

```
pip install -r backend/requirements-research.txt # pinned environment used here
python scripts/inspect_dataset.py # dataset inventory
python ts/verify_camus_conventions.py # orientation, intensity, duplicate checks
python scripts/train_baseline.py --config configs/view_classifier_v1.yaml
python scripts/train_baseline.py --config configs/camus_segmentation_v1.yaml
python scriptsscripts/build_camus_manifest.py # manifest + subject split (seed 20260908)
python scrip/evaluate_baseline.py # one-time test evaluation
python scripts/analyze_test_statistics.py # bootstrap CIs, calibration, HD95/ASSD
python scripts/validate_probabilities.py
python scripts/validate_quality_score.py
python scripts/visualize_predictions.py # percentile-sampled qualitative panels
python scripts/benchmark_inference.py --runs 200 # run on an idle machine
python scripts/make_paper_figures.py
```

IV EXPERIMENTAL RESULTS

All results in this section come from forward passes of the frozen checkpoints over the 400 test frames from 100 patients, except where validation or training values are explicitly labelled.

A. Training behaviour (validation partition)

Fig. 6 shows the learning curves.

- **View classifier**: reached its highest validation accuracy, 0.9925 (validation loss 0.0524), at epoch 8 and was stopped after epoch 13. Validation accuracy fluctuated between 0.92 and 0.99 from epoch 8 onwards, reflecting the small validation set (399 frames). This is why the selected checkpoint's test performance is reported separately below.
- **Segmentation network**: validation macro Dice rose overall, with most fluctuation in the first seven epochs, and peaked at 0.8963 at epoch 21 (LV 0.9312, MYO 0.8626, LA 0.8950). A value of 0.8966 at epoch 22 did not exceed the minimum-improvement threshold and was not selected.

These validation values were used for model selection only and are not the reported performance.

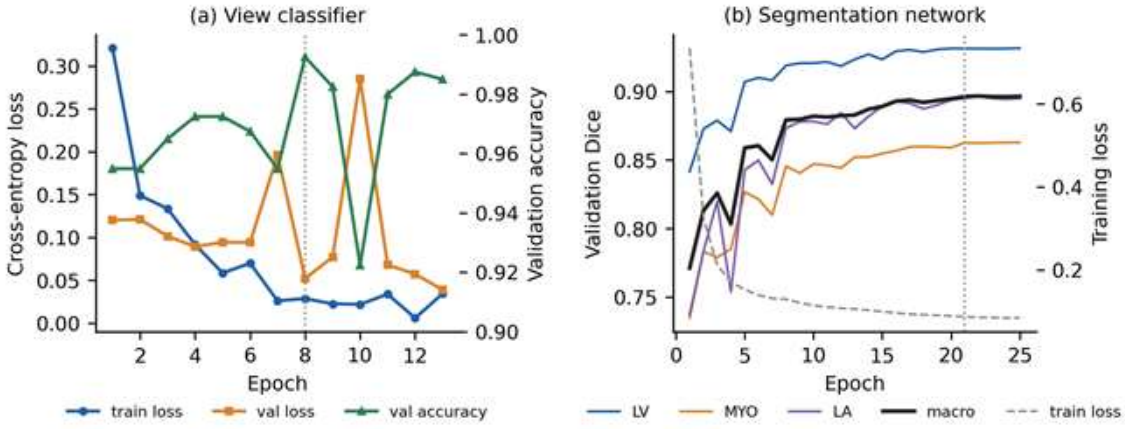


Fig. 6. Learning curves on the training and validation partitions. (a) View classifier. (b) Segmentation network: validation Dice per structure, macro mean and training loss. Dotted lines mark the selected epochs.

Fig. 7. (a) Confusion matrix on the test partition. (b) Reliability by confidence bin: mean confidence and observed accuracy, with the number of frames per bin.

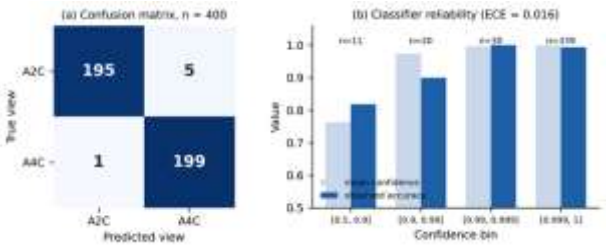
B. View classification (test partition)

The classifier labelled 394 of 400 test frames correctly (Ta-ble III, Fig. 7a):

- Accuracy: 0.985 (95% CI 0.9675–0.9975).
- Macro-F1: 0.985 (0.9675–0.9975).
- ROC-AUC: 0.996 (0.989–1.000).
- Sensitivity / specificity (A4C positive): 0.995 / 0.975.
- Calibration: ECE 0.016, Brier score 0.015 (Fig. 7b).

TABLE III
View classification on the test partition (400 frames, 100 patients)

Class	Precision	Recall	F1	Support
A2C	0.995	0.975	0.985	200
A4C	0.976	0.995	0.985	200
Macro average	0.985	0.985	0.985	400
Accuracy	—	—	0.985 (95% CI 0.9675–0.9975)	400
ROC-AUC	—	—	0.996 (95% CI 0.989–1.000)	400
ECE (10 bins) / Brier score	—	—	0.016 / 0.015	400



The six errors show a structure that aggregate accuracy hides:

- Asymmetric: five of the six were A2C frames predicted as A4C.
- Clustered: they came from four patients. Patients 0020 and 0046 each contributed both of their A2C frames, which suggests patient-specific factors rather than independent random errors.
- Not explained by poor image quality: five of the six er-roneous frames were rated Good by the expert, and all 38 Poor-rated test frames were classified correctly.
- Made with high confidence: confidence on the errone-ous frames ranged from 0.808 to 1.000 (mean 0.934, against 0.993 for correct predictions).

The low overall ECE is therefore dominated by the 339 frames with confidence ≥ 0.999. It does not imply that confi-dence reliably flags individual errors.

C. Segmentation (test partition)

Table IV summarises overall segmentation performance.

TABLE IV
Segmentation performance on the test partition

Structure	Dice, frame mean ± SD	Dice, subject mean ± SD	Dice 95% CI	IoU, frame mean ± SD	IoU 95% CI
LV cavity	0.934 ± 0.047	0.934 ± 0.036	0.926–0.940	0.879 ± 0.070	0.868–0.889
Myocardium	0.873 ± 0.049	0.873 ± 0.036	0.865–0.879	0.777 ± 0.070	0.767–0.787
Left atrium	0.892 ± 0.101	0.892 ± 0.061	0.880–0.904	0.817 ± 0.129	0.801–0.833
Macro	0.900 ± 0.047	0.900 ± 0.033	0.893–0.906	0.824 ± 0.064	0.815–0.833

Structure	HD95 (mm), mean ± SD [median]	HD95 95% CI	ASSD (mm), mean ± SD [median]	ASSD 95% CI
LV cavity	4.01 ± 3.17 [3.38]	3.62–4.49	1.49 ± 0.91 [1.29]	1.36–1.65
Myocardium	4.23 ± 1.89 [3.79]	3.98–4.51	1.64 ± 0.64 [1.51]	1.55–1.74
Left atrium	5.42 ± 8.15 [3.83]	4.59–6.49	1.97 ± 1.99 [1.51]	1.73–2.25

In summary:

- The LV cavity was segmented most accurately and most consistently.
- The myocardium had the lowest overlap (Dice 0.873, IoU 0.777), as expected for a thin structure, where a one-pixel boundary shift removes a large fraction of the over-lap. Its HD95 is nonetheless comparable to that of the LV.
- The left atrium had competitive median performance (Dice 0.921, HD95 3.83 mm) but by far the widest disper-sion (Dice SD 0.101; HD95 SD 8.15 mm), indicating a small number of large failures rather than uniformly lower accuracy.

D. Per-structure performance across views, phases and image quality

Table V breaks performance down by view, cardiac phase and expert quality rating, and reports the exploratory paired tests.

TABLE V
Mean Dice by subgroup (test partition) and exploratory paired comparisons

Subgroup	Frames / patients	LV	MYO	LA	Macro
A2C	200 / 100	0.931	0.878	0.877	0.895
A4C	200 / 100	0.937	0.868	0.907	0.904
End-diastole	200 / 100	0.945	0.869	0.867	0.894
End-systole	200 / 100	0.923	0.877	0.917	0.906
Expert: Good	218 / 69	0.938	0.878	0.891	0.902
Expert: Medium	144 / 54	0.932	0.872	0.890	0.898
Expert: Poor	38 / 16	0.916	0.845	0.908	0.890

Paired comparison (per-patient means, n = 100)	LV: median Δ, p_Holm	MYO: median Δ, p_Holm	LA: median Δ, p_Holm
A2C − A4C	−0.005, 0.070	+0.007, 0.105	−0.017, 0.009
ED − ES	+0.015, 4.9×10^{-11}	−0.010, 0.0023	−0.027, 8.5×10^{-10}

Two-sided Wilcoxon signed-rank tests with Holm correction across structures. Post hoc and exploratory.
The quality subgroups are not independent of patient and are too unbalanced for formal testing (only 16 patients have Poor-rated views), so they are reported descriptively only.

Observed:

- By view: LA segmentation was better in A4C than in A2C (median per-patient difference −0.017, p_Holm = 0.009). LV and myocardium did not differ significantly be-tween views after correction.
- By phase: LV Dice was higher at ED, whereas myocar-dium and LA Dice were higher at ES (all p_Holm ≤ 0.0023).
- By image quality: frames rated Poor had lower mean LV and myocardial Dice than Good frames, but not lower LA Dice.
- Interpretation (hypothesis, not tested here). The phase-dependent differences follow the size of each structure within the cycle: the LV cavity is

largest at ED, while the atrium is largest and the myocardial wall thickest at ES. Larger struc-tures are less sensitive to a fixed boundary error.

-

E. Qualitative results

Fig. 8 shows test frames selected by a fixed rule. For each view, frames were ranked by per-frame mean Dice and sam-pled at the 0th, 10th, 25th, 50th, 75th, 90th and 100th percen-tiles, so the worst frame of each view is included by construc-tion. The figure shows the best, median and worst frame of each view; all 14 panels are provided in the supplementary material (reports/phase5/qualitative/). Each pan-el shows the input, the expert annotation, the prediction over-laid on the input, and the pixel-wise disagreement.

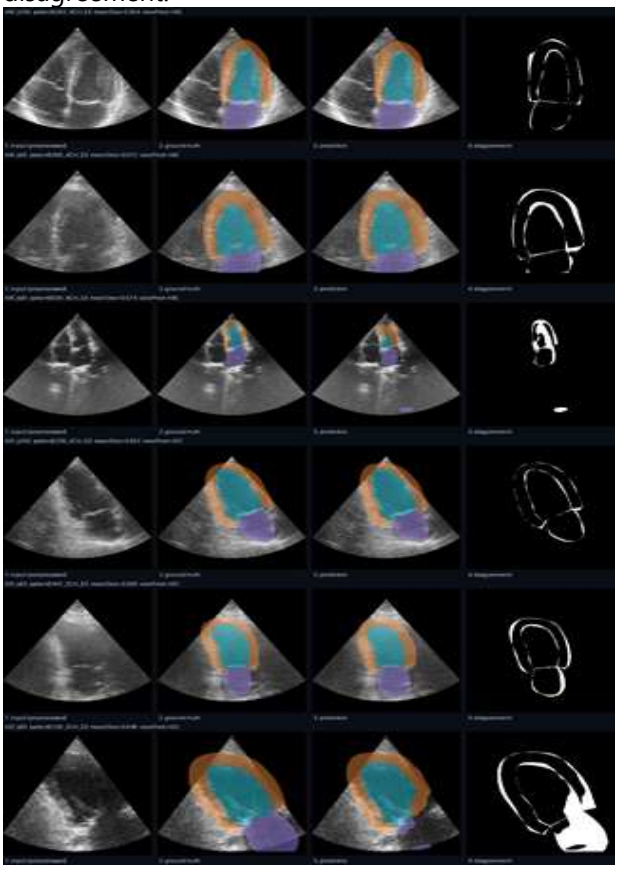


Fig. 8. Test-partition segmentations at fixed Dice percentiles. Rows, top to bottom: A4C 100th percentile (patient0293 ES); A4C 50th percentile (patient0369 ED); A4C 0th percentile (patient0039 ES); A2C 100th percentile (patient0236 ED); A2C 50th percentile (patient0441 ED, mean Dice 0.908); A2C 0th percentile (patient0136 ED, mean Dice 0.648). Columns: input, expert mask, predicted mask overlaid on the input, disagreement map. Panels are shown at the 256 × 256 model resolution and enlarged with nearest-neighbour interpolation for print; no detail is added. LV cavity is shown in cyan, myocardium in orange and left atrium in violet.

What the panels show:

- Where errors concentrate: in the median cases, disa-greement is confined to thin bands along the endocardial and epicardial borders and the mitral annulus. This matches the one-to-two-pixel boundary offsets implied by the ASSD values.
- The worst A2C frame (patient0136 ED): the LV cavity is captured approximately, the myocardium is over-segmented, and the left atrium is almost entirely missed.
- Failure mode: complete or near-complete omission of the atrium on a low-contrast frame. This single mode ex-plains most of the LA dispersion in Table IV.
- Remaining sources of error: the apical myocardium, where the wall is thin and often poorly resolved, and the atrioventricular boundary, where the mitral valve plane must be inferred.
- F. Uncertainty and derived quality score
- Probability checks (64 test frames). Before evaluating Q, we verified that the outputs were well-formed probability dis-tributions:
- softmax outputs summed to one within 2.4×10^{-7};
- all entropies lay in [0, 1];
- mean view entropy was lower for correct predictions (0.030) than for incorrect ones (0.274), although this com-parison rests on few errors.

Agreement with expert ratings. Q was then compared with the expert quality ratings of all 400 test frames (Table VI, Fig. 9). The ratings were never used in training or model selection.

- Ordering: the mean score was correctly ordered across groups (Good 95.15 > Medium 94.78 > Poor 94.32), and the groups differed significantly (Kruskal–Wallis $H = 31.84$, $p = 1.2 \times 10^{-7}$).

- Effect size: the rank correlation with the ratings was weak (ρ = 0.281), the group means differed by less than one point, and the ROC-AUC for identifying Poor frames was 0.683.
- Tracking of model accuracy: Q correlated considerably more strongly with the model's own per-frame Dice (ρ = 0.554, p = 1.4×10^{-33}).
- Calibration: 379 of 400 frames (94.8%) scored ≥ 90 and would receive the highest grade under the interface's original thresholds, whereas experts rated 218 of 400 (54.5%) as Good.

Evaluation of the derived quality score Q (test partition, 400 frames)

Measure	Value
Spearman ρ, *Q* vs expert rating	0.281 ($p = 1.1 \times 10^{-8}$)
Kruskal–Wallis H across ratings	31.84 ($p = 1.2 \times 10^{-7}$)
Mean *Q*: Good / Medium / Poor	95.15 / 94.78 / 94.32
ROC-AUC, Poor vs other	0.683
Spearman ρ, *Q* vs per-frame mean Dice	0.554 ($p = 1.4 \times 10^{-33}$)
Q range / median / SD	63.4–97.1 / 96.1 / 4.5
Frames with $Q \geq 90$	379 (94.8%)
Expert-rated Good	218 (54.5%)
Calibrated to the clinical scale	No

TABLE VI

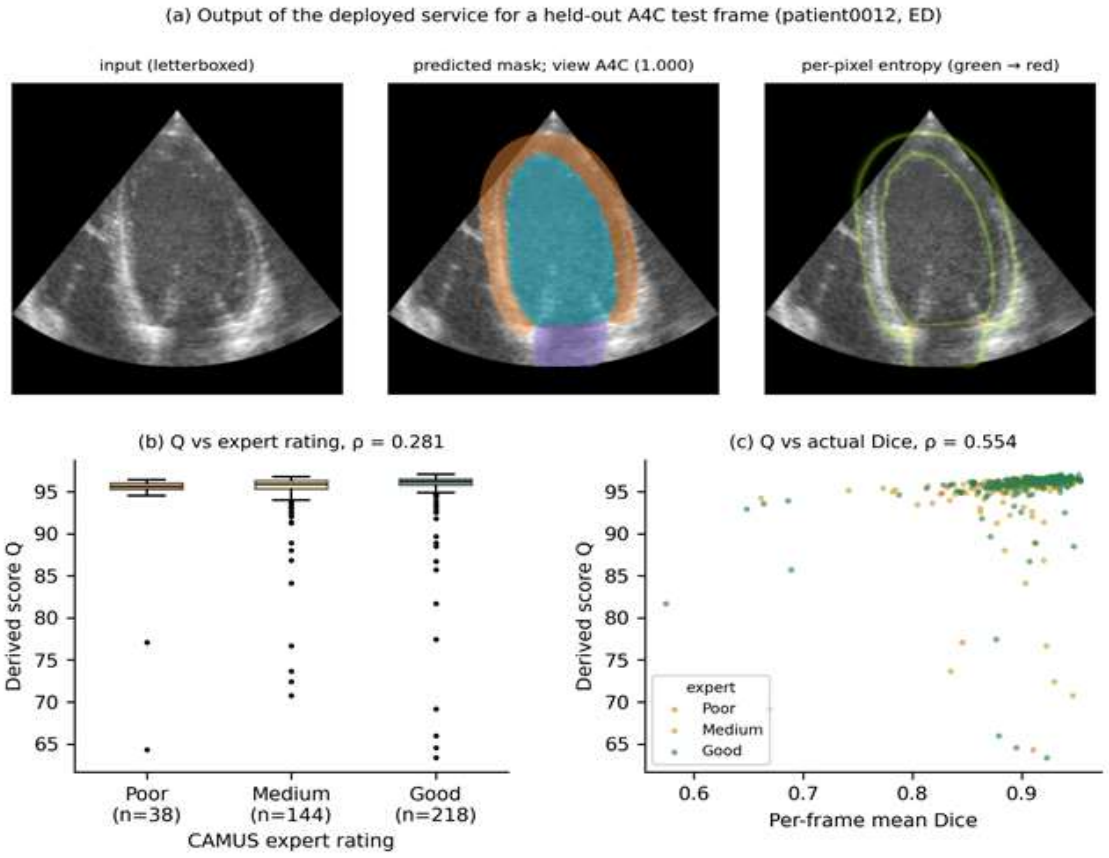


Fig. 9. (a) Service output for a held-out A4C test frame (patient0012 ED): input, predicted mask, and per-pixel entropy map, which concentrates on structure boundaries and the valve plane. (b) Distribution of Q by expert rating. (c) Q against the per-frame mean Dice, coloured by expert rating.

On this evidence, the entropy-derived score is a relative in-dicator of how well the model performed on a frame, not a validated measure of acquisition quality. It is reported and displayed accordingly (Section III-H, Fig. 10).

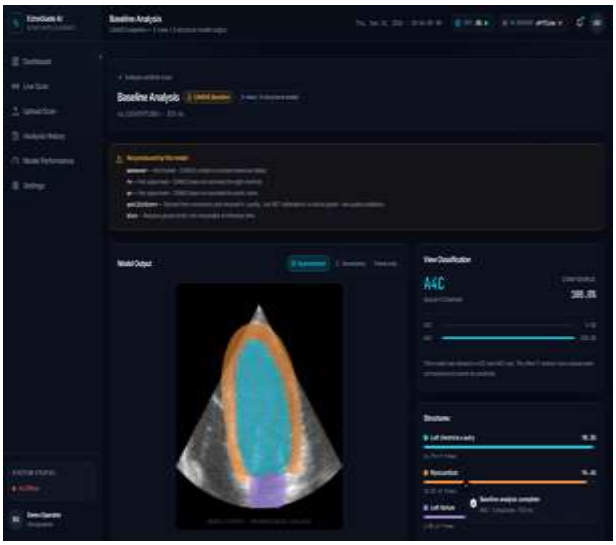


Fig. 10. Analysis view of the application for a held-out A4C test frame (patient0012 ED). It shows the predicted view and confidence, the segmentation overlay, per-structure confidence, and the explicit list of outputs the model does not produce. The values in this figure are one illustrative output, not aggregate results.

G. Computational performance

Model size (load-independent):

- view classifier: 11.17 M parameters, 44.8 MB checkpoint including metadata;
- segmentation network: 1.96 M parameters, 7.9 MB.

Latency, throughput and frame rate: not measured. A benchmarking script for the trained checkpoints exists (scripts/benchmark_inference.py: batch 1, float32, 20 warm-up and 200 timed calls). Its only run is inva-lid because the workstation was not idle: total CPU utilisation was 98.7% at the start, owing to an unrelated workload. Its timings are therefore not reported. No latency, throughput or frame-rate value is reported for EchoGuide AI, and no re-al-time claim is made. A benchmark on an idle machine is required before any such value can be stated.

TABLE VII
Computational characteristics

Quantity	View classifier	Segmentation network	Full service call
Parameters	11,171,266	1,963,860	—
Checkpoint size (with metadata)	44.8 MB	7.9 MB	—
Latency	Not measured	Not measured	Not measured
Throughput / frame rate	Not measured	Not measured	Not measured

H. Ablation studies

No ablation experiments were performed. Each segmen-tation training run takes several hours on the available CPU, and no alternative configuration was trained under the same protocol. Table VIII lists the ablations we consider most in-formative, as future work. No results are implied.

TABLE VIII
Recommended ablations (not performed)

Factor	Variants	Question	Status
Resizing	Letterbox vs stretch vs crop	Does aspect preservation matter for boundary accuracy?	Not performed
Intensity normalisation	dtype maximum vs per-image min–max vs z-score	Does preserving gain help the quality signal?	Not performed
Classifier initialisation	ImageNet vs random; frozen stages	How much does pretraining contribute on 1,199 frames?	Not performed
Classifier backbone	ResNet-18 vs lightweight backbones, e.g. MobileNetV2	Accuracy–latency trade-off for deployment	Not performed
Segmentation width / resolution	Base width 16 vs 32; 256 vs 384 px	Accuracy–cost trade-off	Not performed
Loss	CE vs Dice vs CE + Dice; background in Dice	Effect on thin-structure accuracy	Not performed
Augmentation	With vs without each transform; flips as a negative control	Effect of anatomically constrained augmentation	Not performed
Architecture	Separate networks vs shared encoder with two heads	Does multi-task sharing help at this data scale?	Not performed
Uncertainty estimator	Softmax entropy vs MC dropout vs test-time augmentation vs ensembles [26]	Does Q agree better with experts?	Not performed
Seeds	3–5 training seeds	Run-to-run variance of all reported metrics	Not performed

I. Relation to the base framework

Table IX compares the two studies along the dimensions that determine comparability. Because the datasets, view sets, structure sets, evaluation protocols and hardware all differ, no numerical performance comparison is made and no superiority is implied. Values in the base-framework column are those reported by Jeong et al. [1]; none was reproduced here.

TABLE IX
Base framework [1] and the present study

Dimension	Base framework (as described in its published abstract)	This study
Dataset	43-case maneuver-tagged dataset (intentional standard-to-non-standard transitions) used to train and validate the system; no public release found	CAMUS, public; 500 patients; 100-patient held-out test set
Views	Standard and intentionally non-standard views (the abstract does not state the number of view classes)	A2C and A4C only
Segmentation structures	Cardiac structures used for structure-specific quality	LV cavity, myocardium, LA (no RV, no AV)
Maneuver labels	Seven corrective maneuvers	None available (0 of 1,998 records)
Quality assessment	Entropy-based, structure-specific, label-free; AUC 0.901–0.987 for standard vs non-standard views	Entropy-based, label-free *Q*; evaluated against expert Good/Medium/Poor ratings; weak agreement ($\rho = 0.281$), not calibrated
Model architecture	Multi-task framework combining view classification and segmentation	Two separate networks (ResNet-18; U-Net) sharing preprocessing and serving
Training data	The 43-case maneuver-tagged dataset described above	1,199 frames from 300 patients
Evaluation protocol	Per the original publication	Subject-level split, single test read, subject-level bootstrap CIs
Maneuver prediction	Top-1 accuracy 85.1%, mAP 0.904 (reported by Jeong et al.)	Not trained: no maneuver labels in CAMUS or the other reviewed datasets
Reported inference time	17 ms per frame (reported by Jeong et al.)	Not measured (Section IV-G)
Intended use	Acquisition guidance for novice or trainee users, described by the authors as feasible for real-time use	Research web application with REST/WebSocket service; replay of recorded cycles; no probe hardware; no clinical deployment

5. DISCUSSION

This study set out to establish which parts of an uncertainty-aware echocardiography assistant can be built and verified on public data. Three findings stand out.

Two-view classification and three-structure segmentation are reproducible on CAMUS with a transparent protocol.

- Observed: a macro Dice of 0.900 (95% CI 0.893–0.906) and view accuracy of 0.985 (0.9675–0.9975), from 100 patients who took no part in training or model selection.
- Why the protocol matters: the subject-level partition, leakage check, single test read and fixed qualitative sam-pling rule address common sources of optimistic bias in small-dataset studies [36, 37, 40].
- What this does not show: we deliberately make no nu-merical comparison with published CAMUS results. Those use a different partition and, in several cases, different tar-get definitions (for example, endocardial and epicardial contours instead of a myocardial region) [2]. Our figures are best read as a documented baseline, not as a claim of improvement.
- Scope of the classification result: it concerns a two-class problem and should not be extrapolated to multi-view classification [7, 8].
- Confidence is informative about model error but only weakly informative about acquisition quality.
- Observed: the derived score Q was ordered correctly across expert quality groups, but its agreement with experts was weak ($\rho = 0.281$), whereas its agreement with actual segmentation accuracy was moderate ($\rho = 0.554$).
- The classifier shows the same pattern from a different angle: it was confidently wrong on Good-quality frames and correct on every Poor-quality frame.
-

We propose three non-exclusive explanations, none of which this study can separate (hypotheses):

1. Different constructs. Low entropy indicates that a frame resembles the training distribution. The CAMUS test frames come from the same centre, scanner type and rou-tine acquisition protocol as the training frames [29], so they may be in-distribution for the model even when an expert rates their quality as poor. An expert judges foreshortening, endocardial dropout and gain, which overlap only partially with what makes a network uncertain.
2. Label resolution. CAMUS provides one rating per pa-tient–view pair, so the score is evaluated against a label that cannot vary between the ED and ES frames it covers.
3. A weak uncertainty estimator. Single-pass softmax en-tropy is known to be over-confident [18, 19]. Sampling-based, ensemble or dedicated estimators [26–28] may track image degradation more closely.

These observations do not contradict the base framework's reported AUCs [1], which concern a different task (separating standard from non-standard views) on different data. They do indicate that expert-rated quality within standard apical views is a harder target. Any clinical use of an entropy-derived quality score would require calibration and validation against the specific quality construct intended.

Interpretation of the quality score. The score is explorato-ry and uncalibrated. It tracks model correctness ($\rho = 0.554$) more strongly than expert-rated image quality ($\rho = 0.281$). In its current single-pass form, the label-free score is therefore not an adequate substitute for expert quality assessment. Calibration on held-out validation data and stronger uncertainty estimation (e.g. Monte Carlo dropout, test-time augmentation or ensembles) remain future work.

For probe guidance, the limiting factor is data. The sev-en-maneuver task is defined by the physical motion between frames. CAMUS contains none of that information, and nei-ther did the other public datasets we reviewed (Section II-D). Recent guidance studies obtain it from acquisitions with rec-orded probe motion [20–22] and from institutional data collections [1, 5]. We therefore consider an openly available, tracked, expert-annotated acquisition dataset to be a prerequisite for reproducible research on guidance. Section III-I specifies what such a dataset would need to contain.

On deployment and transparency. The serving architecture makes the supported scope of the models explicit and de-clares unsupported outputs instead of filling them. We regard this as good engineering practice consistent with reporting guidance [38, 39], not as a scientific result. The uncertainty maps show that model uncertainty concentrates on structure boundaries and the valve plane (Fig. 9a). Whether such dis-plays help operators or clinicians was not studied, and the sys-tem should not be described as explainable in the sense of at-tribution-based methods [35].

6. LIMITATIONS

1. Scope. Two of the standard TTE views and three cardi-ac structures were modelled. The right ventricle and aortic valve, which the intended application requires, have no su-pervision in CAMUS.
2. No probe guidance. No maneuver model was trained and no guidance performance is reported. No synthetic maneuver labels were generated.
3. Single public dataset. All data come from one public collection. External validation on data from other centres, vendors and patient populations was not performed, so generalisability is unknown.
4. Still frames only. Evaluation used ED and ES frames. Performance on arbitrary frames of the cycle, on live ac-quisitions, and on deliberately off-axis or non-standard ac-quisitions was not measured.
5. Quality score. The score is uncalibrated and agrees on-ly weakly with expert ratings, whose own inter-rater relia-bility is not reported for this dataset.
6. Single training run per model. Run-to-run variance across seeds was not measured, and no ablations were per-formed (Table VIII). The CIs capture test-sample variabil-ity only.

7. Latency. No valid latency measurement exists (Section IV-G). No latency, throughput, frame-rate or real-time per-formance is claimed.
8. Boundary metrics were computed on the resampled 256 × 256 grid, not at native resolution.
9. No clinical evaluation. No clinical validation, reader study, user study or outcome analysis was performed. The system is a research prototype and is not intended for diag-nostic use.
10. Post hoc analyses. The subgroup comparisons in Ta-ble V were exploratory and not pre-registered.
11. Official CAMUS split. No result is reported on the of-ficial test split. That split overlaps our training and valida-tion partitions (Section III-D), so a comparison would re-quire retraining.
12. Dataset review. The review of public datasets for ma-neuver annotations covered four datasets (Section II-D) and is not exhaustive.

7. CONCLUSION

We presented an openly specified, leakage-controlled deep-learning pipeline for echocardiographic view classification and cardiac structure segmentation on the public CAMUS dataset. On 100 held-out patients it reached a view accuracy of 0.985 and a macro Dice of 0.900, reported with subject-level CIs, calibration, boundary metrics and failure cases.

An entropy-derived, label-free quality score tracked the model's own accuracy moderately well but agreed only weak-ly with expert quality ratings. It is exploratory and uncalibrat-ed, it is not a substitute for expert quality assessment, and it should not be presented as an image-quality grade without further work.

Probe-maneuver prediction could not be trained, because none of the public datasets we reviewed provides probe-maneuver annotations. Open, tracked and expertly annotated acquisition data are, in our assessment, the main prerequisite for reproducible research on AI-guided echocardiographic ac-quisition.

The most valuable next steps are:

- calibration and better uncertainty estimation for the quality signal;
- a valid latency measurement on an idle machine;
- multi-seed training and the ablations in Table VIII;
- external validation;
- collection of a tracked acquisition dataset for guidance.
- Data Availability
- Dataset: CAMUS is available free of charge from its or-ganisers at https://www.creatis.insa-lyon.fr/Challenge/camus/databases.html (accessed 17 September 2026) [29]. Access is governed by the CC BY-NC-SA 4.0 licence and dataset-specific terms:

– use is restricted to non-commercial scientific re-search;
– each user must accept the terms;
– the dataset paper [2] must be cited.

• Derived artefacts: the following generated in this study are kept in the authors' private project repository under data/manifests/ and reports/, and are available from the corresponding author on reasonable request:

– the subject-level partition manifests (frame identifiers and relative paths, with SHA-256 digests);
– per-frame test results;
– statistical outputs.

• Redistribution: CAMUS image files are not redistributed with the code. Users obtain them from the organisers; scripts/export_samples.py regenerates the ex-ample frames used by the software. Figures in this article that show CAMUS frames reproduce them with attribution under the dataset licence.

Code Availability

The source code is maintained in a private Git repository; the version described in this article is tagged v1.0-paper. It is not deposited in a public repository and is not publicly available. It is available from the corresponding author on reasonable request. The repository includes:

- data inspection, manifest construction and verification;
- training, evaluation and statistical analysis;
- figure generation;
- the FastAPI service and the React client.

The trained checkpoints are not stored in the repository, in line with the project's version-control policy. Their SHA-256 digests and configuration hashes (b10c2a700597c1c4 and 6bfae38b8ee1a9db) are recorded in re-ports/phase5/training_summary.json. The checkpoints are likewise available from the corresponding au-thor on reasonable request, subject to the CAMUS licence terms.

Ethics / Data Use Statement

This study used only a publicly released, de-identified da-taset, under its licence terms for non-commercial scientific re-search. According to the dataset's organisers, the examinations were acquired at the University Hospital of St Etienne (France) and included in the CAMUS study within the regulation set by the hospital's local ethical committee, after full anonymization [29]. No new data were collected from human participants by the authors, and no clinical intervention was performed. Ethical approval and informed consent were therefore not required for this study.

Declarations

- Funding: This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.
- Competing interests: The authors declare that they have no known competing financial interests or personal rela-tionships that could have appeared to influence the work reported in this paper.
- Author contributions: Malarkodi G: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. R.K. Soundarayaa: Supervision, Validation, Writing – review & editing. S.M. Madhumathi: Data curation, Visualization, Writing – review & editing.
- Use of generative AI: A generative AI assistant, Claude (Anthropic), accessed through the Claude Code tool, was used in September 2026; the models used included Claude Opus 5 and Claude Sonnet 5. It was used to (i) write and revise the software (the web application and the data-processing, training, evaluation and statistical-analysis scripts); (ii) run those scripts and record their outputs; (iii) generate the figures; (iv) draft and revise the manuscript text; and (v) retrieve bibliographic records and check them against Crossref, PubMed, PMLR and arXiv, and cross-check the manuscript's numbers against the stored result files. The AI system is not an author. All reported results are outputs of the described code run on the CAMUS data, and every number is traceable to a stored result file listed in the evidence inventory. All authors reviewed and approved this disclosure.
- Reporting guidelines: the manuscript was prepared with reference to CLAIM [38] and TRIPOD+AI [39].

Acknowledgment

The authors thank the creators of the CAMUS dataset for making it publicly available [2]. The authors also thank the management of P.S.R. Engineering College, Sivakasi, for its encouragement and support.

REFERENCES

1. [2] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. Grenier, C. Lartizien, J. D'hooge, L. Lovstakken, and O. Bernard, "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography," IEEE Trans. Med. Imaging, vol. 38, no. 9, pp. 2198–2210, 2019, doi: 10.1109/TMI.2019.2900516.
2. [3] H. N. Pettersen, S. Sæbø, D. Pasdeloup, et al., "Real-time deep learning-based image guiding and automated left ventricular measurements to reduce test-retest variability," Open Heart, 2025, doi: 10.1136/openhrt-2025-003783.
3. [4] V. Holmstrøm, E. Smistad, S. B. Stølen, et al., "Real-time global longitudinal strain during echocardiography: a deep learning platform for improved workflow," J. Am. Soc. Echocardiogr., 2025.

4. [5] B. Trost, L. Rodrigues, C. Ong, et al., "Artificial intelligence empowers novice users to acquire diagnostic-quality echocardiography," JACC: Adv., 2025, doi: 10.1016/j.jacadv.2025.102005.
5. [6] C. Schissel, D. Trotzky, N. Avisar, et al., "Artificial intelligence in breaking the learning curve for echocardiography: a secondary analysis of a multicentre trial," Eur. Heart J. Digit. Health, vol. 7, no. 4, art. no. ztag065, 2026, doi: 10.1093/ehjdh/ztag065.
6. [7] X. Li, H. Zhang, J. Yue, et al., "A multi-task deep learning approach for real-time view classification and quality assessment of echocardiographic images," Sci. Rep., vol. 14, 2024, doi: 10.1038/s41598-024-71530-z.
7. [8] G. E. Jansen, B. D. de Vos, M. A. Molenaar, et al., "Automated echocardiography view classification and quality assessment with recognition of unknown views," J. Med. Imaging, vol. 11, no. 5, art. no. 054002, 2024, doi: 10.1117/1.JMI.11.5.054002.
8. [9] M. Vukadinovic et al., "Comprehensive echocardiogram evaluation with view primed vision language AI," Nature, 2025, doi: 10.1038/s41586-025-09850-x.
9. [10] M. Christensen, M. Vukadinovic, N. Yuan, et al., "Vision–language foundation model for echocardiogram interpretation," Nat. Med., vol. 30, pp. 1481–1488, 2024, doi: 10.1038/s41591-024-02959-y.
10. [11] G. Holste, E. K. Oikonomou, M. Tokodi, et al., "Complete AI-enabled echocardiography interpretation with multitask deep learning," JAMA, vol. 334, no. 4, 2025, doi: 10.1001/jama.2025.8731.
11. [12] E. K. Oikonomou, A. Vaid, G. Holste, et al., "Artificial intelligence-guided detection of under-recognised cardiomyopathies on point-of-care cardiac ultrasonography: a multicentre study," Lancet Digit. Health, vol. 7, pp. e113–e123, 2025, doi: 10.1016/S2589-7500(24)00249-8.
12. [13] X. Deng, H. Wu, R. Zeng, and J. Qin, "MemSAM: Taming segment anything model for echocardiography video segmentation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 9622–9631.
13. [14] J. Hu, W. Xue, J. Cheng, Y. Liu, W. Zhuo, and D. Ni, "EchoONE: Segmenting multiple echocardiography planes in one model," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025, pp. 5207–5216.
14. [15] W. Lu, Y. Wang, W. Dai, Y. Wu, H. Xu, and D. Kong, "Echo-ODE: A dynamics modeling network with neural ODE for temporally consistent segmentation of video echocardiograms," Front. Physiol., vol. 16, art. no. 1629121, 2025, doi: 10.3389/fphys.2025.1629121.
15. [16] S. Olaisen et al., "Automatic measurements of left ventricular volumes and ejection fraction by artificial intelligence: clinical validation in real time and large databases," Eur. Heart J. Cardiovasc. Imaging, vol. 25, no. 3, pp. 383–395, 2024, doi: 10.1093/ehjci/jead280.
16. [17] L. Du, W. Xue, Z. Qi, Z. Shi, G. Guo, X. Yang, D. Ni, and J. Yao, "Deep learning framework for interpretable quality control of echocardiography video," Med. Phys., 2025, doi: 10.1002/mp.17722.
17. [18] C. Penso, L. Frenkel, and J. Goldberger, "Confidence calibration of a medical imaging classification system that is robust to label noise," IEEE Trans. Med. Imaging, vol. 43, no. 6, pp. 2050–2060, 2024, doi: 10.1109/TMI.2024.3353762.
18. [19] B. Murugesan, S. A. Vasudeva, B. Liu, H. Lombaert, I. Ben Ayed, and J. Dolz, "Neighbor-aware calibration of segmentation networks with penalty-based constraints," Med. Image Anal., vol. 101, art. no. 103501, 2025, doi: 10.1016/j.media.2025.103501.
19. [20] Z. Jiang, Y. Bi, M. Zhou, Y. Hu, M. Burke, and N. Navab, "Intelligent robotic sonographer: Mutual information-based disentangled reward learning from few demonstrations," Int. J. Robot. Res., vol. 43, no. 7, pp. 981–1002, 2024, doi: 10.1177/02783649231223547.
20. [21] H. Jiang, Z. Sun, N. Jia, et al., "Cardiac Copilot: Automatic probe guidance for echocardiography with world model," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, LNCS, Springer, 2024, pp. 190–199, doi: 10.1007/978-3-031-72378-0_18.

21. [22] Y. Yue, Y. Wang, H. Jiang, P. Liu, S. Song, and G. Huang, "EchoWorld: Learning motion-aware world models for echocardiography probe guidance," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025, pp. 25993–26003, doi: 10.1109/CVPR52734.2025.02421.
22. [23] R. Azad et al., "Medical image segmentation review: The success of U-Net," IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, pp. 10076–10095, 2024, doi: 10.1109/TPAMI.2024.3435571.
23. [24] F. Isensee, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. Maier-Hein, and P. F. Jaeger, "nnU-Net revisited: A call for rigorous validation in 3D medical image segmentation," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, LNCS, vol. 15009, Springer, 2024, pp. 488–498, doi: 10.1007/978-3-031-72114-4_47.
24. [25] B. Lambert, F. Forbes, S. Doyle, H. Dehaene, and M. Dojat, "Trustworthy clinical AI solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis," Artif. Intell. Med., vol. 150, art. no. 102830, 2024.
25. [26] K.-C. Kahl, C. T. Lüth, M. Zenk, K. Maier-Hein, and P. F. Jaeger, "ValUES: A framework for systematic validation of uncertainty estimation in semantic segmentation," in Proc. Int. Conf. Learn. Represent. (ICLR), 2024.
26. [27] M. Zenk, D. Zimmerer, F. Isensee, J. Traub, T. Norajitra, P. F. Jäger, and K. Maier-Hein, "Comparative benchmarking of failure detection methods in medical image segmentation: Unveiling the role of confidence aggregation," Med. Image Anal., vol. 101, art. no. 103392, 2025, doi: 10.1016/j.media.2024.103392.
27. [28] A. Judge, T. Judge, N. Duchateau, R. A. Sandler, J. Z. Sokol, O. Bernard, and P.-M. Jodoin, "Domain adaptation of echocardiography segmentation via reinforcement learning," in Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, LNCS, vol. 15009, Springer, 2024.
28. [29] CREATIS, "CAMUS project: Cardiac Acquisitions for Multi-structure Ultrasound Segmentation — Datasets." [Online]. Available: https://www.creatis.insa-lyon.fr/Challenge/camus/databases.html (accessed Sep. 17, 2026).
29. [30] Stanford AIMI, "EchoNet-Dynamic: Cardiac ultrasound dataset." [Online]. Available: https://echonet.github.io/dynamic/ (accessed Oct. 24, 2026).
30. [31] Tufts Medical Echocardiogram Dataset (TMED), "TMED-2 dataset." [Online]. Available: https://tmed.cs.tufts.edu/tmed_v2.html (accessed Sep. 17, 2026).
31. [32] A. Degerli et al., "HMC-QU dataset." [Online]. Available: https://www.kaggle.com/aysendegerli/hmcqu-dataset (accessed Oct. 24, 2026).
32. [33] LeapLabTHU, "EchoWorld" (source code repository). [Online]. Available: https://github.com/LeapLabTHU/EchoWorld (accessed Sep. 17, 2026).
33. [34] J. Yu et al., "Multitask learning 1997–2024: Part I fundamentals," Harvard Data Sci. Rev., vol. 7, no. 3, 2025, doi: 10.1162/99608f92.462f48a2.
34. [35] D. Muhammad and M. Bendechache, "Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis," Comput. Struct. Biotechnol. J., vol. 24, pp. 542–560, 2024, doi: 10.1016/j.csbj.2024.08.005.
35. [36] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, "Data leakage inflates prediction performance in connectome-based machine learning models," Nat. Commun., vol. 15, art. no. 1829, 2024, doi: 10.1038/s41467-024-46150-w.
36. [37] V. M. Young, S. Gates, L. Y. Garcia, and A. Salardini, "Data leakage in deep learning for Alzheimer's disease diagnosis: A scoping review of methodological rigor and performance inflation," Diagnostics, vol. 15, no. 18, art. no. 2348, 2025, doi: 10.3390/diagnostics15182348.
37. [38] A. S. Tejani et al., "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update," Radiol. Artif. Intell., vol. 6, no. 4, art. no. e240300, 2024, doi: 10.1148/ryai.240300.
38. [39] G. S. Collins, K. G. M. Moons, P. Dhiman, et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ,

vol. 385, art. no. e078378, 2024, doi: 10.1136/bmj-2023-078378.

39. [40] L. Maier-Hein, A. Reinke, P. Godau, et al., "Metrics reloaded: recommendations for image analysis validation," Nat. Methods, vol. 21, no. 2, pp. 195–212, 2024, doi: 10.1038/s41592-023-02151-z.
40. [41] C. Raptis, E. Karavasilis, G. Anastasopoulos, and A. Adamopoulos, "Comparative analysis of conventional CNN v's ImageNet pretrained ResNet in medical image classification," Information, vol. 15, no. 12, art. no. 806, 2024, doi: 10.3390/info15120806.
41. [42] J. Ansel et al., "PyTorch 2: Faster machine learning through dynamic Python bytecode transformation and graph compilation," in Proc. 29th ACM Int. Conf. Archit. Support Program. Lang. Oper. Syst. (ASPLOS), vol. 2, 2024, pp. 929–947, doi: 10.1145/3620665.3640366.
42. [43] A. Reinke et al., "Understanding metric-related pitfalls in image analysis validation," Nat. Methods, vol. 21, no. 2, pp. 182–194, 2024, doi: 10.1038/s41592-023-02150-0.