1828
AKADÉMIAI KIADÓ

Deep learning models for automated view and orientation classification and cardiac cycle phase detection in echocardiography

IMAGING

ORIGINAL ARTICLE

ÁDÁM SZIJÁRTÓ[1], THOMAS AKOS SZEIER[2], TETSUJI KITANO[3], YOSUKE NABESHIMA[4], BÁLINT KÁROLY LAKATOS[1], ALEXANDRA FÁBIÁN[1], ZSUZSANNA LADÁNYI[1], LUCA SZÁVAI[1], MÁTÉ TOLVAJ[1], ANDREA FERENCZ[1], TÍMEA KATALIN TURSCHL[1], BÁLINT MAGYAR[5], ZSOLT BAGYURA[1], FEDERICO M. ASCH[6], KARIMA ADDETIA[7], MASAAKI TAKEUCHI[8], BÉLA MERKELY[1], ATTILA KOVÁCS[1,9] and MÁRTON TOKODI[1,9]*

[1] Heart and Vascular Center, Semmelweis University, Budapest, Hungary

[2] Independent Researcher, Budapest, Hungary

[3] Department of Cardiology, Mie University Hospital, Tsu, Japan

[4] Department of Cardiovascular Medicine, Saga University, Saga, Japan

[5] Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest, Hungary

[6] MedStar Health Research Institute, Washington, District of Columbia, USA

[7] Noninvasive Cardiac Imaging Laboratory, University of Chicago, Chicago, Illinois, USA

[8] Department of Cardiovascular Medicine, Tobata General Hospital, Kitakyushu, Japan

[9] Department of Experimental Cardiology and Surgical Techniques, Semmelweis University, Budapest, Hungary

Received: December 13, 2025 • Revised manuscript received: April 11, 2026 • Accepted: May 2, 2026

ABSTRACT

Background: Deep learning (DL) is increasingly used for the automated analysis of echocardiographic videos. These end-to-end pipelines often incorporate models targeting common preprocessing tasks to ensure the selection of videos and frames suitable for their primary analytic objective. However, publicly available models supporting these core steps remain limited.

Objectives: We aimed to develop three DL models for distinct preprocessing tasks: Model-V for view classification, Model-O for orientation classification, and Model-CC for identifying end-diastolic (ED) and end-systolic (ES) frames.

Methods: Model-V was trained and internally validated on 13,864 videos and externally evaluated on 5,650 videos. Model-O and Model-CC were trained and internally validated on 5,513 and 3,108 apical 4-chamber view (A4C) videos, respectively, and both were externally tested on the same set of 1,200 A4C videos.

Results: Model-V discriminated A4C from non-A4C with balanced accuracies of 0.991 (0.987–0.994) and 0.895 (0.884–0.907), and Model-O distinguished Stanford from Mayo orientation with balanced accuracies of 0.995 (0.990–1.000) and 0.960 (0.948–0.971) during internal and external testing, respectively. Model-CC identified ED frames with mean absolute errors (MAEs) of 4.041 (3.878–4.210) and 3.117 (3.044–3.193) frames and ES frames with MAEs of 2.799 (2.708–2.894) and 4.827 (4.738–4.917) frames in the internal and external test sets, while failing to identify only 7.3 and 4.2% of ED frames and 1.2 and 3.6% of ES frames, respectively.

Conclusions: The proposed models demonstrated good performance across key preprocessing tasks; thus, they constitute readily integrable components for end-to-end DL pipelines in automated echocardiographic analysis.

IMAGING (2026)
DOI: 10.1556/1647.2026.00404

*Corresponding author. Heart and Vascular Center, Semmelweis University, 68 Városmajor Street, Budapest 1122, Hungary.
Tel.: +36 30-877-4434.
E-mail: tokmarton@gmail.com
E-mail: tokodi.marton@semmelweis.hu

AKJournals

KEYWORDS

echocardiography, deep learning, artificial intelligence, view classification, cardiac cycle phase detection

Introduction

Deep learning (DL) has been increasingly used for the automated analysis and interpretation of echocardiograms, offering the potential to streamline workflows, reduce operator dependence, and improve the reproducibility of measurements [1–3]. DL models designed for high-level tasks, such as assessing systolic or diastolic function or predicting disease progression and adverse outcomes, typically rely on a series of preprocessing steps that transform raw echocardiographic images and videos into a standardized format [4–8]. In addition to conventional preprocessing operations, such as isolating relevant regions by cropping or masking, standardizing spatial resolution through resizing or downsampling, and normalizing pixel intensities, several other tasks can be considered as part of preprocessing in a broader sense [9, 10].

View classification represents one such task. Although each echocardiographic view has its own value, the apical 4-chamber view (A4C) is particularly informative because it simultaneously visualizes both atria and ventricles, making it indispensable for assessing chamber size and function [11]. Consequently, the A4C view is frequently used as the sole or primary input for DL models targeting the automated assessment of atrial and ventricular structure and function, and accurate selection of A4C is therefore a prerequisite for these applications [4, 5, 12–15]. Importantly, centers differ in their preferred left-right orientation when acquiring A4C recordings: most use the Stanford orientation, in which the left ventricle (LV) appears on the right side of the image and the right ventricle (RV) on the left, whereas some centers prefer the Mayo orientation, where the LV is displayed on the left side and the RV on the right. Given that orientation differences may adversely affect models that are not orientation-agnostic (i.e., not trained to handle both orientations), reliable orientation classification is crucial for enabling horizontal flipping before A4C videos are supplied to those models.

Another important preprocessing step is temporal standardization, which may involve identifying key frames within the cardiac cycle, specifically end-diastole (ED) and end-systole (ES) [4, 12, 16]. By cutting echocardiographic videos at these specific time points, they can be split into individual cardiac cycles, allowing for temporal alignment and providing a standardized input for the subsequent primary analysis step. This standardization is not only relevant for DL models but also benefits other quantitative image analysis techniques, such as the radiomic analysis of echocardiographic videos, which frequently rely on ED and ES frames to extract meaningful and reproducible features [17, 18].

Despite the indisputable importance of the aforementioned preprocessing tasks, publicly available and thoroughly validated models for performing them remain scarce. Motivated by this, we developed, extensively validated, and released three dedicated DL models for view classification, orientation classification, and ED/ES frame identification (Fig. 1), providing the research community with reliable, ready-to-use components for constructing robust end-to-end DL pipelines for echocardiographic video analysis.

Methods

Preprocessing of echocardiographic videos

Echocardiographic videos were exported in a Digital Imaging and Communications in Medicine (DICOM) file format. All frames were extracted from these files, converted to single-channel grayscale images, and cropped using the bounding box of the sector-shaped region of interest. Next, a binary mask was generated by tracking frame-to-frame changes in the intensity value of each pixel. Pixels that changed their intensity value by less than five or in fewer than 10% of the total frames were set to black, while all other pixels were set to white. To avoid multiple disjoint white patches in the binary mask, only the contour with the largest area was retained, and a convex hull was fitted around it. Then, the bounding box enclosing all white pixels in the binary mask was used to crop the mask and each frame of the corresponding video.

View and orientation classifiers – Model-V and Model-O

To determine the view and orientation of echocardiographic videos, two classifiers of the same architecture were trained: Model-V for view classification and Model-O for orientation classification. Both models operated at the frame level, and frames and binary masks used as input were resized to 256×256 pixels. To derive video-level classifications, frame-wise predictions were averaged in each video. The architecture used for both models was ResNext50 [19]. The models were trained with a batch size of 32. To improve model generalization, several data augmentation techniques were applied: (1) random rotation with up to 10° in either direction, (2) random cropping while ensuring that at least 80% of the original frame was retained, and (3) horizontal flipping with a probability of 0.5. Both models were trained for 80 epochs, with early stopping implemented based on validation performance, using a patience of 10 epochs to prevent overfitting. Additional model-specific details are presented in the following subsections.

View classifier – Model-V. Model-V aimed to predict the view of echocardiographic videos. It was designed to classify each frame of the given video into the following three categories: (1) standard A4C, (2) RV-focused A4C, or (3) other view.

The model was trained, validated, and tested on an internal dataset of 13,864 videos (2,659 [19%] standard A4C, 1,152 [8%] RV-focused A4C, and 10,053 [73%] other views) from 1,031 transthoracic echocardiographic examinations performed between November 2013 and March 2021 at the Heart and Vascular Center of Semmelweis University (Budapest, Hungary) (Supplementary Tables 1 and 2). For training, validation, and testing, this dataset was split at the video level into three sets in an approximately 65:15:20 ratio. The model was also tested externally on 5,650 videos (891 [16%] standard A4C, 309 [5%] RV-focused A4C, and 4,450 [79%] other views) from 150 healthy adults enrolled in the World Alliance of Societies of Echocardiography (WASE) Normal Values Study [20] and 300 participants of the Budakalász Epidemiology Study (Supplementary Tables 3 and 4) [21]. All videos of the internal and external datasets were reviewed and labeled by an experienced echocardiographer.

The output of Model-V is a probability distribution across three classes. A frame is classified as a non-A4C view if the probability assigned to the non-A4C (i.e., 'other') class exceeds a predefined threshold. Experiments were performed using both the default threshold of 0.5 and an optimized threshold determined on the internal validation set using Youden's J statistic. If the probability does not exceed the selected threshold, the frame is classified as either a standard or an RV-focused A4C view, depending on which of these two classes has the higher predicted probability.

Model performance was evaluated using the area under the receiver operating characteristic curve, accuracy, balanced accuracy, precision, recall, and F1 score. Metrics were computed with both the default and optimized thresholds for the 3-class multiclass classification and for discriminating the combined A4C classes from the other view class (i.e., binary classification).

Orientation classifier – Model-O. Model-O was trained to classify whether the A4C videos were acquired in the Stanford orientation (LV on the right and RV on the left side) or the Mayo orientation (LV on the left and RV on the right side).

A dual-center dataset comprising 5,513 A4C videos (3,527 [64%] with a Mayo and 1,986 [36%] with a Stanford orientation) from 1,418 transthoracic echocardiographic studies was used to train, validate, and test the model internally (Supplementary Tables 1 and 2). These studies were acquired between November 2013 and March 2021 at the Heart and Vascular Center of Semmelweis University (Budapest, Hungary) or between January 2014 and December 2022 at the University Hospital of the University of Occupational and Environmental Health (Kitakyushu, Japan). The dual-center dataset was split at the video level into training, validation, and testing sets in an 80:10:10 ratio. The model was also tested externally on 1,200 A4C videos (563 [47%] with a Mayo and 637 [53%] with a Stanford orientation) from the same 450 individuals used for the external validation of Model-V (Supplementary Tables 3 and 4). All videos of the dual-center (i.e., internal) and external datasets were reviewed and labeled by an experienced echocardiographer.

Model performance was evaluated using the area under the receiver operating characteristic curve, accuracy, balanced accuracy, precision, recall, and F1 score. Performance metrics were calculated using both the default threshold of 0.5 and an optimized threshold determined on the internal validation set using Youden's J statistic.

Identifying end-diastolic and end-systolic frames – Model-CC

To identify ED and ES frames, a DL model – Model-CC – was trained using Ultrasound Video Transformers (UVT), a previously published transformer-based architecture [16]. UVT comprises the following components: a ResNet Autoencoder (ResNetAE) [22] for extracting essential spatial features, a Bidirectional Encoder Representations from Transformers (BERT) [23] model adapted for token classification to capture the temporal dependencies between frames, and a regressor composed of fully connected layers to provide frame-wise predictions, which are then passed through a tanh activation function to map them to the range $[-1, 1]$. This design enables UVT to process videos of any length. Each frame of the echocardiographic videos was resized to a resolution of 128×128 pixels before being processed by the model. Of note, the binary masks were not used as input. Model-CC was trained with a batch size of 16 for 100 epochs, using early stopping based on validation performance with a patience of 5 epochs to prevent overfitting.

Model-CC was trained, validated, and tested on an internal dataset containing 3,108 A4C videos from 982 transthoracic echocardiographic examinations performed between November 2013 and March 2021 at the Heart and Vascular Center of Semmelweis University (Budapest, Hungary) (Supplementary Tables 1 and 2). For training, validation, and testing, this dataset was split at the video level into three sets in a 70:10:20 ratio. The model was externally tested on the same 1,200 A4C videos from 450 individuals used for the external validation of Model-O (Supplementary Tables 3 and 4). Additionally, Model-CC was tested on 20 A4C videos from 20 patients in atrial fibrillation at the time of image acquisition (Supplementary Tables 5 and 6), who underwent transthoracic echocardiographic examination at the Heart and Vascular Center of Semmelweis University (Budapest, Hungary).

All videos were reviewed and annotated by an experienced echocardiographer. Within a given cardiac cycle, the frames with the largest and smallest LV areas were annotated as ED and ES, respectively. In each video, up to three cardiac cycles were annotated in the internal dataset and up to five in the external and atrial fibrillation datasets. As a result, the internal dataset contained 8,454 and 8,454 annotated ED and ES frames, the external dataset contained 4,324 and 4,213 annotated ED and ES frames, and the atrial fibrillation dataset contained 72 and 72 annotated ED and ES frames, respectively. For comparison with inter-reader variability, the external dataset was also annotated by a second reader who was blinded to the first reader's annotations. To avoid supplying unannotated ED or ES frames to the model during training, only frames from each annotated ED frame to the subsequent annotated ES frame were used. A reversed copy of each frame sequence was then inserted immediately after the original sequence. If the total number of frames exceeded 128 after this procedure, the video was randomly cropped to 128 frames.

Model-CC's predictions were analyzed using signal processing techniques to identify the peak values corresponding to ED and ES frames. ED frames were identified as local maxima in the model's prediction signal. To qualify as an ED frame, the local maximum must have met the following criteria: (1) its value was at least 10% of the largest value in the prediction signal, (2) its prominence was at least 15% of the absolute difference between the highest and lowest values of the prediction signal., and (3) the heart rate estimated from the mean time difference between consecutive ED frames did not exceed 200 beats per minute. ES frames were identified using a similar approach, but applied to the negative of the model's prediction signal.

The performance of Model-CC was evaluated based on (1) the temporal and frame index differences (i.e., mean absolute error [MAE] and root mean squared error) between the predicted and reference ED/ES frames, (2) the percentage of ED/ES frames not identified by the model (i.e., when no ED/ES was predicted within 400 ms – the duration of a cardiac cycle at a heart rate of 150 beats per minute – from the corresponding reference frame), and (3) the correlation and agreement between heart rates estimated from the predicted ED/ES frames and the actual heart rates, quantified using intraclass correlation coefficients, coefficients of determination, and Bland-Altman-derived bias and limits-of-agreement width. To evaluate the effects of frame rate and actual heart rate on performance, their associations with errors were assessed using Spearman's correlation in the external dataset. Furthermore, Model-CC's performance was also compared with that of a previously published, publicly available DL model for multibeat cardiac cycle phase detection [24]. Lastly, to evaluate the impact of Model-CC's error on a downstream clinical task, changes in ED and ES LV areas and LV ejection fraction (LVEF) derived from predicted rather than human-selected reference frames were quantified. In these experiments, the LV in the selected frames was segmented using a previously published DL model [25].

Software packages

Data preprocessing and DL algorithms were implemented in Python (version 3.9.16, Python Software Foundation, Wilmington, Delaware, USA) using the PyTorch (version 2.0.1) Lightning (version 2.1.3) libraries. Models were trained and evaluated on a computer equipped with a single NVIDIA GeForce RTX 3090 graphics processing unit.

Code and model availability

To ensure transparency, reproducibility, and the free use of the models for research purposes, the source code is publicly available at https://github.com/szadam96/dl-models-for-echo-video-preprocessing. In addition to the scripts required for preprocessing, model training, and evaluation, all model weights are also provided in the repository.

Ethical approval

The study protocol conforms with the principles outlined in the Declaration of Helsinki, and it was approved by the Regional and Institutional Committee of Science and Research Ethics of Semmelweis University (approval number: 190/2020) and the ethics committees of all other participating centers.

Results

Performance of the view classifier – Model-V

Using the default threshold, Model-V achieved balanced accuracies of 0.929 (95% CI: 0.913–0.945) and 0.802 (95% CI: 0.783–0.822) in the 3-class classification task during internal and external testing (Table 1), whereas in binary classification, it discriminated A4C from non-A4C with balanced accuracies of 0.989 (95% CI: 0.985–0.993) and 0.864 (95% CI: 0.852–0.877) in the internal and external test sets, respectively (Table 2). After threshold optimization, Model-V achieved balanced accuracies of 0.931 (95% CI: 0.915–0.947) and 0.835 (95% CI: 0.816–0.853) for 3-class classification (Table 1), and balanced accuracies of 0.991 (95% CI: 0.987–0.994) and 0.895 (95% CI: 0.884–0.907) for binary classification on the internal and external test sets, respectively (Table 2). Other performance metrics are reported in Tables 1 and 2, and confusion matrices are shown in Fig. 2.

Analysis of performance across subgroups defined by age (<40 years, 40–64.9 years, and ≥65 years), sex (male and female), race (White, Asian, and Black; external test set only), body mass index (BMI; <25 kg/m^2, 25–29.9 kg/m^2, and ≥30 kg/m^2), and LVEF (<50% and ≥50%; internal test set only) showed that Model-V performed best in individuals younger than 40 years and in Asian and Black individuals, with performance slightly decreasing as BMI increased, while remaining relatively consistent across sex and LVEF categories (Supplementary Tables 7, 8, 9, and 10).

Performance of the orientation classifier – Model-O

Using the default threshold, Model-O classified video orientation with balanced accuracies of 0.995 (95% CI: 0.990–1.000) and 0.957 (95% CI: 0.947–0.968) in the internal and external test sets, respectively. After threshold optimization, it achieved balanced accuracies of 0.995 (95% CI: 0.990–1.000) and 0.960 (95% CI: 0.948–0.971) on these two test sets (Table 3). Additional performance metrics are reported in Table 3, and confusion matrices are shown in Fig. 3.

Subgroup analyses showed consistent performance across age, sex, BMI, and LVEF categories (Supplementary Tables 11 and 12).

Table 1. Performance of Model-V in the internal and external test sets (3-class classification)

	Accuracy	Balanced accuracy	Micro-avg. precision	Macro-avg. precision	Micro-avg. recall	Macro-avg. recall	Micro-avg. F1	Macro-avg. F1	Micro-avg. AUC	Macro-avg. AUC
Internal test set										
Default threshold	0.966 (0.959–0.972)	0.929 (0.913–0.945)	0.966 (0.959–0.972)	0.910 (0.893–0.928)	0.966 (0.959–0.972)	0.929 (0.913–0.945)	0.966 (0.959–0.972)	0.919 (0.903–0.935)	0.996 (0.995–0.998)	0.994 (0.992–0.996)
Optimized threshold	0.966 (0.960–0.973)	0.931 (0.915–0.947)	0.966 (0.960–0.973)	0.909 (0.892–0.927)	0.966 (0.960–0.973)	0.931 (0.915–0.947)	0.966 (0.960–0.973)	0.920 (0.904–0.936)	–	–
External test set										
Default threshold	0.926 (0.920–0.932)	0.802 (0.783–0.822)	0.926 (0.920–0.932)	0.892 (0.876–0.908)	0.926 (0.920–0.932)	0.802 (0.783–0.822)	0.926 (0.920–0.932)	0.841 (0.824–0.857)	0.988 (0.986–0.989)	0.977 (0.972–0.981)
Optimized threshold	0.934 (0.928–0.940)	0.835 (0.816–0.853)	0.934 (0.928–0.940)	0.879 (0.863–0.895)	0.934 (0.928–0.940)	0.835 (0.816–0.853)	0.934 (0.928–0.940)	0.854 (0.839–0.870)	–	–

Performance metrics are reported with 95% confidence intervals calculated from 10,000 stratified bootstrap resamples.
AUC – area under the receiver operating characteristic curve.

Performance of the cardiac cycle phase detection model – Model-CC

Model-CC identified ED frames with MAEs of 4.041 (95% CI: 3.878–4.210) and 3.117 (95% CI: 3.044–3.193) frames and ES frames with MAEs of 2.799 (95% CI: 2.708–2.894) frames and 4.827 (95% CI: 4.738–4.917) frames in the internal and external test sets, respectively (Table 4). Although these values fall within an acceptable range, they were higher than the mean absolute inter-reader differences, which were 0.740 (95% CI: 0.663–0.822) frames for ED frames and 1.663 (95% CI: 1.569–1.763) frames for ES frames in the external test set. Importantly, the model was unable to identify only 7.3 and 4.2% of the ED frames and 1.2 and 3.6% of the ES frames in the internal and external test sets, respectively (Table 4). In the atrial fibrillation dataset, Model-CC achieved similar performance in identifying ES frames (MAE: 4.143 [95% CI: 3.357–5.057] frames, unidentified: 2.8%) but lower performance in identifying ED frames (MAE: 6.433 [95% CI: 5.200–7.733] frames, unidentified: 16.7%) compared with the external test set (Supplementary Table 13).

Model-CC's absolute error (in frames) showed moderate positive correlations with frame rate for both ED and ES frames ($\rho = 0.533$ and $\rho = 0.665$, both $P < 0.001$). After conversion to milliseconds by normalizing to frame rate, these associations became very weak for ED frames ($\rho = -0.088$, $P < 0.001$) and non-significant for ES frames ($\rho = 0.009$, $P = 0.577$). The model's absolute error (in frames) showed weak negative correlations with heart rate for both ED and ES frames ($\rho = -0.146$ and $\rho = -0.221$, both $P < 0.001$). These correlations remained weak but became positive for both ED and ES frames ($\rho = 0.205$ and $\rho = 0.192$, both $P < 0.001$) after normalizing errors to the length of the cardiac cycle.

In the internal test set, the model tended to perform better for ED frames in individuals younger than 40 years, females, and those with lower BMI and preserved LVEF, and for ES frames in individuals younger than 40 years and those with preserved LVEF (Supplementary Tables 14 and 15). In the external test set, performance was highest among individuals aged 40–64.9 years and White individuals for both ED and ES frames (Supplementary Tables 16 and 17).

The actual heart rate could be estimated with MAEs of 4.106 (95% CI: 3.586–4.698) and 1.911 (95% CI: 1.665–2.187) beats per minute based on the predicted ED frames and with MAEs of 2.460 (95% CI: 2.190–2.759) and 1.757 (95% CI: 1.517–2.030) beats per minute based on the predicted ES frames in the internal and external test sets, respectively (Table 4). Moreover, strong agreement and correlations were observed between the predicted and actual heart rates in both test sets (Table 4, Fig. 4).

In comparison with the previously published model by Lane et al. [24], Model-CC showed slightly higher errors (MAE: 3.117 [95% CI: 3.044–3.193] vs. 2.249 [95% CI: 2.168–2.333] frames and 4.827 [95% CI: 4.738–4.917] vs. 3.684 [95% CI: 3.594–3.772] frames for ED and ES frames, respectively) but no meaningful difference in the proportion

Table 2. Performance of Model-V in the internal and external test sets (binary classification)

	Accuracy	Balanced accuracy	Precision	Recall	F1	AUC
Internal test set						
Default threshold	0.987 (0.983–0.991)	0.989 (0.985–0.993)	0.962 (0.950–0.974)	0.993 (0.987–1.000)	0.977 (0.970–0.984)	0.998 (0.997–0.999)
Optimized threshold	0.988 (0.983–0.992)	0.991 (0.987–0.994)	0.960 (0.948–0.972)	0.997 (0.995–1.000)	0.987 (0.971–0.986)	–
External test set						
Default threshold	0.935 (0.929–0.941)	0.864 (0.852–0.877)	0.942 (0.928–0.957)	0.741 (0.717–0.765)	0.829 (0.812–0.846)	0.979 (0.975–0.983)
Optimized threshold	0.946 (0.941–0.952)	0.895 (0.884–0.907)	0.928 (0.913–0.944)	0.808 (0.785–0.830)	0.864 (0.849–0.879)	–

Performance metrics are reported with 95% confidence intervals calculated from 10,000 stratified bootstrap resamples.
AUC – area under the receiver operating characteristic curve.

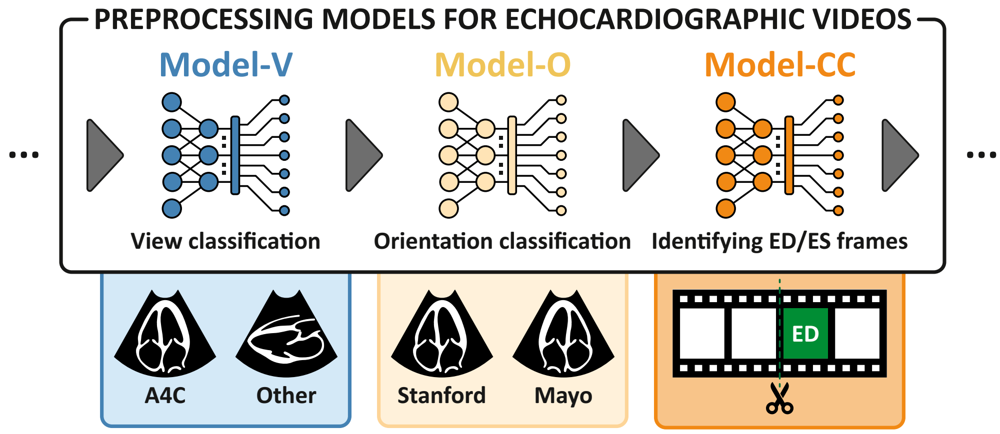

Fig. 1. DL models for automated view and orientation classification and cardiac cycle phase detection in echocardiography. In the present study, we developed and validated three DL models targeting preprocessing tasks commonly incorporated into DL pipelines for the automated analysis of echocardiograms: Model-V for view classification (A4C vs. other views), Model-O for orientation classification (Stanford vs. Mayo orientation), and Model-CC for identifying ED and ES frames. A4C – apical 4-chamber view, DL – deep learning, ED – end-diastolic, ES – end-systolic

of unidentified frames (4.2% vs. 3.6% and 3.6% vs. 3.7% for ED and ES frames, respectively; Supplementary Table 18). Interestingly, the lower error of the previously published model did not translate into superior performance in heart rate estimation compared with Model-CC (Supplementary Table 18).

Segmenting predicted rather than human-selected reference frames yielded mean differences of 4.294 (95% CI: 3.936–4.653)% and 4.727 (95% CI: 4.307–5.148)% for the ED and ES LV areas in the internal test set, and 5.183 (95% CI: 4.767–5.622)% and 8.145 (95% CI: 7.664–8.635)% for the ED and ES LV areas in the external test set. Mean absolute differences in LVEF were 3.877 (95% CI: 3.498–4.255) percentage points when calculated from LV areas segmented in predicted ED but reference ES frames, 4.339 (95% CI: 3.966–4.721) percentage points when calculated using reference ED but predicted ES frames, and 5.271 (95% CI: 4.791–5.756) percentage points when calculated using predicted ED and ES frames in the internal test set. Corresponding values in the external test set were 3.729 (95% CI: 3.369–4.117) percentage points, 5.611 (95% CI: 5.224–6.018) percentage points, and 6.178 (95% CI: 5.691–6.708) percentage points, respectively.

Discussion

In the present study, we introduced three DL models targeting preprocessing tasks commonly incorporated into DL pipelines for the automated analysis of echocardiograms: Model-V for view classification, Model-O for orientation classification, and Model-CC for identifying ED and ES frames. Across both internal and external test sets, all three models demonstrated good performance. Although the models' performance showed some degradation in certain subgroups, particularly those with a higher likelihood of poor acoustic windows and suboptimal image quality (e.g., obese patients), it remained acceptable across all subsets. We made the code and model weights available in a public repository (https://github.com/szadam96/dl-models-for-echo-video-preprocessing), providing the research community with reliable, ready-to-use components for constructing

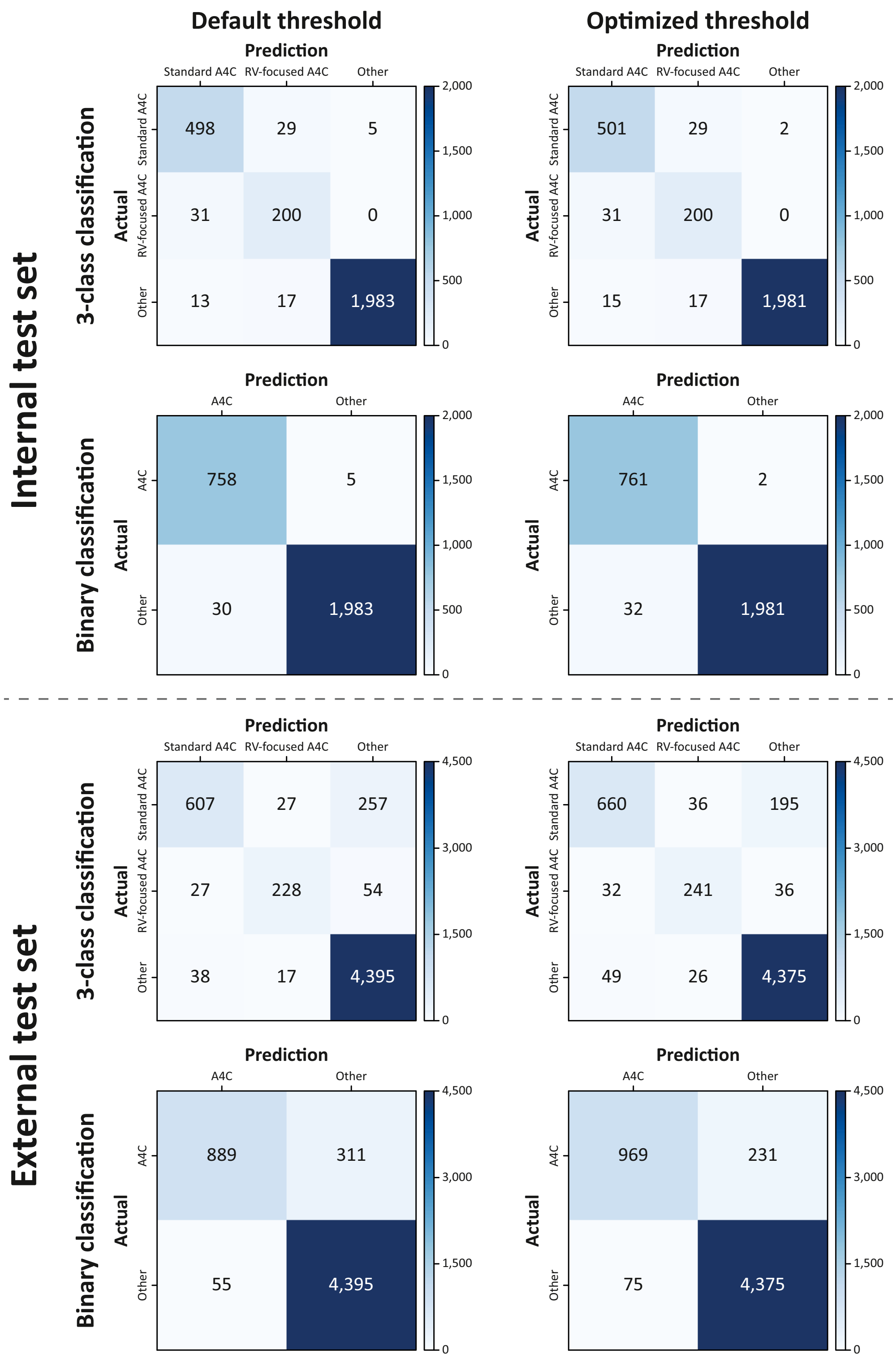

Fig. 2. Confusion matrices showing the classification performance of Model-V. Confusion matrices are presented for both the binary and 3-class view classification tasks in the internal test set (upper panel) and the external test set (lower panel). A4C – apical 4-chamber view, RV – right ventricle

Table 3. Performance of Model-O in the internal and external test sets

	Accuracy	Balanced accuracy	Precision	Recall	F1	AUC
Internal test set						
Default threshold	0.996 (0.993–1.000)	0.995 (0.990–1.000)	1.000 (1.000–1.000)	0.990 (0.980–1.000)	0.995 (0.990–1.000)	1.000 (1.000–1.000)
Optimized threshold	0.996 (0.993–1.000)	0.995 (0.990–1.000)	1.000 (1.000–1.000)	0.990 (0.980–1.000)	0.995 (0.990–1.000)	–
External test set						
Default threshold	0.956 (0.945–0.967)	0.957 (0.947–0.968)	0.926 (0.908–0.945)	0.984 (0.975–0.993)	0.954 (0.943–0.966)	0.987 (0.981–0.993)
Optimized threshold	0.958 (0.981–0.993)	0.960 (0.948–0.971)	0.934 (0.916–0.952)	0.980 (0.968–0.993)	0.957 (0.945–0.968)	–

Performance metrics are reported with 95% confidence intervals calculated from 10,000 stratified bootstrap resamples.
AUC – area under the receiver operating characteristic curve.

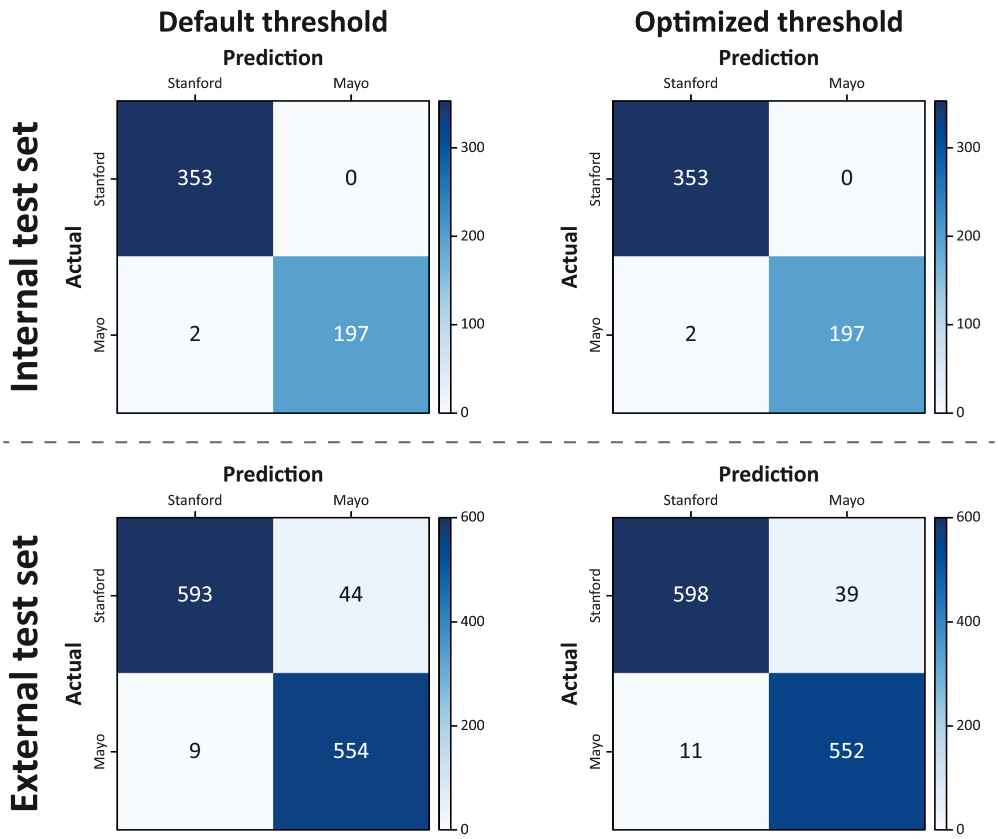

Fig. 3. Confusion matrices showing the classification performance of Model-O. Confusion matrices are presented for orientation classification in the internal test set (upper panel) and the external test set (lower panel)

end-to-end DL pipelines for echocardiographic video analysis.

An essential and advantageous attribute of the proposed models is their ability to process echocardiographic videos of arbitrary length, as well as their modular design, which allows them to be used as standalone tools or to be integrated into existing pipelines, either individually or sequentially as part of a multi-step preprocessing chain.

Besides the modular design and public availability of the models, another major strength of our study is the use of an extensive external test set, comprising videos sampled from the WASE Normal Values Study and the Budakalász Epidemiology Study. The former was a prospective, multicenter, cross-sectional study that enrolled 2,262 healthy adults between 2016 and 2019 across 19 centers in 15 countries spanning six continents, with the overarching goal of establishing reference ranges for echocardiographic parameters [20, 26]. In contrast, the latter study was a community-based, voluntary screening program conducted in central Hungary between 2011 and 2014, with a total of

Table 4. Performance of Model-CC in the internal and external test sets

	MAE (frames)	RMSE (frames)	MAE (ms)	RMSE (ms)	Not found	MAE – HR prediction (bpm)	RMSE – HR prediction (bpm)	ICC – HR prediction	R^2 – HR prediction
Internal test set									
End-diastolic frames	4.041 (3.878–4.210)	5.253 (4.975–5.544)	76 (73–78)	96 (91–101)	7.3% (124/1,704)	4.106 (3.586–4.698)	9.778 (7.414–12.198)	0.830 (0.774–0.888)	0.552 (0.318–0.737)
End-systolic frames	2.799 (2.708–2.894)	3.414 (3.266–3.577)	53 (51–55)	63 (61–66)	1.2% (21/1,704)	2.460 (2.190–2.759)	5.452 (4.217–6.916)	0.938 (0.910–0.960)	0.865 (0.788–0.917)
External test set									
End-diastolic frames	3.117 (3.044–3.193)	3.992 (3.824–4.172)	65 (63–66)	78 (76–81)	4.2% (183/4,324)	1.911 (1.665–2.187)	4.954 (3.940–5.964)	0.915 (0.879–0.945)	0.827 (0.753–0.888)
End-systolic frames	4.827 (4.738–4.917)	5.663 (5.541–5.790)	97 (96–99)	107 (105–109)	3.6% (153/4,213)	1.757 (1.517–2.030)	4.803 (3.411–6.167)	0.926 (0.887–0.960)	0.839 (0.739–0.918)

Performance metrics are reported with 95% confidence intervals calculated from 10,000 stratified bootstrap resamples.
bpm – beats per minute, ICC – intraclass correlation coefficient, HR – heart rate, MAE – mean absolute error, RMSE – root mean squared error, R^2 – coefficient of determination.

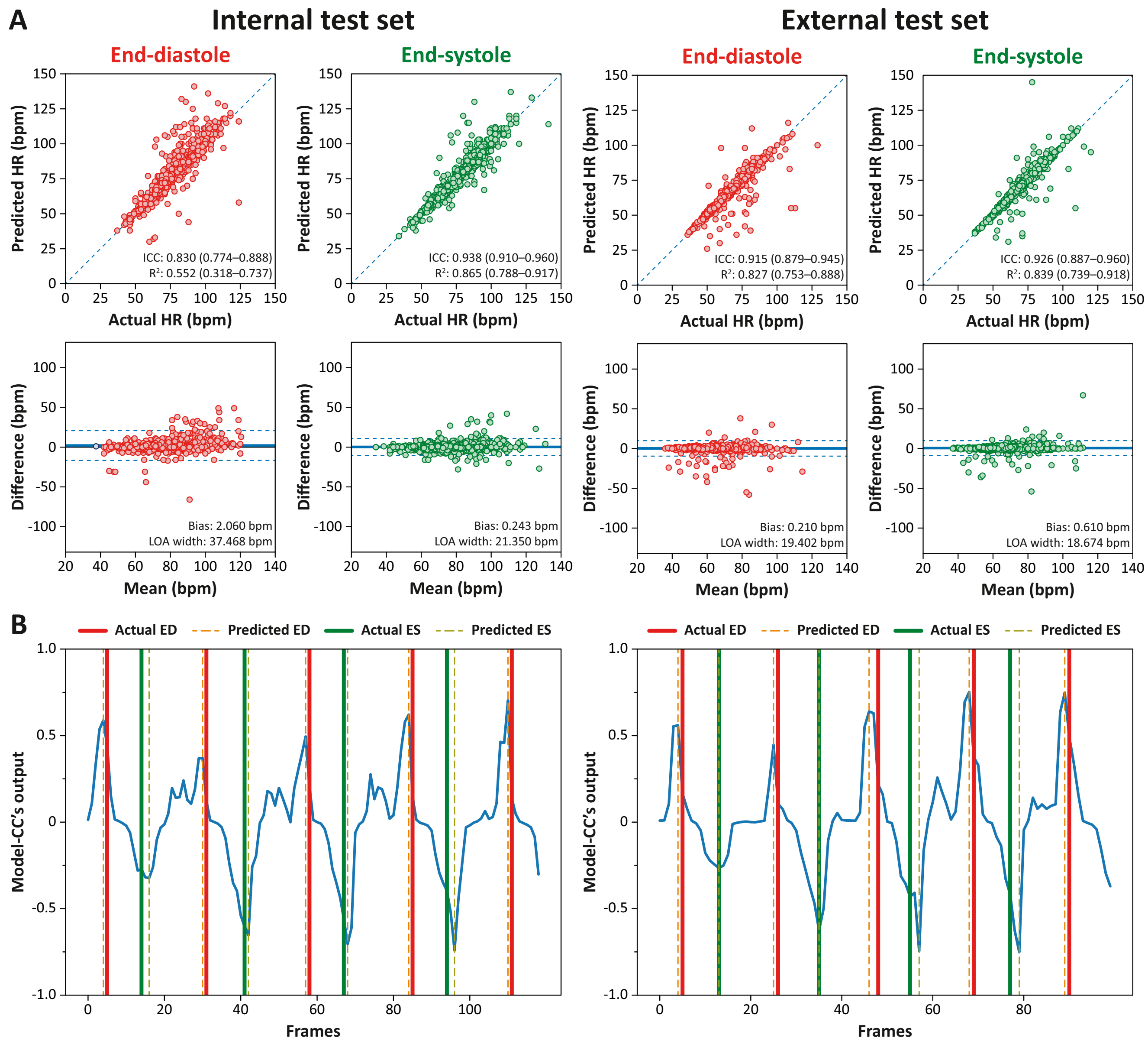

Fig. 4. Performance of Model-CC in the internal and external test sets. (A) Performance of Model-CC evaluated based on the correlation and agreement between heart rates estimated from the predicted ED/ES frames and the actual heart rates, visualized using scatter and Bland-Altman plots. (B) Outputs of Model-CC for two example videos. bpm – beats per minute, ED – end-diastolic, ES – end-systolic, HR – heart rate, ICC – intraclass correlation coefficient, LOA – limits of agreement, R^2 – coefficient of determination

2,420 adult participants [21]. Thus, the external test set derived from the data collected in these studies provided an exceptionally diverse testing environment, characterized by wide variability in ultrasound vendors, acquisition protocols, technical settings, patient demographics (i.e., age, sex, and race/ethnicity), and cardiovascular risk factors.

For many years, view classification has been a crucial preprocessing step in automated echocardiographic analysis pipelines [25, 27]. Although reliance on this step may decrease somewhat with the emergence of view-agnostic models [28], accurate view identification will likely remain essential in workflows where views are analyzed separately in downstream steps. Likewise, orientation classification will remain highly relevant, particularly when orientation-sensitive models (e.g., those relying on the detection of specific anatomical landmarks) are deployed at institutions where the left–right orientation of A4C echocardiographic videos is not standardized or differs from that of the training center. Taken together, these considerations underscore that both Model-V and Model-O address tasks that will remain important for the foreseeable future and thus constitute valuable additions to the armamentarium of publicly available tools for echocardiographic video analysis.

Given that echocardiographic videos may vary substantially in length (i.e., may contain different numbers of cardiac cycles) and are often acquired without electrocardiographic gating (i.e., they may begin and end at arbitrary time points in the cardiac cycle), accurate identification of ED and ES frames is essential for temporal standardization during preprocessing. Several different approaches have been employed previously to accomplish this task, including manifold learning [29], non-negative matrix factorization [30], convolutional neural networks [31], and hybrid architectures that combine convolutional and recurrent neural networks [24]. In this context, the transformer-based UVT architecture used in Model-CC represents a notable advancement, offering improved capabilities to model temporal dependencies in echocardiographic videos [16]. Nevertheless, Model-CC still exhibited higher prediction errors than inter-reader variability or the DL model of Lane et al. [24], even though its performance fell within a clinically acceptable range. This was also confirmed by a downstream analysis, which showed that using Model-CC-identified frames instead of human-selected ones for LV segmentation resulted in only a 3–7 percentage-point difference in LVEF. Notably, it was also demonstrated that frame rate had a negligible impact, while heart rate had a weak impact on Model-CC's performance, providing evidence of its relative resilience to these factors.

Limitations

Despite the strengths discussed above, the present study has some limitations that should be acknowledged. First, Model-V and Model-CC were trained on data originating from a single tertiary center, whereas Model-O was trained on a dataset from two tertiary centers. As such, the training datasets primarily reflect the patient populations and acquisition practices of these institutions. Nevertheless, all three models demonstrated acceptable performance during external testing, supporting their generalizability and cross-healthcare-system reliability. Second, the external test set included only individuals with preserved LVEF, precluding the comprehensive assessment of model performance in patients with reduced LVEF. Although such patients were included in internal testing, additional studies are needed to confirm model robustness in this population. Third, Model-V was designed to discriminate A4C from non-A4C views only, rather than performing complete multi-view classification. This design choice was intentional as A4C is one of the most important echocardiographic views, serving as the foundation for numerous DL applications and clinical tasks, including the assessment of atrial and ventricular sizes and function. Focusing the model exclusively on A4C identification reduced model complexity, lowered the required number of training examples per class, decreased computational cost, and potentially enabled more accurate detection of this single high-value view. Fourth, although Model-CC demonstrated acceptable overall performance, it is far from being perfect. A substantial generalization gap was observed between internal and external test sets, likely reflecting differences in acquisition settings and patient populations, although performance remained within an acceptable range. In addition, prediction error was higher for ED frame identification in atrial fibrillation, indicating reduced accuracy in the presence of irregular rhythm. Accordingly, caution is warranted when applying this model in such patients. Fifth, when analyzing the impact of Model-CC's error on a downstream clinical task, imperfections in the publicly available DL model used for segmentation [25] may have contributed to the observed differences in LV areas and LVEF and should be considered when interpreting these results. Moreover, although this analysis provides valuable insight into the impact of Model-CC's prediction error in a relevant use case, its clinical implications should ultimately be evaluated in the context of the intended downstream application prior to deployment. Finally, although all three models were evaluated on a large and diverse external dataset, additional validation across different ultrasound systems, acquisition protocols, clinical settings, and specific patient populations would be required to provide further evidence of their robustness and generalizability.

Conclusions

We successfully developed three DL models to perform key preprocessing tasks (i.e., view and orientation classifications and ED/ES frame identification) commonly required in end-to-end pipelines for automated echocardiogram analysis. Notably, all three models demonstrated good performance during internal and external testing, underscoring their suitability as ready-to-use components for such pipelines. By making the source code and model weights publicly available, we aimed to support the research community and promote further innovation in this domain.

Authors contribution: Ádám Szijártó, Attila Kovács, and Márton Tokodi conceptualized and designed the study. Tetsuji Kitano, Yosuke Nabeshima, Bálint Károly Lakatos, Alexandra Fábián, Zsuzsanna Ladányi, Luca Szávai, Máté Tolvaj, Andrea Ferencz, Tímea Katalin Turschl, Zsolt Bagyura, Federico M. Asch, Karima Addetia, Masaaki Takeuchi, Attila Kovács, and Márton Tokodi contributed to data acquisition and annotation. Ádám Szijártó, Thomas Akos Szeier, Bálint Magyar, and Márton Tokodi developed the source code, conducted the experiments, and performed the statistical analyses. Béla Merkely, Attila Kovács, and Márton Tokodi interpreted the results. Ádám Szijártó and Márton Tokodi drafted the manuscript. Béla Merkely and Attila Kovács critically reviewed the initial draft and contributed substantially to the refinement of the manuscript. All authors reviewed the final version of the manuscript and agreed to submit it to IMAGING for publication.

Sources of funding: Project number RRF-2.3.1-21-2022-00004 (MILAB) has been implemented with support from the European Union. 2024-1.2.3-HU-RIZONT-2024-00057 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the 2024-1.2.3-HU-RIZONT funding scheme. Attila Kovács has received grant support from the National Research, Development and Innovation Office (NKFIH) of Hungary (FK 142573) and has been supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. Márton Tokodi has been supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

Disclosures: Ádám Szijártó, Thomas Akos Szeier, and Bálint Magyar are employees of Argus Cognitive and receive financial compensation for their work (unrelated to the content of the present manuscript). Bálint Károly Lakatos and Alexandra Fábián report personal fees from Argus Cognitive (unrelated to the content of the present manuscript). Masaaki Takeuchi has received grants from Philips Medical Systems (unrelated to the content of the present manuscript). Béla Merkely has received grants from Boston Scientific and Medtronic, as well as personal fees from Biotronik, Abbott, AstraZeneca, Novartis, and Boehringer Ingelheim (all unrelated to the content of the present manuscript). Attila Kovács reports personal fees from Argus Cognitive and CardioSight (unrelated to the content of the present manuscript). All other authors have reported that they have no relationships relevant to the content of this paper to disclose.

Béla Merkely, Attila Kovács and Márton Tokodi are members of the Editorial Board of the journal, therefore they did not take part in the review process in any capacity and the submission was handled by a different member of the editorial board. The submission was subject to the same process as any other manuscript and editorial board membership had no influence on editorial consideration and the final decision.

Ethical statement: The presented study was conducted in accordance with the principles outlined in the Declaration of Helsinki, and its protocol was approved by the Regional and Institutional Committee of Science and Research Ethics of Semmelweis University (approval number: 190/2020) and the ethics committees of all other participating centers.

Data availability: The datasets used for model training and validation are not publicly available due to patient confidentiality, ethical restrictions, and institutional data-sharing constraints. Data access may be considered upon reasonable request to the corresponding author, subject to applicable ethical approvals and institutional policies.

ACKNOWLEDGEMENTS

Some of the methods and results presented in this article are also included in the Ph.D. thesis of Ádám Szijártó.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1556/1647.2026.00404.

REFERENCES

[1] Myhre, PL, Grenne, B, Asch, F, Delgado, V, Khera, R, Lafitte, S, et al.: Artificial intelligence-enhanced echocardiography in cardiovascular disease management. Nat Rev Cardiol 2025; 23(3): 164–82.

[2] Kagiyama N, Tokodi M, Sengupta PP: Machine learning in cardiovascular imaging. Heart Fail Clin 2022; 18(2): 245–58.

[3] Tokodi M, Kovács A: A new hope for deep learning-based echocardiogram interpretation: the DROIDs you were looking for. J Am Coll Cardiol 2023; 82(20): 1949–52.

[4] Szijártó, Á, Merkely, B, Kovács, A, Tokodi, M, QUEST-EF Investigators. Deep learning-enabled echocardiographic assessment of biventricular ejection fractions: the dual-task QUEST-EF model. Eur Heart J Cardiovasc Imaging 2025; 26(8): 1402–5.

[5] Tokodi, M, Magyar, B, Soós, A, Takeuchi, M, Tolvaj, M, Lakatos, B, et al.: Deep learning-based prediction of right ventricular ejection fraction using 2D echocardiograms. JACC Cardiovasc Imaging 2023; 16(8): 1005–18.

[6] Chen, X, Yang, F, Zhang, P, Lin, X, Wang, W, et al.: Artificial Intelligence-assisted left ventricular diastolic function assessment and grading: multiview versus single view. J Am Soc Echocardiogr 2023; 36(10): 1064–78.

[7] Valsaraj, A, Kalmady, S, Sharma, V, Frost, M, Sun, W, Sepehrvand, N, et al.: Development and validation of

echocardiography-based machine-learning models to predict mortality. EBioMedicine 2023; 90: 104479.
[8] Oikonomou, EK, Holste, G, Yuan, N, Coppi, A, McNamara, R, Haynes, N, et al.: A multimodal video-based AI biomarker for aortic stenosis development and progression. JAMA Cardiol 2024; 9(6): 534–44.
[9] Sengupta, PP, Shrestha, S, Berthon, B, Messas, E, Donal, E, Tison, G, et al.: Proposed requirements for cardiovascular imaging-related machine learning evaluation (PRIME): a checklist: reviewed by the American college of cardiology healthcare innovation council. JACC: Cardiovasc Imaging 2020; 13(9): 2017–35.
[10] Kagiyama, N, Tokodi, M, Hathaway, Q, Arnaout, R, Davies, R, Dey, D, et al.: PRIME 2.0: proposed requirements for cardiovascular imaging-related multimodal-AI evaluation: an updated checklist. JACC Cardiovasc Imaging 2025; 19(2): 225–51.
[11] Mitchell, C, Rahko, P, Blauwet, L, Canaday, B, Finstuen, J, Foster, M, et al.: Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the American society of echocardiography. J Am Soc Echocardiogr 2019; 32(1): 1–64.
[12] Ouyang, D, He, B, Ghorbani, A, Yuan, N, Ebinger, J, Langlotz, C, et al.: Video-based AI for beat-to-beat assessment of cardiac function. Nature 2020; 580(7802): 252–6.
[13] Kwan, AC, Chang, E, Jain, I, Theurer, J, Tang, X, Francisco, N, et al.: Deep learning-derived myocardial strain. JACC Cardiovasc Imaging 2024; 17(7): 715–25.
[14] Hu, J, Olaisen, S, Smistad, E, Dalen, H, Lovstakken, L: Automated 2-D and 3-D left atrial volume measurements using deep learning. Ultrasound Med Biol 2024; 50(1): 47–56.
[15] Tokodi, M, He, B, Szijártó, Á, Ferencz, A, Shiida, K, Tolvaj, M, et al.: Artificial intelligence-enabled echocardiographic assessment of right ventricular function. medRxiv 2026.
[16] Reynaud, H, Vlontzos, A, Hou, B, Beqiri, A, Leeson, P, Kainz, B: Ultrasound video transformers for cardiac ejection fraction estimation. In: Medical image computing and computer assisted intervention – MICCAI 2021, 2021. Cham: Springer International Publishing.
[17] Moon, I, Lee, J, Lee, S, Jeong, D, Jeon, J, Jang, Y, et al.: Artificial Intelligence-enhanced analysis of echocardiography-based radiomic features for myocardial hypertrophy detection and etiology differentiation. Circ Cardiovasc Imaging 2025; 18(5): e017436.
[18] Hathaway, QA, Yanamala, N, Siva, N, Adjeroh, D, Hollander, J, Sengupta, P: Ultrasonic texture features for assessing cardiac remodeling and dysfunction. J Am Coll Cardiol 2022; 80(23): 2187–201.
[19] Xie, S, Girshick, R, Dollár, P, Tu, Z, He, K: Aggregated residual transformations for deep neural networks. In: 2017 IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[20] Asch, FM, Banchs, J, Price, R, Rigolin, V, Thomas, J, Weissman, N, et al.: Need for a global definition of normative echo values-rationale and design of the World alliance of Societies of echocardiography normal values study (WASE). J Am Soc Echocardiogr 2019; 32(1): 157–162.e2.
[21] Bagyura, Z, Kiss, L, Édes, E, Lux, Á, Polgár, L, Soós, P, et al.: Cardiovascular screening programme in the central Hungarian region. The Budakalász study. Orvosi Hetilap, 2014; 155(34): 1344–52.
[22] He, K, Zhang, X, Ren, S, Sun, J: Identity mappings in deep residual networks. In: Computer vision – ECCV 2016, Cham: Springer International Publishing.
[23] Devlin, J, Chang, MW, Toutanova, K: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv [cs.CL], 2019.
[24] Lane, ES, Azarmehr, N, Jevsikov, J, Howard, J, Shun-Shin, M, Cole, G, et al.: Multibeat echocardiographic phase detection using deep neural networks. Comput Biol Med 2021; 133: 104373.
[25] Zhang, J, Gajjala, S, Agrawal, P, Tison, G, Hallock, L, Beussink-Nelson, L, et al.: Fully automated echocardiogram interpretation in clinical practice. Circulation 2018; 138(16): 1623–35.
[26] Asch, FM, Miyoshi, T, Addetia, K, Citro, R, Daimon, M, Desale, S, et al.: Similarities and differences in left ventricular size and function among races and nationalities: results of the world alliance societies of echocardiography normal values study. J Am Soc Echocardiogr 2019; 32(11): 1396–1406.e2.
[27] Ghorbani, A, Ouyang, D, Abid, A, He, B, Chen, J, Harrington, R, et al.: Deep learning interpretation of echocardiograms. npj Digital Med 2020; 3(1): 10.
[28] Holste, G, Oikonomou, E, Tokodi, M, Kovács, A, Wang, Z, Khera, R: Complete AI-enabled echocardiography interpretation with multitask deep learning. Jama 2025; 334(4): 306–18.
[29] Gifani, P, Behnam, H, Shalbaf, A, Sani, Z: Automatic detection of end-diastole and end-systole from echocardiography images using manifold learning. Physiol Meas 2010; 31(9): 1091–103.
[30] Yuan, B, Chitturi, S, Iyer, G, Li, N, Xu, X, Zhan, R, et al.: Machine learning for cardiac ultrasound time series data, in SPIE Medical Imaging, 2017. 101372D.
[31] Farhad, M, Masud, M, Beg, A, Ahmad, A, Ahmed, L, Memon, S: Cardiac phase detection in echocardiography using convolutional neural networks. Scientific Rep 2023; 13(1): 8908.