SEMMELWEIS EGYETEM

DOKTORI ISKOLA

Ph.D. értekezések

3454.

SZIJÁRTÓ ÁDÁM

Szív- és érrendszeri betegségek élettana és klinikuma
című program

Programvezető: Dr. Merkely Béla, egyetemi tanár

Témavezető: Dr. Kovács Attila, egyetemi docens

Dr. Tokodi Márton, egyetemi adjunktus

DEEP LEARNING-ENABLED ECHOCARDIOGRAPHIC ASSESSMENT OF BIVENTRICULAR EJECTION FRACTIONS

PhD thesis

Szijártó Ádám

Semmelweis University Doctoral School
Cardiovascular Medicine and Research Division

Supervisors: Attila Kovács, MD, DSc,
Márton Tokodi, MD, PhD

Official reviewers: Bence Ágg, MD, PhD
Márton Áron Goda, PhD

Head of the Complex Examination Committee: Alán Alpár, MD, DSc

Members of the Complex Examination Committee: László Cervenák, MD, PhD
Róbert Langer, MD, PhD

Budapest

2026

Contents

List of Abbreviations

2D – two-dimensional

2DE – two-dimensional echocardiography

3D – three-dimensional

3DE – three-dimensional echocardiography

A4C – apical 4-chamber

ACHD – adult congenital heart disease

AI – artificial intelligence

AUC – area under the receiver operating characteristic curve

BERT – Bidirectional Encoder Representations from Transformers

CNN – convolutional neural network

CT – computed tomography

DHZ – German Heart Center Munich

DICOM – Digital Imaging and Communications in Medicine

DL – deep learning

E – early mitral inflow velocity

EF – ejection fraction

e’ – early diastolic mitral annular velocity

HF – heart failure

HR – hazard ratio

ICC – intraclass correlation coefficient

LV – left ventricle

LVEF – left ventricular ejection fraction

MAE – mean absolute error

ML – mmachine learning

MRI – magnetic resonance imaging

NLP – natural language processing

NPV – negative predictive value

PET – positron emission tomography

pp – percentage point

PPV – positive predictive value

QUEST-EF – QUantification of Echocardiographic STudies – Ejection Fraction

R^2 – coefficient of determination

RMSE – root mean squared error

ROI – region of interest

RV – right ventricle

RVEF – right ventricular ejection fraction

SSL – self-supervised learning

TR – tricuspid regurgitation

UVT – Ultrasound Video Transformer

VideoMAE – video masked autoencoder

ViT – vision transformer

ViViT – video vision transformer

WASE – World Alliance Societies of Echocardiography

1. Introduction

1.1 Limitations of convolutional neural networks

Ever since the inception of convolutional neural networks (CNNs),[1] we have witnessed unprecedented advancements in neural image processing. CNNs enable the use of large volumes of high-dimensional visual data for automatic pattern recognition by hierarchically extracting spatially invariant features through localized filter operations. For many years, CNNs were the most effective method for automatic classification, regression, or segmentation tasks. However, with the increasing amount of data and size of the models, CNNs' limitations became apparent, as their sole focus on local structures restricts their ability to learn global dependencies.[2]

1.2 Transformers

Even though the transformer architecture was introduced specifically for sequence-to-sequence natural language processing (NLP) tasks, it has almost completely taken over all the other domains and modalities in the machine learning (ML) realm in recent years.[3–5] Practically all the commercially available artificial intelligence (AI) models today rely on this architecture in some way.

Previous state-of-the-art architectures for processing sequential data, such as text, relied on recurrence to analyze sequences in their natural order, while maintaining a single internal representation of the entire sequence.[6, 7] While these methods proved highly effective for shorter sequences, a key limitation of recurrent neural networks was their tendency to place more importance on inputs later in the sequence, progressively decreasing the importance of the beginning of the sequence.[5]

Transformers overcome this limitation by employing the self-attention mechanism.[5] This approach involves generating three learned inner representations for each token of the sequence and applying the non-parametric scaled dot-product attention operation, which models a relation or "attention" between all the tokens regardless of their proximity. Since this methodology contains no recurrence or convolution, positional

encodings must be added to the initial input tokens to utilize their relative and absolute positions.

By creating a relationship matrix between all input elements, transformers can learn the global structures of the input, making them highly scalable, with models reaching hundreds of billions in parameter size.[8, 9]

1.3 Transformers for vision

Not long after the transformer architecture was introduced in NLP, its extension was developed for the image domain in the form of Vision Transformers (ViTs).[10–12] Although the previously proposed self-attention mechanism could be applied in this domain with little to no modification, pixel-wise tokenization was deemed infeasible in current hardware, due to its quadratic complexity with respect to the sequence size.[10] To overcome this, patch embedding was introduced, considerably reducing the sequence size. This approach divides the input image into fixed-size, non-overlapping rectangular patches, which are flattened into a one-dimensional sequence. The resulting sequence, combined with positional embeddings, serves as the input tokens for the model. Patch embedding can be easily extended to videos, where, along with the spatial dimensions, the temporal dimension is also included with the patching, creating three-dimensional (3D) cuboids as input tokens.[13, 14]

ViTs address one of CNNs' major limitations (i.e., the tendency to prioritize local structures while ignoring global ones) through the inherently global nature of the self-attention module, while still extracting local information via patch and positional embeddings.[10] However, due to weaker reliance on localities, ViTs require more training data than CNNs.[15] This challenge is typically addressed through pre-training on a large-scale dataset, which is followed by fine-tuning to perform a specific task.

Although natural image datasets are abundant, specialized medical imaging suffers from a scarcity of high-quality annotated data, mainly due to the time-consuming nature of manual labeling and the need for expert knowledge. Thus, supervised pre-training is unfeasible in this domain. Simultaneously, in routine clinical practice, imaging modalities generate a vast amount of data without explicit labels or annotations, alongside

labeled data originally intended for specific tasks but also potentially useful for alternative purposes.

1.4 Self-supervised learning

Self-supervised learning (SSL) is an ML paradigm, originally developed for NLP,[8, 16] that enables models to learn meaningful representations from unlabeled data. It is achieved by defining ML tasks, for which the input and output can be generated from the raw data in an automated fashion. Although these tasks may substantially differ from the model's intended purpose, they help the model learn inner representations that are also useful for the final task. Following the self-supervised training phase, components of the model exclusively dedicated to this task are replaced with new ones responsible for the final task. Then, either only the latter component or the whole model is trained in a supervised manner.

1.4.1 Contrastive learning

In computer vision, multiple approaches have been introduced to leverage SSL. One common approach is contrastive learning, in which the model is tasked with creating an embedded representation that places similar data points close together.[17–19] It is trained using a positive pair, typically the same data sample with different augmentations, and a set of negative examples drawn from the dataset at each training step. Using a contrastive loss function, the positive pairs are "pulled together" while the negative examples are "pushed apart" in the embedding space. This approach enables the model to capture meaningful semantic features without requiring explicit labels. However, contrastive learning also has some limitations. Since positive pairs are derived from the same data sample and differ only by augmentations, the method is highly sensitive to the type and quality of those augmentations. If these are not set correctly, the model only learns to differentiate the augmentations without extracting important semantic information from the images.[20] Moreover, for contrastive learning to perform optimally, a large number of negative samples is required at each training step, making it resource-intensive because more data must be held in memory simultaneously. This issue becomes even more

pronounced when transitioning from images to videos, due to the significantly larger input size.

1.4.2 Masked autoencoders

Another popular approach for visual SSL is masked autoencoders.[21–24] This technique employs an encoder–decoder architecture that aims to reconstruct the masked or missing parts of an image or video, enabling the encoder to capture meaningful representations of the input. After self-supervised training, the decoder is replaced with a classifier or regressor head for further supervised training. Because of this, masked autoencoders often employ an asymmetric design with the encoder being significantly larger than the decoder.

Due to their unique design, ViTs are well-suited for integration with the masked autoencoder paradigm.[21] The patch embedding and non-parametric attention function, which makes the model invariant to sequence length, allow masking at the token level rather than the pixel level. Consequently, instead of feeding a modified input with obstructed areas, the masked tokens are simply excluded from the input, constraining the model to process only real tokens and enabling the use of high masking ratios. Moreover, skipping the masked tokens significantly reduces computational costs, allowing for training larger models or using larger batch sizes.

Video masked autoencoder (VideoMAE)[23] is a natural extension of masked autoencoder into the video domain, utilizing cube embedding and a video vision transformer (ViViT).[14] A notable innovation in this pipeline is the tube masking strategy, which samples the tokens consistently along the temporal axis (i.e., the same tokens are masked in all the frames) instead of randomly sampling tokens across all dimensions. Tube masking accounts for temporal redundancy and prevents information leakage, allowing the model to focus on high-level information.

VideoMAE v2, an improved version of the pipeline, was introduced recently by adding a second mask for the decoder, which reconstructs only a portion of the originally masked tokens.[24] This modification substantially reduced the resources required for pre-training without negatively affecting the model's overall performance.[24]

1.5 Application of transformers and self-supervised learning in medical imaging

Medical imaging is among the most widely studied areas of application for deep learning (DL). DL techniques have achieved remarkable success in various medical image analysis tasks, including reconstruction and enhancement of images, image registration and alignment across modalities, organ and lesion segmentation, disease detection and classification, and prognosis or outcome prediction.[25–30] Their applications span multiple imaging modalities, including X-ray, computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and ultrasound.[31]

Unsurprisingly, recent advances in DL architectures, such as transformers, have been rapidly adopted in medical imaging.[31] However, these architectures' requirement for enormous datasets necessitated the use of pre-training techniques. While some researchers conducted successful experiments with supervised pre-training on natural images,[32] self-supervised pre-training was often found to be a more viable solution due to the special characteristics of the imaging modalities.[33] Using these methods, transformers were able to achieve better performance in multiple medical image analysis tasks than CNN-based models.[34]

Although there are many examples where transformers were used for medical image analysis, the application of this architecture to videos is still in its infancy. Hybrid architectures have been frequently employed, in which individual frames are first processed by a CNN-based model, and the resulting embedded sequence of frames is then processed by a transformer, leveraging its sequential modeling capabilities.[35–37] Nevertheless, there are far fewer examples of using fully transformer-based approaches that process the spatio-temporal input in its entirety, thereby harnessing all the benefits of the architecture.[38]

1.6 Deep learning-enabled echocardiographic assessment of ejection fraction

Echocardiography plays a pivotal role in modern cardiology as a non-invasive, safe, and widely accessible imaging modality that provides real-time visualization of cardiac structure and function.[39] It is indispensable for diagnosing and managing a broad spectrum of cardiovascular conditions, including valvular diseases,[40] cardiomyopathies,[41] heart failure,[42] and congenital heart diseases.[43]

A key functional parameter measured using echocardiography is ejection fraction (EF), defined both for the left and right ventricles (LV and RV).[44] EF quantifies the volumetric fraction of blood ejected from the ventricle during systole in each cardiac cycle. LVEF serves as a critical indicator of LV systolic function and is widely used to classify heart failure phenotypes and guide therapeutic decisions.[45] Although RVEF is assessed less frequently using echocardiography, its clinical value has been increasingly recognized in various cardiovascular conditions, such as pulmonary hypertension, tricuspid regurgitation, or congenital heart disease.[46, 47]

To this day, two-dimensional (2D) echocardiography (2DE) remains the most commonly used imaging modality for assessing LV and RV systolic function. Nevertheless, this imaging modality has inherent limitations: it requires multiple views to measure LVEF appropriately and does not enable the quantification of RVEF. Three-dimensional echocardiography (3DE) offers clear incremental value over 2DE by directly measuring ventricular volumes without geometric assumptions, thereby reducing variability and correlating better with cardiac MRI, the gold standard modality for volumetric assessment of cardiac chambers.[48, 49] Despite its advantages, cardiac MRI is less accessible and more costly, highlighting the clinical value of echocardiographic EF assessment. However, recent surveys have revealed that 3DE is still underutilized for assessing ventricular volumes and EFs, most commonly due to the lack of dedicated training, time constraints, and the complexity of post-processing.[50, 51]

Due to the importance of EF in clinical cardiology and the challenges related to the manual annotation and interpretation of echocardiographic images and videos, AI-enabled automated estimation of EF has been the focus of many studies in recent years

(Table 1).[53] Similar to the trend observed in general medical imaging-related research, we could also witness a steady shift from the application of CNNs to transformers in this task.

One of the first entirely CNN-based pipelines for the fully automated interpretation of echocardiograms was EchoCV.[52] It was trained on more than 14,000 echocardiograms and was intended for multiple tasks, including view classification, image segmentation, cardiac structure and function quantification, and disease detection. Importantly, it achieved a median absolute deviation of 6.0 percentage points in predicting LVEF during internal validation.

EchoNet, another CNN-based model, was trained on a single-center dataset comprising more than 2.6 million echocardiographic images from 2,850 patients to identify cardiac structures, estimate cardiac function, and predict systemic risk factors.[53] It showed a mean absolute error (MAE) of 7.0 percentage points in predicting LVEF during internal validation.

The same research group also developed another CNN-based pipeline, EchoNet-Dynamic, which used a segmentation-based approach to identify cardiac cycles and then performed direct regression to provide beat-to-beat LVEF predictions from apical 4-chamber (A4C) view echocardiographic videos.[54] The model achieved MAEs of 4.1 and 6.0 percentage points on the internal and external test sets, respectively. Besides the novel DL model, the training and internal test sets containing 10,030 A4C videos were also published, encouraging additional research in the field.[55] Moreover, a subsequent blinded, randomized clinical trial proved that the initial evaluation of LVEF by a modified version of this model was non-inferior and even superior to sonographer-guided initial assessment.[56]

RVENet was the first model that attempted to estimate 3DE-derived RVEF from A4C echocardiographic videos.[57] Using an ensemble of multiple custom spatial-temporal CNN models trained on a single-center dataset containing nearly 3,000 videos, this model achieved MAEs of 4.5 and 5.5 percentage points during internal and external validation, respectively. Moreover, the RVENet-predicted RVEF values were also associated with major cardiac events, independent of age, sex, and LV systolic function. To promote further innovation, the internal data set was made publicly available.[58]

With the advent of transformers, more and more studies have employed hybrid architectures that combine CNNs and transformers, processing individual video frames with CNNs and the resulting embedded frame sequence with transformers. The first notable model utilizing this approach was Ultrasound Video Transformer (UVT),[36] which was trained on the EchoNet-Dynamic dataset[55] to predict LVEF and identify the end-dyastolic (ED) and end-systolic (ES) frames in A4C videos. In predicting LVEF, UVT achieved an MAE of 5.3 percentage points during internal validation

By customizing the aforementioned hybrid architecture, EchoCoTr outperformed both the Echonet-Dynamic and UVT models on the EchoNet-Dynamic dataset with an MAE of 3.9 percentage points for LVEF prediction.[59] However, it is important to note that neither UVT nor EchoCoTr was validated on an independent external dataset.

Another example of this hybrid architecture is PanEcho, which was trained on 1.2 million multi-view echocardiographic videos to estimate 39 labels and measurements and predicted LVEF with MAEs of 4.2 and 4.5 percentage points on the internal and external test sets, respectively.[35]

In contrast to hybrid CNN-transformer models, EchoPrime employs a fully transformer-based design to process multiple videos from an echocardiographic study and generate a complete textual report as output.[38] The model was trained on 1.2 million video-report pairs using contrastive learning for video encoding. Despite not being trained directly for LVEF prediction, EchoPrime achieved MAEs of 4.8 and 4.1 percentage points in the internal and external test sets, respectively.

Table 1 Examples of previously published deep learning models for the automated echocardiographic assessment of ejection fractions

Model	Input	Prediction target	Architecture	Performance in EF prediction (MAE)	
				Internal	External
EchoCV[52]	Multi-view	LVEF + other variables	CNN	6.0 pp*	-
EchoNet[53]	Multi-view	LVEF + other variables	CNN	7.0 pp	
EchoNet-Dynamic[54]	A4C	LVEF	CNN	4.1 pp	6.0 pp
RVENet[57]	A4C	RVEF	CNN	4.5 pp	5.5 pp
UVT[36]	A4C	LVEF + ED/ES frames	CNN + transformer	5.3 pp	-
EchoCoTr[59]	A4C	LVEF	CNN + transformer	3.9 pp	-
PanEcho[35]	Multi-view	LVEF + other variables	CNN + transformer	4.2 pp	4.5 pp
EchoPrime[38]	Multi-view	Textual report that includes LVEF	Transformer	4.8 pp	4.1 pp

* Median absolute deviation was used as the performance metric.

A4C – apical 4-chamber view, ED – end-diastolic, EF – ejection fraction, ES – end-systolic, LVEF – left ventricular ejection fraction, MAE – mean absolute error, RVEF – right ventricular ejection fraction, CNN – convolutional neural network, pp – percentage point

2. Objectives

1. Adapting VideoMAE for analyzing echocardiographic videos

VideoMAE is a state-of-the-art SSL method designed specifically for video transformers. Nevertheless, this method was developed for the general domain of natural videos.[60] Accordingly, we hypothesized that echocardiography-specific modifications could markedly improve the method's performance, and we proposed a novel region-of-interest (ROI)-aware masking strategy that accounts for the characteristic sector-shaped ROI during mask sampling.

2. Developing an end-to-end DL pipeline to predict LVEF and RVEF from A4C echocardiographic videos

The transformer architecture has taken over computer vision research in recent years, owing to its ability to capture global dependencies. However, due to the size and complexity of this architecture, more data is required for its effective training compared to earlier architectures such as CNNs. This challenge is particularly pronounced for high-dimensional inputs such as echocardiographic videos, where high-quality annotated datasets are scarce. SSL is an ML paradigm that offers a panacea for this issue by enabling the model to learn meaningful inner representations from larger unannotated datasets, while only requiring a smaller dataset of annotated samples for fine-tuning on the final task (e.g., the automated assessment of echocardiographic parameters). Accordingly, we proposed QUEST-EF (QUantification of Echocardiographic STudies – Ejection Fraction), a dual-task DL pipeline trained with SSL incorporating our novel ROI-aware masking strategy for automated quantification of 3DE-derived LVEF and RVEF from A4C echocardiographic videos. QUEST-EF was designed as a fully automated end-to-end DL pipeline that also includes a complex preprocessing module. Besides internal testing, we sought to evaluate QUEST-EF across a diverse spectrum of acquired and congenital cardiac diseases and various geographic regions. Last, we also analyzed the associations between the predictions and clinical outcomes in a patient cohort with mixed cardiac diseases and a low-risk, community-based cohort.

3. Methods

3.1 Preprocessing of echocardiographic videos

The preprocessing phase comprised three distinct steps: (1) algorithmically cleaning and cropping the frames, (2) confirming that the video is an A4C and determining the orientation of the A4C, and (3) splitting each video into cardiac cycles.

3.1.1 Cleaning and cropping of frames

Echocardiographic videos were exported in a Digital Imaging and Communications in Medicine (DICOM) file format. All the frames were extracted from these files and converted to single-channel gray-scale images, then cropped using the bounding box of the sector-shaped ROI. Next, a binary mask was created by tracking frame-to-frame changes in each pixel's intensity value. Pixels that changed their intensity value by less than 5 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 was used to crop the binary mask and each frame of the corresponding video.

3.1.2 Confirming view and determining orientation

To confirm the view and determine orientation, 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 used as input were resized to 256x256 pixels. To derive a video-level classification, the frame-wise predictions were averaged. The architecture used for both models was ResNext50.[61] 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 degrees 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 task-specific details are presented in the following subsections.

3.1.2.1 View classification – Model-V

Model-V aimed to predict the view of 2DE videos. It was designed to classify each frame of the given video into 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). For training, validation, and testing, the dataset was split in an approximately 65:15:20 ratio at the video level. 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)[62] study and 300 participants of the Budakalász Epidemiology Study.[63] 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. If the probability of being a non-A4C (i.e., other) view exceeds a threshold that was determined on the internal validation set using Youden's J statistic, the frame is classified as a non-A4C. Otherwise, it is classified as either a standard or an RV-focused A4C, depending on which of these two classes has the higher probability.

Model performance was evaluated using accuracy, balanced accuracy, precision, recall, and F1 score. These metrics were reported both for the three-class multiclass classification and for discriminating the combined A4C classes from the other view class (i.e., binary classification). For multiclass classification, macro-averaged metrics were reported.

3.1.2.2 Orientation classification – Model-O

Model-O was trained to classify whether the A4C videos were acquired in Stanford (LV on the right and RV on the left side) or 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. 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 dataset was split in an 80:10:10 ratio at the video level for training, validation, and testing. The model was also tested externally in 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. All videos of the dual-center and external datasets were reviewed and labeled by an experienced echocardiographer.

Model performance was evaluated using accuracy, balanced accuracy, precision, recall, and F1 score. The threshold used for dichotomizing the probability predicted was optimized in the internal validation set using Youden's J statistic.

3.1.3 Splitting videos into cardiac cycles

Given that echocardiographic videos may substantially differ in length and contain different numbers of cardiac cycles, each video was split into segments containing frames from exactly one cardiac cycle. For this task, a DL model – Model-CC – was trained using Ultrasound Video Transformers (UVT), a previously published spatio-temporal architecture.[36]

3.1.3.1 Model architecture

UVT comprises the following components: a ResNet Autoencoder (ResNetAE)[64] for extracting essential spatial features, a Bidirectional Encoder Representations from Transformers (BERT)[8] 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 128x128 pixels before being processed by the model. The model was trained with a batch size of 16. Both models were trained for 100 epochs, with early stopping implemented based on validation performance, using a patience of 5 epochs to prevent overfitting.

3.1.3.2 Datasets used for model training and evaluation

Model-CC was trained, validated, and tested on an internal dataset containing 3,108 A4C videos from 991 transthoracic echocardiographic examinations performed between November 2013 and March 2021 at the Heart and Vascular Center of Semmelweis University (Budapest, Hungary). For training, validation, and testing, the dataset was split in a 70:10:20 ratio at the video level. In addition, the model was externally tested on the same 1,200 A4C videos from 450 individuals that were used for the external validation of Model-O. All videos of the internal and external datasets 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 dataset. As a result, the internal dataset contained 8,454 and 8,454 annotated ED and ES frames, whereas the external dataset contained 4,324 and 4,213 annotated ED and ES frames, respectively. To ensure that no unannotated ED or ES frames were supplied 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.

3.1.3.3 Identifying end-diastolic and end-systolic frames based on the prediction signal

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 is at least 10% of the largest value in the prediction signal.
2. Its prominence is at least 15% of the absolute difference between the highest and lowest values of the prediction signal.
3. The heart rate estimated from the mean time difference between each consecutive ED frames does not exceed 200/min. The following formula was used to estimate heart rate:

$$heart\ rate = 60 \times \frac{frame\ rate}{mean\ number\ of\ frames\ between\ consecutive\ ED\ frames}$$

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 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 differences between heart rates estimated from the predicted ED/ES frames and the actual heart rate.

3.2 Self-supervised pre-training

To achieve high performance in predicting LVEF and RVEF, we first performed self-supervised pre-training using VideoMAE on a large unlabeled echocardiographic dataset before proceeding to the supervised training phase in a smaller labeled dataset. [14]Although such pre-training techniques have achieved outstanding results in various computer vision tasks,[23, 24] the conventional masking strategy of the VideoMAE pipeline may perform sub-optimally in the specific domain of echocardiographic videos. Motivated by this, we proposed a novel ROI-aware masking method that considers the unique characteristics of this domain.

3.2.1 Previous masking strategies

In VideoMAE[23], the input videos $\mathrm{I} \in \mathbb{R}^{\mathrm{T}\times\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ are transformed into a token sequence using cube embedding $S = \Phi(I)$. Then, a binary masking map $\mathrm{M_e}$ is generated for the sequence using a custom tube masking strategy to mask certain parts. The unmasked tokens $\boldsymbol{S^u} = \{S_i\}_{i\in\neg M_e}$ are used to generate a latent representation of the input $\mathrm{Z} = \Phi_{\mathrm{enc}}(\mathrm{S^u})$. This encoded representation is then combined with the learnable masking tokens A, and the decoder is tasked with reconstructing the masked tokens in the pixel space as $\hat{I} = \Phi_{dec}(Z, A)$.

VideoMAEv2[24] introduced the concept of dual masking. By applying a mask to the decoder as well, the computational cost could be significantly reduced during pre-training, while performance remained comparable to that of the single masking method. A new mask M_d is calculated (with $\mathrm{M_d} \subset \mathrm{M_e}$) which is then used to limit the number of learnable tokens supplied to the decoder $\Phi_{\mathrm{dec}}(\mathrm{Z}, \mathrm{A^m})$ where $A^m = \{A_i\}_{i\in M_d}$.

3.2.2 Challenges in analyzing echocardiographic videos

In a typical medical ultrasound video, only a fraction of pixels in each frame represent the actual image, while the remaining pixels display textual information and technical markers that, although useful for clinicians during the examination, are considered noise for DL algorithms.

By cropping each frame using a rectangular bounding box, some irrelevant pixels can be discarded. However, due to the sector shape of the ROI in echocardiographic videos, a significant portion of the remaining pixels still contains no valuable information.

3.2.3 Concept and technical details of ROI-aware masking

To overcome the challenge detailed above, we proposed a novel ROI-aware masking strategy that considers the peculiar shape and location of the ROI in the echocardiographic videos. Using the method described in subsection 3.1.1, a binary segmentation mask $I_s \in R^{HxW}$ could be created around the ROI.

As the ROI is static across all frames of the given video, we could stack the mask along the temporal dimension to map the same binary mask to each input frame.

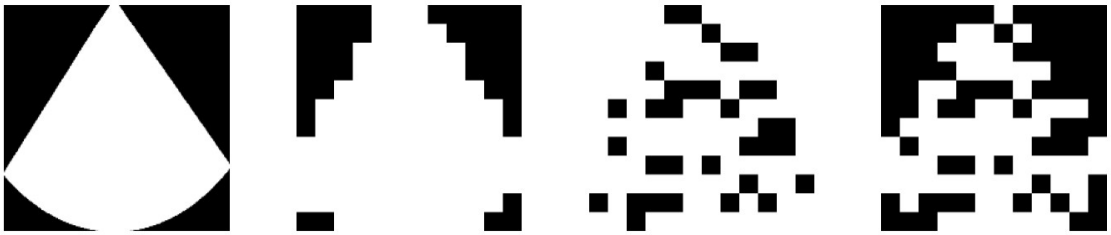
Cube embedding[14] was then applied to this stacked mask, resulting in a token masking map M_s that masked the embedded patches outside the segmentation mask. Using this new map, we defined the encoder ($\widehat{M_e}$) and decoder masks ($\widehat{M_d}$) as follows:

$$\widehat{M_e} = M_e \cup \neg M_s$$

$$\widehat{M_d} = M_e \cap M_s$$

where M_e was the tube mask defined for the original VideoMAE[23]. The encoder mask ensured that no unmasked tokens outside the ROI were used by the encoder, whereas the decoder mask ensured that the learnable tokens fed to the decoder were limited to the remaining unmasked parts of the ROI. Figure 1 shows how the encoder and decoder masking maps were generated using the binary segmentation mask of the ROI.

Using this novel masking strategy, the encoder was forced to learn the representation of the entire ROI, while the decoder was only tasked with reconstructing the relevant parts of the frames. Additionally, the ROI mask was also applied to the encoder during downstream training. This was a crucial step as the model had never been exposed to tokens outside the ROI in the pre-training phase, so supplying such tokens to the model during the supervised training phase could have potentially diminished its performance. The pre-training pipeline with the ROI-aware masking strategy is demonstrated in Figure 2.



(a) (b) (c) (d)

Figure 1 (a) The binary segmentation mask generated for each frame of an example echocardiographic video (I_s). (b) The same binary segmentation mask downsampled to the resolution of the cube embedding (M_s). (c) The encoder mask generated from the downsampled segmentation mask using the conventional tube masking method ($\widehat{M_e}$). (d) The decoder mask generated by intersecting the encoder mask with the downsampled segmentation mask ($\widehat{M_d}$).

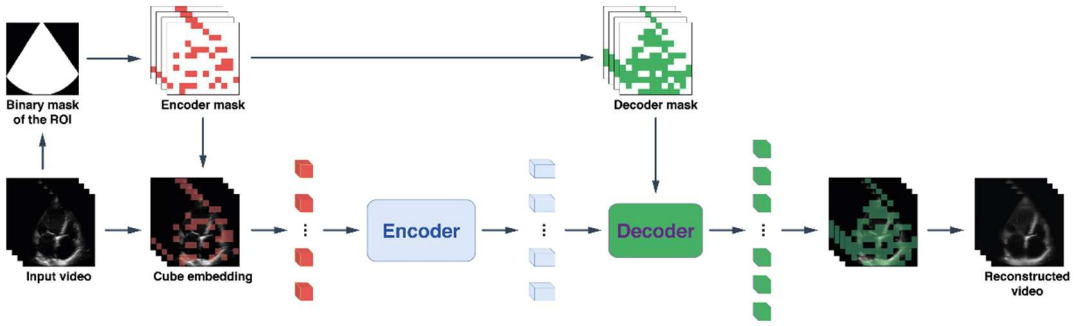


Figure 2 Schematic illustration of self-supervised pre-training with the novel ROI-aware masking strategy.

3.2.4 Evaluating the performance of ROI-aware masking

We evaluated the proposed ROI-aware masking strategy in comparison with simple tube masking and no pre-training, both quantitatively (by evaluating performance in downstream tasks) and qualitatively (by visually inspecting the reconstructed frames).

3.2.4.1 Pre-training using the novel ROI-aware masking strategy

We pre-trained a VideoMAE using a large dataset of unlabeled echocardiographic videos acquired at the Heart and Vascular Center of Semmelweis University between 2006 and 2023. From this dataset, 29,876 A4C videos were selected with the help of Model-V, corresponding to 15,661 transthoracic echocardiographic studies and encompassing 57,691 cardiac cycles in total. To validate the pre-training process, 450 videos were held out as a validation set. Due to the nature of the pre-training task, a larger validation set was not deemed necessary, as the primary focus was on training the model rather than directly testing its reconstruction performance. In our experiments, we used a ViViT-b 16x2 architecture,[14] which processed input videos of 16 gray-scale frames (192x192 pixels each) sampled from a single cardiac cycle with the help of Model-CC. During pre-training, we applied a 75% masking ratio using ROI-aware masking. The model was pre-trained for 100 epochs with a batch size of 2x30. The learning rate was set to 0.0005 initially and adjusted using a Cosine Annealing learning rate scheduler,[65] gradually reducing it to a minimum of 0.00001. To ensure temporal consistency and

improve generalization, we applied augmentations, including a random crop that retained at least 90% of the frames and random rotations with a maximum deviation of 15 degrees in either direction.

3.2.4.2 Evaluating the impact of ROI-aware masking on model performance in downstream tasks

In the subsequent supervised training phase, we replaced the decoder of the pre-trained model with either a regressor or a classification head (feed-forward neural networks, each containing one hidden layer with 512 neurons) and fine-tuned separate models on the smaller publicly available RVENet dataset[58] for two downstream tasks: (1) predicting LVEF, and (2) assigning a primary diagnosis (healthy, athlete, heart failure with reduced LVEF, aortic valve disease, or mitral valve disease) to each video. This dataset originally comprised 3,583 A4C videos (in DICOM file format) from 944 transthoracic echocardiographic studies of healthy volunteers and patients with a wide variety of cardiac diseases. Nevertheless, we only used the videos with 3DE-derived LVEF values available (n=3,523) and the videos of patients within one of the five abovementioned primary diagnosis categories (n=2,545) in the two downstream tasks, respectively. These subsets of videos were split into training, validation, and testing sets in an approximately 70:10:20 ratio at the study level. We also performed a series of experiments by progressively decreasing the sample size of the training set in 10% steps while leaving the validation and testing sets unchanged. The supervised training phase lasted for 40 epochs, with a batch size of 8. The learning rate was set to 0.0005 initially and adjusted using a Cosine Annealing learning rate scheduler,[65] gradually reducing it to a minimum of 0.00001. The same augmentation techniques applied during pre-training were also used in this phase. Layer normalization and dropout with a probability of 0.3 were applied to the regressor and classifier heads. The model's performance was evaluated using MAE, root mean squared error (RMSE), and the coefficient of determination (R^2) for the regression task (i.e., predicting LVEF) and accuracy for the classification task (i.e., predicting the primary diagnosis).

3.3 Supervised training for predicting ejection fractions

After performing pre-training as described in subsection *3.2.4.1*, we proceeded to the supervised training phase by discarding the decoder of the pre-trained model and attaching a feed-forward regressor network (containing one hidden layer with 512 neurons) to the encoder. Separate models were trained to predict LVEF and RVEF. Accordingly, the final QUEST-EF pipeline consisted of two encoders, each initialized with the pre-trained weights, and each connected to a regressor head.

3.3.1 Datasets used for training and evaluation

The LVEF prediction model of QUEST-EF was trained on the publicly available EchoNet-Dynamic dataset[55] comprising 10,030 A4C videos with 2DE-derived labels (only LVEF) and a dual-center 3DE dataset comprising 5,341 A4C videos with 3DE-derived labels (both LVEF and RVEF), whereas the RVEF prediction model was trained only on the latter dataset. For training, validation, and testing, the dual-center dataset was split in a 70:15:15 ratio at the patient level to avoid data leakage (i.e., all videos of a given patient were assigned to either the training, the validation, or the test set). The EchoNet-Dynamic dataset was only included in the training set of the LVEF prediction model.

Beyond testing QUEST-EF internally on 15% of the dual-center dataset (i.e., the internal test set), its performance was also evaluated in a labeled external test set, which included (1) 238 A4C videos of 238 patients with mixed cardiac diseases from an Italian center of whom 187 had available data regarding heart failure hospitalizations and all-cause mortality during follow-up, (2) 177 A4C videos of 90 adults with congenital heart disease from a British center, (3) 183 A4C videos (with LVEF labels only) of 183 patients with mixed cardiac diseases from a German center, (4) 20 A4C videos (with RVEF labels only) of 20 patients with severe tricuspid regurgitation from another German center, and (5) 4,695 A4C videos of 901 healthy adults enrolled in the WASE study.[66] Last, the associations between the predictions and 10-year all-cause mortality were also investigated in a Hungarian, low-risk, community-based cohort (1,166 unlabeled A4C videos of 1,166 individuals)(Figure 3).[63]

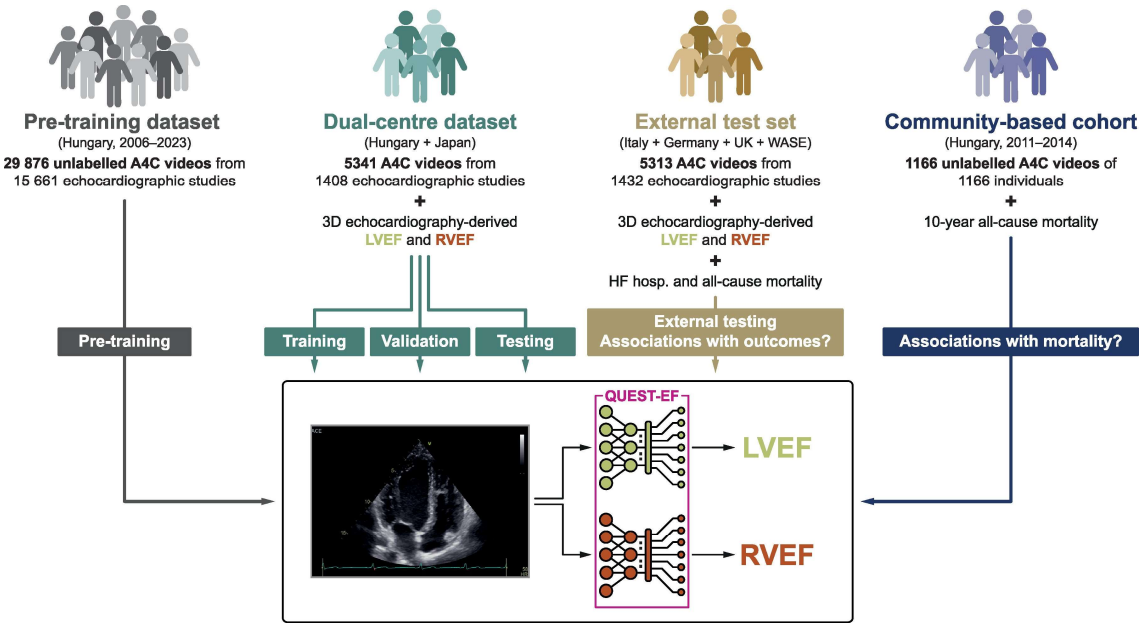


Figure 3 Training and testing datasets used for the development and testing of QUEST-EF.

A4C - apical four-chamber view; hosp. - hospitalization; LVEF - left ventricular ejection fraction; RVEF - right ventricular ejection fraction; WASE - World Alliance of Societies of Echocardiography.

3.3.2 Addressing the imbalanced distribution of ejection fraction values

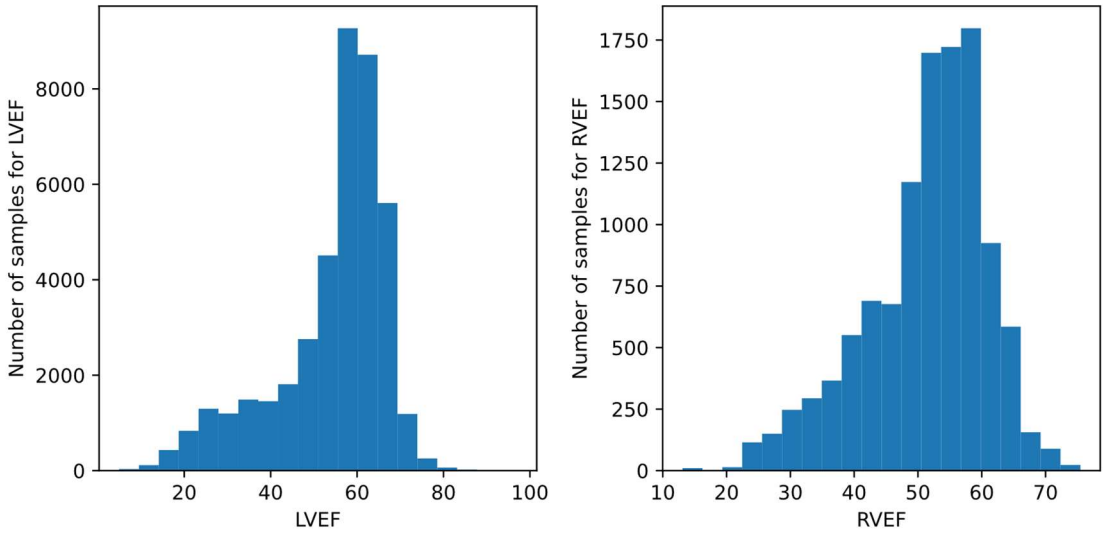


Figure 4 Histograms depicting the distributions of the ground truth LVEF and RVEF values in the dual-center dataset. The Y-axis represents the number of unique input samples (i.e., cardiac cycles) in each bin, and the X-axis shows the corresponding ejection fraction values.

LVEF – left ventricular ejection fraction, RVEF – right ventricular ejection fraction

Given the imbalanced distributions of the ground truth LVEF and RVEF values (Figure 4), we used the following weighted loss function during training to assign greater importance to underrepresented samples:

$$\mathrm{L} = \frac{1}{N}\sum_{i=1}^{N} w_Y(Y_i)\left(\widehat{Y}_i - Y_i\right)^2$$

where Y_i was the ground truth value of the i^{th} sample, $\widehat{Y}_i$ was the predicted value, and w_Y was a weight function based on the distribution of the true values Y in the training set. The weight function $w_Y(x)$ was defined as:

$$w_Y(x) = \frac{1 - \alpha f_Y(x)}{\frac{1}{N}\sum_{y_i \in Y}\left(1 - \alpha f_Y(y_i)\right)}$$

where f_Y represented the probability density function of Y, estimated using kernel density estimation (KDE)[67] with a Gaussian kernel, and α was a value between 0 and 1 determining the strength of the weighting. In our studies, we used an α value of 0.7 (Figure 5).

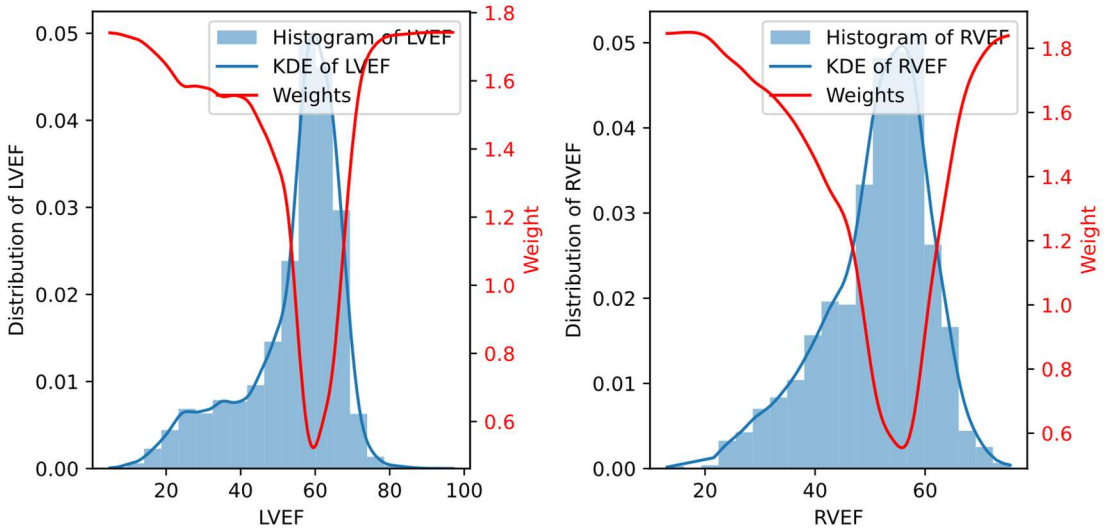


Figure 5 Density functions calculated by the kernel density estimator and the outputs of the weighting functions with $\alpha = 0.7$ in the dual-center dataset. The Y-axis represents the distribution of the input samples (i.e., cardiac cycles) in blue, as well as the calculated weight in red, and the X-axis shows the corresponding ejection fraction values.

3.3.3 Training parameters

The supervised training phase lasted for 40 epochs, with a batch size of 2x25. The learning rate of the regressor heads was set to 0.0005 initially and adjusted using a Cosine Annealing learning rate scheduler,[65] gradually reducing it to a minimum of 0.00001. For the encoders, we used a learning rate set to 10% of that of the regressor heads. The same augmentation techniques applied during pre-training (i.e., random cropping and rotation) were also used in this phase to further enhance the model's robustness. Layer normalization and dropout with a probability of 0.3 were applied to the regressor heads.

3.3.4 Evaluation metrics

QUEST-EF's performance in predicting the continuous LVEF and RVEF values was evaluated using the MAE, RMSE, and R^2. Blant-Altman analyses were performed to assess the bias and limits of agreement (LOA). Although QUEST-EF was trained to perform regression, we also evaluated its performance in secondary classification tasks: to identify patients with LV dysfunction (3DE-derived LVEF <50%) and RV dysfunction (3DE-derived RVEF <45%). In these tasks, the area under the receiver operating characteristic curve (AUC), accuracy, specificity, sensitivity, negative predictive value, and positive predictive value were used as performance metrics.

3.4 Explainability

Besides achieving prime performance in predicting LVEF and RVEF, we also sought to explore how QUEST-EF makes its final predictions using the tools of explainability.

Due to its transformer architecture, the performance of QUEST-EF could be assessed on subsets of cube-embedded patches corresponding to distinct cardiac structures in the video. Based on the change in performance compared with that obtained using the entire video, the relative importance of each structure could be determined: the smaller the drop in performance, the more important the structure. The following formula was used to calculate an importance score for each structure:

$$Importance\ score = 1 - \frac{MAE_S - MAE_O}{MAE_O}$$

where MAE_S is the MAE when only cube-embedded patches corresponding to the given cardiac structures were used, and MAE_O is the MAE when all cube-embedded patches (i.e., the entire video) were used. In our experiments, we utilized a previously published segmentation model[52] to automatically identify five cardiac structures – LV blood pool, LV myocardium (LV_{MYO}), left atrium (LA), RV, and right atrium (RA) – in the A4C videos (Figure 6). A cube-embedded patch was assigned to a segment (i.e., cardiac

structure) if at least 75% of its pixels overlapped with it.

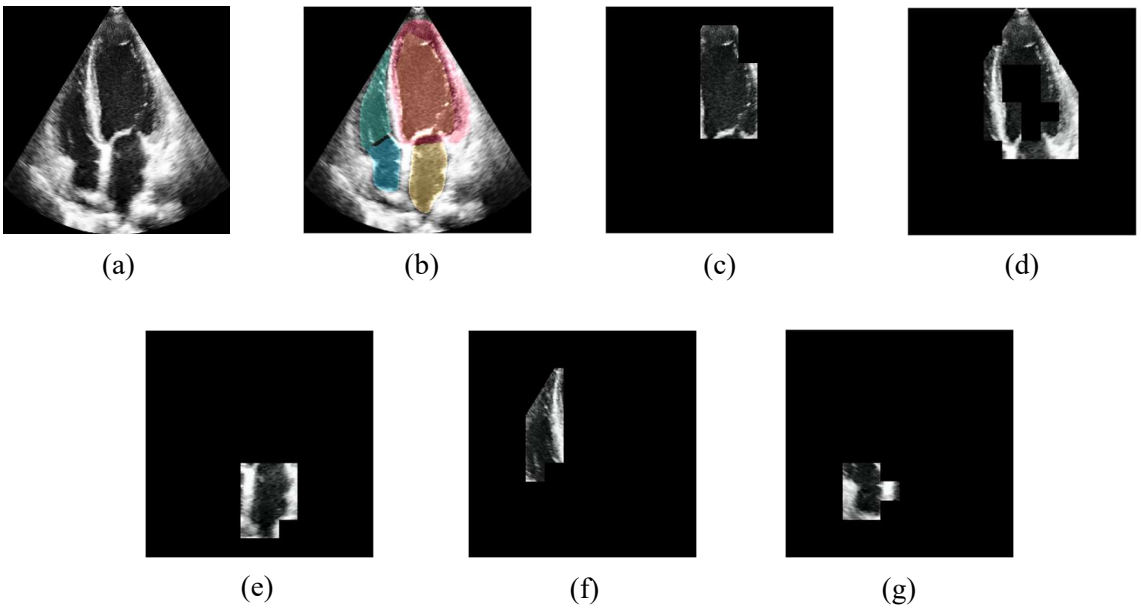


Figure 6 Segmentation results and the corresponding patches. (a) original frame, (b) original frame with overlayed segments, (c) left ventricular blood pool, (d) left ventricular myocardium, (e) left atrium, (f) right ventricle, (g) right atrium.

Additionally, to visualize the semantic reasoning of QUEST-EF, we used the attention-guided CAM method,[68] which leverages self-attention scores to highlight the parts of the frames the model focuses on when predicting LVEF and RVEF.

3.5 Statistical analysis

Statistical analyses were performed in R (version 4.1.2, R Foundation for Statistical Computing, Vienna, Austria). For each performance metric, 95% confidence intervals (CIs) were calculated from 10,000 stratified bootstrap resamples. Univariable and multivariable Cox proportional hazards models were used to compute hazard ratios (HRs) with 95% CIs. A p-value of <0.05 was considered statistically significant.

3.6 Software packages used for deep learning

Data preprocessing and DL algorithms were implemented in Python (version 3.9.16, Python Software Foundation, Wilmington, Delaware, U.S.A.) using the PyTorch (version 2.0.1) Lightning (version 2.1.3) libraries.

3.7 Code and model availability

To ensure transparency and reproducibility, the source code of QUEST-EF is publicly available at https://github.com/szadam96/quest-ef. Besides the scripts required for preprocessing, model training, and evaluation, all model weights are also accessible in this repository. The implementation of the ROI-aware masking strategy is provided separately at https://github.com/szadam96/ROI-aware-masking.

We recognized the importance of providing users with a ready-to-use tool. To achieve this, in addition to publishing the source code of QUEST-EF, we developed an intuitive web interface for the model to facilitate further testing and allow free use for research purposes. The website is available at http://quest-ef.com/.

3.8 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. Obtaining informed consent was waived due to the retrospective nature of the analyses.

4. Results

4.1 Performance of the models used in preprocessing

4.1.1 View classification – Model-V

In the three-class classification task, Model-V achieved balanced accuracies of 0.931 (95% CI: 0.915-0.947) and 0.835 (95% CI: 0.816-0.853) during internal and external testing (Table 2), whereas in binary classification, it disciminated A4C from non-A4C with balanced accuracies of 0.991 (95% CI: 0.987-0.994) and 0.895 (95% CI: 0.884-0.907) in the internal and external test sets, respectively (Table 3).

Table 2 Performance of Model-V in the three-class classification task

	Accuracy	Balanced accuracy	Precision	Recall	F1
Internal test set	0.966 (0.960-0.973)	0.931 (0.915-0.947)	0.909 (0.892-0.927)	0.931 (0.915-0.947)	0.920 (0.904-0.936)
External test set	0.934 (0.928-0.940)	0.835 (0.816-0.853)	0.879 (0.863-0.895)	0.835 (0.816-0.853)	0.854 (0.839-0.870)

Each performance metric is reported with a 95% confidence interval calculated from 10,000 stratified bootstrap resamples.

Table 3 Performance of Model-V in binary classification

	Accuracy	Balanced accuracy	Precision	Recall	F1
Internal test set	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	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)

Each performance metric is reported with a 95% confidence interval calculated from 10,000 stratified bootstrap resamples.

4.1.2 Orientation classification – Model-O

In classifying the orientation of the echocardiographic videos, Model-O achieved balanced accuracies of 0.995 (95% CI: 0.990-1.000) and 0.960 (95% CI: 0.948-0.971) in the internal and external test sets (Table 4).

Table 4 Performance of Model-O in classifying orientation

	Accuracy	Balanced accuracy	Precision	Recall	F1
Internal test set	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	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)

Each performance metric is reported with a 95% confidence interval calculated from 10,000 stratified bootstrap resamples.

4.1.3 Splitting videos into cardiac cycles – 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 5). More importantly, the model was unable to identify 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 5). 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 6). Seeing the more reliable identification of ES compared with ED frames, we defined cardiac cycles for QUEST-EF as starting and ending with the predicted ES frames.

Table 5 Performance of Model-CC in identifying end-diastolic and end-systolic frames

	MAE (frames)	RMSE (frames)	MAE (ms)	RMSE (ms)	Not identified
Internal test set					
End-diastole	4.041 (3.878-4.210)	5.253 (4.975-5.544)	76 (73-78)	96 (91-101)	7.277 % (124/1,704)
End-systole	2.799 (2.708-2.894)	3.414 (3.266-3.577)	53 (51-55)	63 (61-66)	1.232 % (21/1,704)
External test set					
End-diastole	3.117 (3.044-3.193)	3.992 (3.824-4.172)	65 (63-66)	78 (76-81)	4.232 % (183/4,324)
End-systole	4.827 (4.738-4.917)	5.663 (5.541-5.790)	97 (96-99)	107 (105-109)	3.632 % (153/4,213)

Each performance metric is reported with a 95% confidence interval calculated from 10,000 stratified bootstrap resamples.

MAE – mean absolute error, R^2 – coefficient of determination, RMSE – root mean squared error

Table 6 Performance of Model-CC in predicting heart rate

	MAE (bpm)	RMSE (bpm)	ICC	R^2
Internal test set				
End-diastole	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-systole	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-diastole	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-systole	1.757 (1.517-2.030)	4.803 (3.411-6.167)	0.926 (0.887-0.960)	0.839 (0.739-0.918)

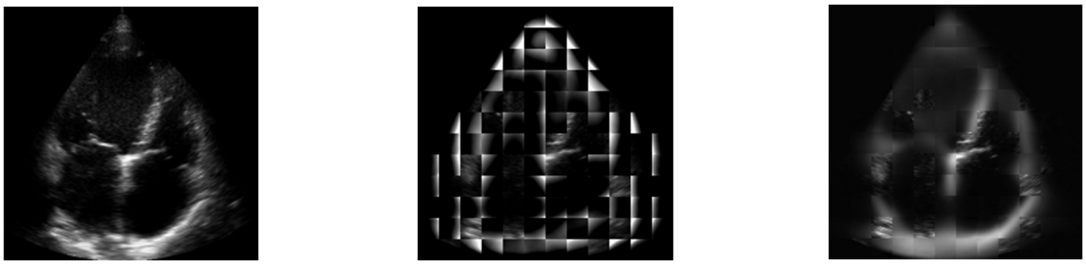
Each performance metric is reported with a 95% confidence interval calculated from 10,000 stratified bootstrap resamples.

bpm – beats per minute, ICC – intraclass correlation coefficient, MAE – mean absolute error, R^2 – coefficient of determination, RMSE – root mean squared error

4.2 ROI-aware masking

4.2.1 Improvement in video reconstruction

As shown in Figure 7, training with the proposed ROI-aware masking strategy enabled the model to capture the more granular structure of the heart seen in the original frame, whereas the model trained with tube masking focused on correctly identifying and reconstructing the border of the ROI and learned only the local black-to-white gradients. These findings confirmed that by allowing tokens consisting of irrelevant pixels to be fed into the encoder and reconstructed by the decoder, a considerable amount of resources is wasted on learning the representation of such tokens.



(a) (b) (c)

Figure 7 (a) The original frame of an example echocardiographic video, (b) the same frame reconstructed using the conventional tube masking method, and (c) the proposed ROI-aware masking method.

4.2.2 Impact on performance in downstream tasks

For LVEF prediction, the model without pre-training achieved an MAE of 6.212 (95% CI: 5.913-6.503) percentage points (Table 7). Pre-training with the traditional tube masking strategy provided only a marginal improvement (MAE: 5.933 [95% CI: 5.659-6.234]), whereas pre-training with the ROI-aware masking strategy outperformed both, yielding an MAE of 4.373 (95% CI: 4.196-4.570) (Table 7). For predicting the primary diagnosis, the tube masking strategy performed worse than the model with no pre-training (accuracy: 0.422 [95% CI: 0.392-0.451] vs. 0.581 [95% CI: 0.551-0.610]), while the ROI-aware masking strategy substantially improved performance, reaching an accuracy of 0.800 (95% CI: 0.776-0.824) (Table 2).

Table 7 Performance of the different pre-training strategies in the downstream tasks

	Predicting LVEF (regression)			**Predicting primary diagnosis** (classification)
	MAE	**RMSE**	**R^2**	**Accuracy**
No pre-training	6.212 (5.913-6.503)	8.490 (8.064-8.902)	0.408 (0.353-0.462)	0.581 (0.551-0.610)
Tube masking	5.933 (5.659-6.234)	8.223 (7.804-8.692)	0.444 (0.385-0.499)	0.422 (0.392-0.451)
ROI-aware masking	**4.373 (4.196-4.570)**	**5.863 (5.590-6.138)**	**0.717 (0.682-0.747)**	**0.800 (0.776-0.824)**

Each performance metric is reported with a 95% confidence interval calculated from 10,000 stratified bootstrap resamples.

LVEF – left ventricular ejection fraction, ROI – region-of-interest

We also analyzed the effect of progressively reducing the size of the labeled training set on model performance. Although performance declined across all three pre-training approaches, the model pre-trained with the novel ROI-aware masking strategy achieved the best results in both tasks, with an even more pronounced advantage over the other strategies as the training set size decreased (Figure 8).

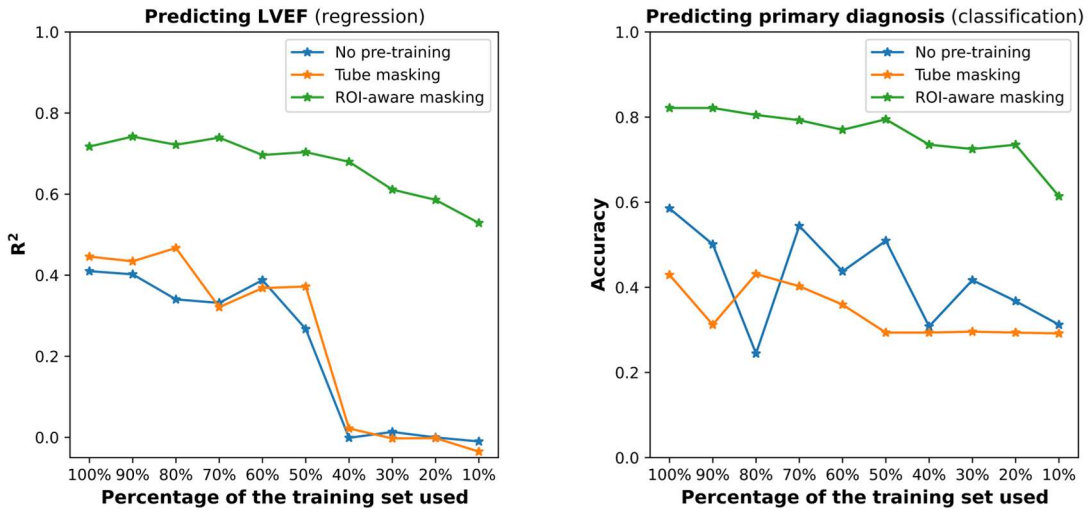


Figure 8 Effect of progressively reducing the sample size of the labeled training set on the model's performance.

LVEF – left ventricular ejection fraction, ROI – region-of-interest

4.3 Performance of QUEST-EF

4.3.1 Performance in predicting LVEF and RVEF

In predicting LVEF, QUEST-EF achieved study-level MAEs of 4.564 (95% CI: 4.105-5.087) and 4.601 (95% CI: 4.409-4.797) percentage points in the internal and external test sets, respectively (Table 8 and Figure 9). Bland-Altman analysis showed a study-level bias of 0.130 (95% CI: -0.648 to 0.908) percentage points with upper and lower LOAs of 11.739 (95% CI: 10.407-13.070) and -11.478 (95% CI: -12.810 to -10.146) percentage points in the internal, and a study-level bias of -1.471 (95% CI: -1.772 to -1.170) percentage points with upper and lower LOAs of 9.832 (95% CI: 9.317-10.347) and -12.774 (95% CI: -13.289 to -12.260) percentage points in the external test set (Figure 10).

Table 8 Performance of QUEST-EF in predicting left ventricular ejection fraction

	MAE (pp)	**RMSE** (pp)	**ICC**	**R^2**
Internal test set				
Video-level performance	4.875 (4.605-5.165)	6.368 (6.016-6.769)	0.888 (0.869-0.904)	0.795 (0.763-0.823)
Study-level performance	4.564 (4.105-5.087)	5.911 (5.342-6.671)	0.895 (0.861-0.921)	0.813 (0.758-0.855)
External test set				
Video-level performance	4.641 (4.537-4.746)	6.009 (5.880-6.150)	0.523 (0.488-0.558)	0.033 (-0.046 to 0.112)
Study-level performance	4.601 (4.409-4.797)	5.950 (5.707-6.214)	0.747 (0.714-0.779)	0.537 (0.477-0.592)

Each performance metric is reported with a 95% confidence interval calculated from 10,000 stratified bootstrap resamples.

ICC – intraclass correlation coefficient, MAE – mean absolute error, pp – percentage point, RMSE – root mean squared error

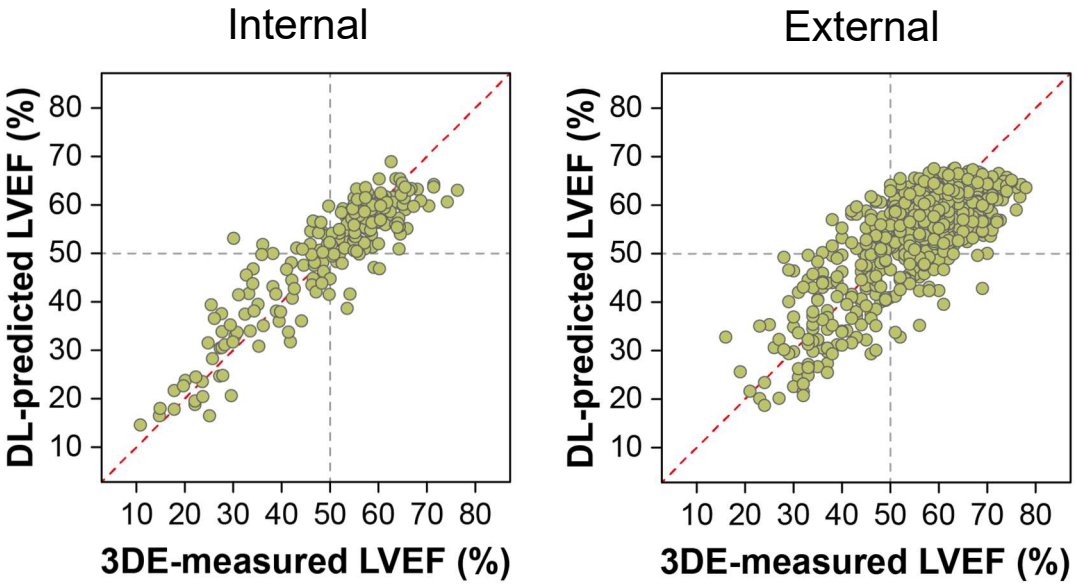


Figure 9 Scatter plots showing predicted vs. reference LVEF in the internal and external test sets.

DL – deep learning, LVEF – left ventricular ejection fraction

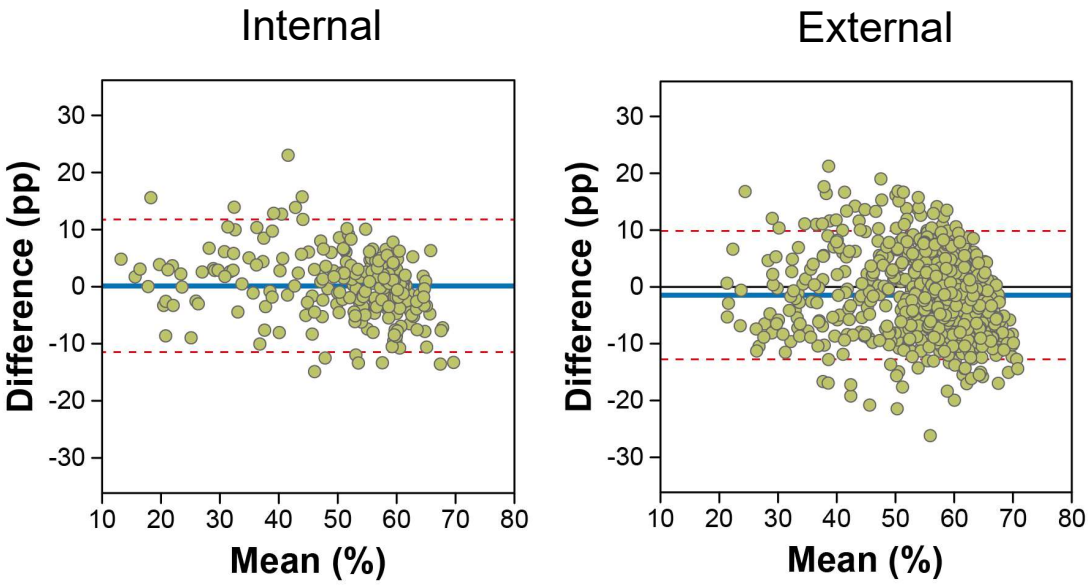


Figure 10 Bland-Altman plots showing the agreement between the predicted and reference LVEF in the internal and external test sets.

pp – percentage point

Table 9 Performance of QUEST-EF in predicting right ventricular ejection fraction

	MAE (pp)	**RMSE** (pp)	**ICC**	**R^2**
Internal test set				
Video-level performance	5.182 (4.902-5.479)	6.720 (6.368-7.138)	0.693 (0.653-0.727)	0.488 (0.426-0.543)
Study-level performance	4.815 (4.329-5.409)	6.265 (5.630-7.102)	0.709 (0.634-0.770)	0.534 (0.427-0.622)
External test set				
Video-level performance	5.579 (5.463-5.698)	7.033 (6.896-7.178)	0.169 (0.139-0.201)	-0.370 (-0.430 to -0.312)
Study-level performance	5.417 (5.183-5.653)	6.841 (6.568-7.138)	0.331 (0.270-0.392)	-0.055 (-0.153 to 0.041)

Each performance metric is reported with a 95% confidence interval calculated from 10,000 stratified bootstrap resamples.

ICC – intraclass correlation coefficient, MAE – mean absolute error, pp – percentage point, RMSE – root mean squared error

In predicting RVEF, QUEST-EF achieved MAEs of 4.815 (95% CI: 4.329-5.409) and 5.417 (95% CI: 5.183-5.653) on the study level in the internal and external test sets, respectively (Table 9 and Figure 11). Bland-Altman analysis showed a study-level bias of -0.211 (95% CI: -1.047 to 0.625) percentage points with upper and lower LOAs of 12.090 (95% CI: 10.660-13.521) and -12.513 (95% CI: -13.943 to -11.082) percentage points in the internal, and a study-level bias of 0.157 (95% CI: -0.223 to 0.537) percentage points with upper and lower LOAs of 13.568 (95% CI: 12.919-14.217) and -13.254 (95% CI: -13.903 to -12.605) percentage points in the external test set (Figure 12).

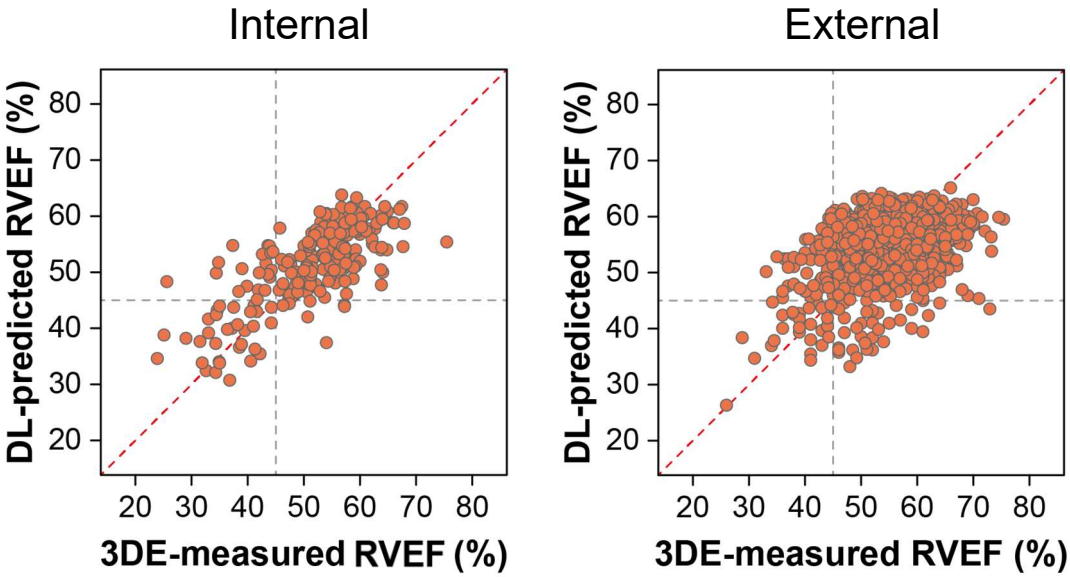


Figure 11 Scatter plots showing predicted vs. reference RVEF in the internal and external test sets.

DL – deep learning; RVEF – right ventricular ejection fraction

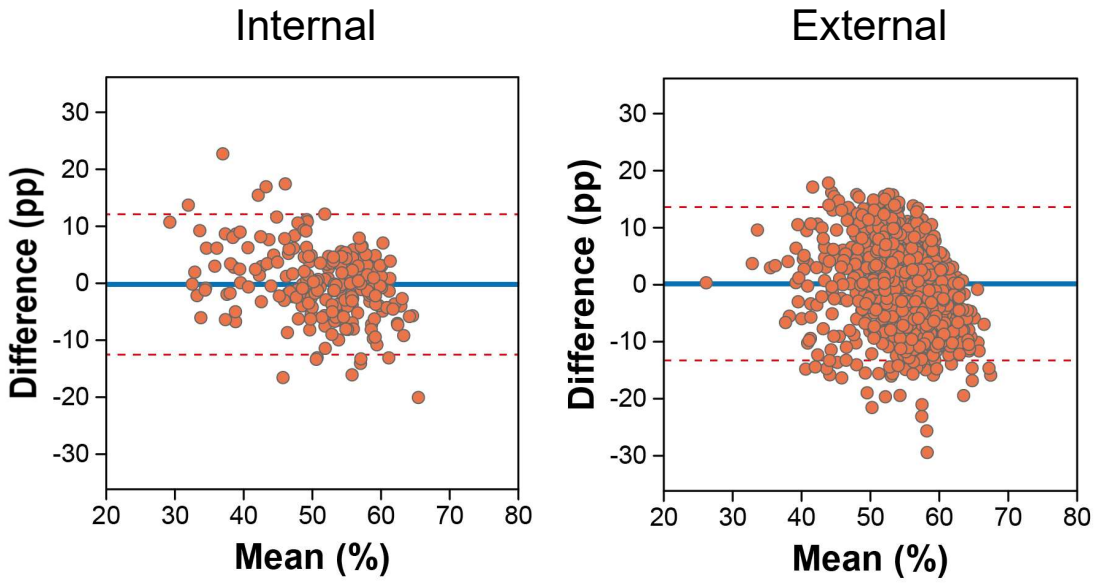


Figure 12 Bland-Altman plots showing the agreement between the predicted and reference RVEF in the internal and external test sets.

pp – percentage point

4.3.2 Performance in predicting LV and RV dysfunction

In the internal test set, QUEST-EF identified LV and RV systolic dysfunction (i.e. LVEF <50% and RVEF <45%) with AUCs of 0.944 (95% CI: 0.911-0.969) and 0.879 (95% CI: 0.821-0.928), respectively (Table 10 and Figure 13), whereas, in the labeled external test set, it achieved AUCs of 0.940 (95% CI: 0.919-0.958) and 0.791 (95% CI: 0.736-0.841) in these tasks (Table 11 and Figure 14).

Table 10 Performance of QUEST-EF in predicting left ventricular dysfunction (left ventricular ejection fraction <50%)

	AUC	Accuracy	Specificity	Sensitvity	NPV	PPV
Internal test set						
Video-level	0.941 (0.925-0.955)	0.873 (0.850- 0.895)	0.914 (0.890-0.938)	0.811 (0.767-0.851)	0.882 (0.859-0.906)	0.859 (0.824-0.894)
Study-level	0.944 (0.911-0.969)	0.871 (0.827- 0.911)	0.921 (0.871-0.964)	0.788 (0.694-0.871)	0.878 (0.832-0.922)	0.861 (0.789-0.929)
External test set						
Video-level	0.908 (0.887-0.928)	0.940 (0.933- 0.946)	0.953 (0.947-0.958)	0.627 (0.561-0.693)	0.984 (0.981-0.987)	0.357 (0.320-0.394)
Study-level	0.940 (0.919-0.958)	0.930 (0.917- 0.943)	0.958 (0.947-0.968)	0.730 (0.661-0.793)	0.962 (0.953-0.971)	0.710 (0.652-0.769)

Accuracy, specificity, sensitivity, NPV, and PPV were calculated for the cutoff value of 50%. 95% confidence intervals were calculated from 10,000 stratified bootstrap resamples.

AUC – area under the receiver operating characteristic curve, NPV – negative predictive value, PPV – positive predictive value

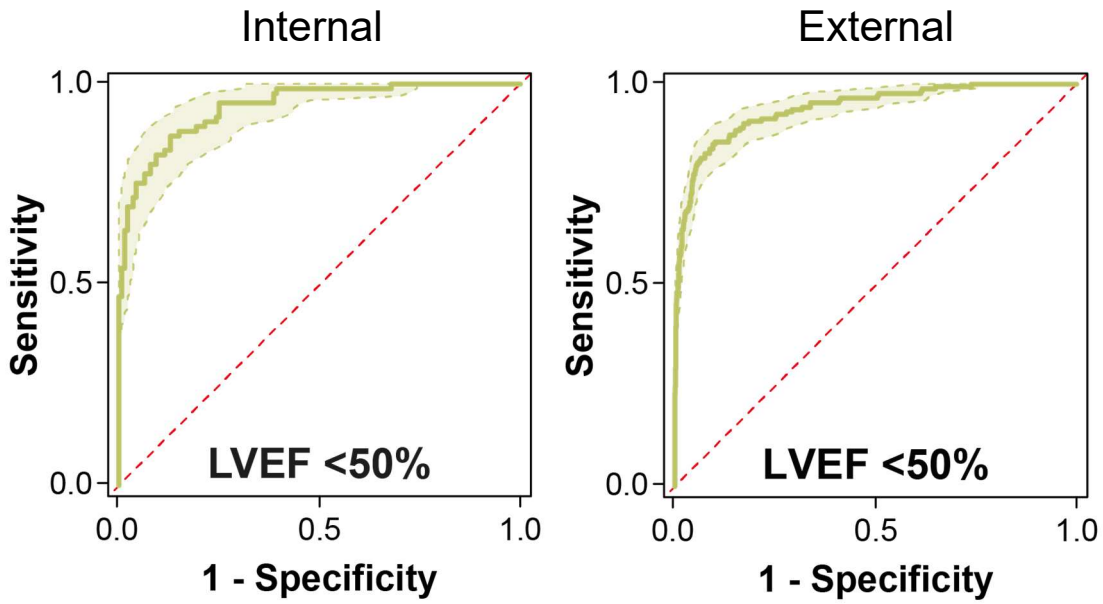


Figure 13 Receiver operating characteristic curves for predicting LV dysfunction
LV – left ventricular, LVEF – left ventricular ejection fraction

Table 11 Performance of QUEST-EF in predicting right ventricular dysfunction (right ventricular ejection fraction <45%)

	AUC	Accuracy	Specificity	Sensitvity	NPV	PPV
Internal test set						
Video-level	0.870 (0.839-0.898)	0.864 (0.842-0.885)	0.936 (0.917-0.954)	0.633 (0.566-0.699)	0.891 (0.873-0.909)	0.756 (0.699-0.814)
Study-level	0.879 (0.821-0.928)	0.877 (0.836-0.913)	0.958 (0.928-0.988)	0.615 (0.481-0.750)	0.890 (0.856-0.924)	0.825 (0.708-0.935)
External test set						
Video-level	0.689 (0.649-0.728)	0.939 (0.934-0.944)	0.972 (0.967-0.976)	0.167 (0.119-0.219)	0.965 (0.963-0.967)	0.202 (0.148-0.260)
Study-level	0.791 (0.736-0.841)	0.924 (0.913-0.935)	0.971 (0.960-0.980)	0.299 (0.207-0.402)	0.949 (0.942-0.956)	0.443 (0.323-0.550)

Accuracy, specificity, sensitivity, NPV, and PPV were calculated for the cutoff value of 45%. 95% confidence intervals were calculated from 10,000 stratified bootstrap resamples.

AUC – area under the receiver operating characteristic curve, NPV – negative predictive value, PPV – positive predictive value

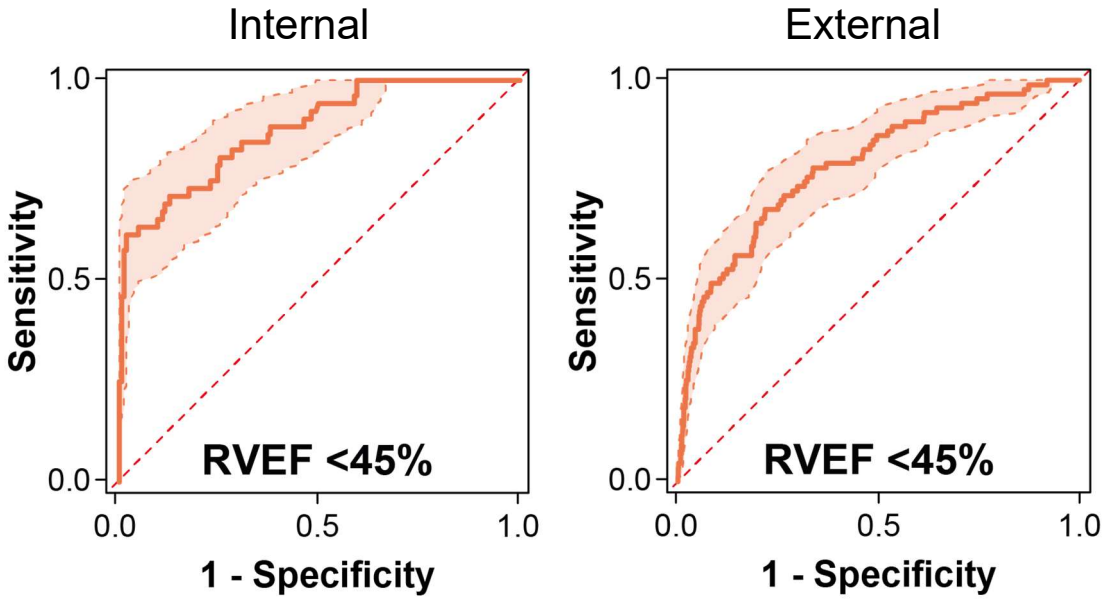


Figure 14 Receiver operating characteristic curves for predicting RV dysfunction
RV – right ventricular, RVEF – right ventricular ejection fraction

4.3.3 Associations with outcomes

Among the 187 patients with available outcome data [28 (15.0%) died or were hospitalized due to heart failure during the median follow-up duration of 1.1 (interquartile range: 0.5-1.6) years], the QUEST-EF-predicted EF values were associated with the composite endpoint of heart failure hospitalization or all-cause death [LVEF – adjusted HR (aHR): 0.945 (95% CI: 0.913-0.979), p=0.002; RVEF – aHR: 0.927 (95% CI: 0.877-0.979), p=0.006], independent of age and sex (Table 12). In the community-based cohort [10-year all-cause mortality rate: 131/1,166 (11.2%)], the predictions were also associated with 10-year all-cause death [LVEF – aHR: 0.947 (95% CI: 0.924-0.970), p<0.001; RVEF – aHR, 0.877 (95% CI: 0.845-0.909), p<0.001], independent of the Framingham Risk Score[69] and LV filling pressures estimated by E/e′ ratio (Table 13).

Table 12 Associations of the QUEST-EF-predicted left and right ventricular ejection fraction values with the composite endpoint of heart failure hospitalizations or all-cause death among patients with available outcome data in the labeled external test set

	Univariable Cox regression		**Multivariable Cox regression**			
			Model 1		Model 2	
	HR (95% CI)	P-value	HR (95% CI)	P-value	HR (95% CI)	P-value
Age	1.057 (1.015-1.102)	0.008	1.050 (1.008-1.093)	0.018	1.045 (1.002-1.089)	0.038
Male sex	0.495 (0.234-1.046)	0.066	0.397 (0.179-0.879)	0.023	0.502 (0.232-1.082)	0.079
Predicted LVEF	0.958 (0.929-0.988)	0.006	0.945 (0.913-0.979)	0.002		
Predicted RVEF	0.929 (0.886-0.975)	0.003			0.927 (0.877-0.979)	0.006

CI – confidence interval, HR – hazard ratio, LVEF – left ventricular ejection fraction, RVEF – right ventricular ejection fraction

Table 13 Associations of the QUEST-EF-predicted left and right ventricular ejection fraction values with 10-year all-cause death in the community-based cohort

	Univariable Cox regression		**Multivariable Cox regression**			
			Model 1		Model 2	
	HR (95% CI)	P-value	HR (95% CI)	P-value	HR (95% CI)	P-value
Framingham RS	1.036 (1.028-1.043)	<0.001	1.031 (1.022-1.039)	<0.001	1.027 (1.018-1.036)	<0.001
Average E/e’	1.243 (1.182-1.306)	<0.001	1.130 (1.040-1.227)	0.004	1.078 (0.987-1.176)	0.094
Predicted LVEF	0.931 (0.911-0.952)	<0.001	0.947 (0.924-0.970)	<0.001		
Predicted RVEF	0.843 (0.817-0.870)	<0.001			0.877 (0.845-0.909)	<0.001

CI – confidence interval, E – early mitral inflow velocity, e’ – early diastolic mitral annular velocity, HF – heart failure, HR – hazard ratio, LVEF – left ventricular ejection fraction, RS – risk score, RVEF – right ventricular ejection fraction

4.3.4 Explainability

Based on our custom segmentation-based method, the LV myocardium and blood pool were the most important segments for predicting LVEF, whereas the remaining three segments (LA, RV, and RA) contributed marginally (Table 14). For RVEF prediction, the RA was the most important segment, followed by the LV myocardium, RV, LV blood pool, and LA in decreasing order of importance; however, the differences in importance scores across segments were less pronounced than in the LVEF prediction task (Table 15).

Table 14 Performance of QUEST-EF in predicting left ventricular ejection fraction using subsets of cube-embedded patches corresponding to different cardiac structures

	MAE (pp)	**RMSE** (pp)	**R^2**	**Importance score**
Entire video	4.875 (4.605-5.165)	6.368 (6.016-6.769)	0.795 (0.763-0.823)	
LA	9.964 (9.304-10.621)	13.793 (12.886-10.621)	0.008 (-0.064 to 0.078)	-0.125 (-0.302 to 0.045)
LV	8.397 (7.860-8.943)	11.294 (10.560-12.044)	0.334 (0.252-0.411)	0.210 (0.074-0.337)
LV_{MYO}	8.028 (7.503-8.576)	10.882 (10.118-11.678)	0.382 (0.316-0.457)	0.290 (0.163-0.413)
RA	10.339 (9.643-11.047)	14.336 (13.453-15.241)	-0.071 (-0.141 to -0.002)	-0.204 (-0.385 to -0.030)
RV	9.989 (9.308-10.685)	13.883 (13.021-14.718)	-0.004 (-0.082 to 0.072)	-0.128 (-0.301 to 0.036)

95% confidence intervals were calculated from 10,000 stratified bootstrap resamples.
LA – left atrium, LV – left ventricle, LV_{MYO} – left ventricular myocardium, RA – right atrium, RV – right ventricle

Table 15 Performance of QUEST-EF in predicting right ventricular ejection fraction using subsets of cube-embedded patches corresponding to different cardiac structures

	MAE (pp)	**RMSE** (pp)	**R^2**	**Importance score**
Entire video	5.182 (4.902-5.479)	6.720 (6.368-7.138)	0.488 (0.426-0.543)	
LA	7.718 (7.262-8.183)	10.041 (9.455-10.632)	-0.151 (-0.261 to -0.048)	0.470 (0.367-0.567)
LV	7.700 (7.243-8.163)	10.023 (9.423-10.651)	-0.148 (-0.284 to -0.030)	0.473 (0.366-0.573)
LV_{MYO}	7.169 (6.761-7.594)	9.265 (8.752-9.787)	0.019 (-0.057 to 0.091)	0.579 (0.491-0.663)
RA	6.922 (6.511-7.344)	9.123 (8.551-9.703)	0.050 (-0.049 to 0.140)	0.628 (0.534-0.714)
RV	7.479 (7.005-7.974)	10.165 (9.390-11.011)	-0.183 (-0.385 to -0.022)	0.517 (0.411-0.619)

95% confidence intervals were calculated from 10,000 stratified bootstrap resamples.
LA – left atrium, LV – left ventricle, LV_{MYO} – left ventricular myocardium, RA – right atrium, RV – right ventricle

To qualitatively assess the explainability of QUEST-EF, we visualized self-attention maps for both LVEF and RVEF predictions using the attention-guided CAM[68] method on a representative sample from the internal test set. Despite the inherent subjectivity of their visual interpretation, the self-attention maps appeared consistent with the quantitative findings: for LVEF prediction, QUEST-EF focused primarily on the LV myocardium, whereas for RVEF prediction, its attention was more diffuse and not confined to a single segment (Figure 15).

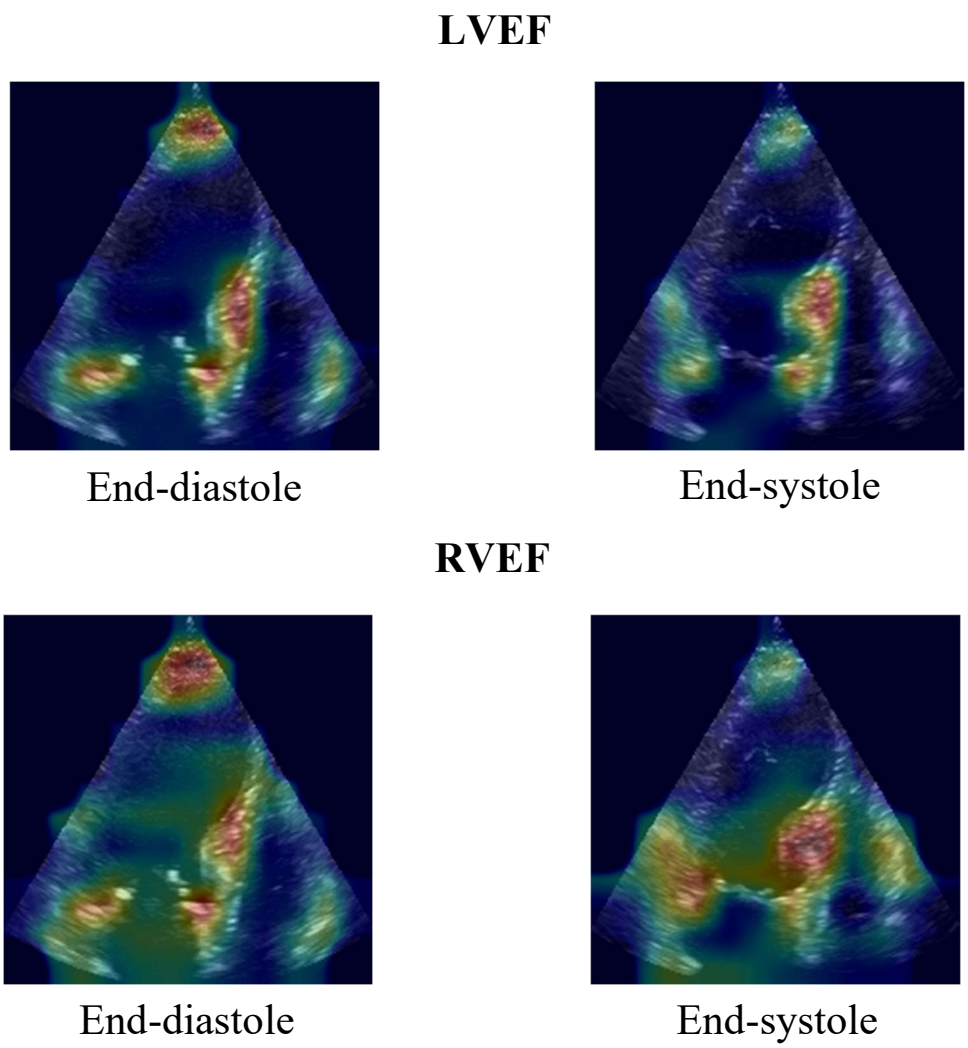


Figure 15 Heatmaps illustrating QUEST-EF's attention on distinct cardiac structures in end-systolic and end-diastolic frames of an example video.
LVEF – left ventricular ejection fraction, RVEF – right ventricular ejection fraction

5. Discussion

5.1 ROI-aware masking

Transformers have rapidly overtaken the computer vision field as the state-of-the-art architecture, owing to their exceptional performance and scalability compared with previous models. This methodological shift, however, has been less pronounced in medical image analysis, as transformers generally require large amounts of data for effective training. SSL approaches, such as masked autoencoders, enable these models to leverage vast collections of unlabeled data. Yet, as we demonstrated, masking strategies developed for natural images do not translate well to the domain of echocardiographic videos.

To address this limitation, we introduced a novel ROI-aware masking strategy designed explicitly for VideoMAEs. By focusing on ROIs and ignoring irrelevant areas of the frames, this approach markedly improved pre-training efficiency and led to superior performance in both image reconstruction and downstream tasks compared with conventional tube masking and no pre-training. Notably, the advantage of the ROI-aware masking strategy became even more pronounced as the amount of labeled training data decreased, underscoring its effectiveness in data-limited settings.

Although we tested the proposed ROI-aware masking strategy in the context of echocardiography, it has strong potential for generalization to other medical and non-medical domains where a substantial portion of pixels in the input frames do not carry relevant information with respect to the DL task.

5.2 QUEST-EF

LVEF and RVEF are the most important non-invasive surrogates of systolic cardiac function, serving as strong determinants of symptoms, functional capacity, quality of life, and clinical outcomes. To support clinicians in the echocardiographic assessment of these parameters, we developed QUEST-EF, a vendor-independent and segmentation-free DL-based solution capable of accurately predicting biventricular EFs using a single,

routinely acquired A4C video. While designing QUEST-EF, we intentionally chose a segmentation-free approach to enable QUEST-EF to extract both direct and indirect indicators of ventricular function from all visualized anatomical structures, rather than restricting it to outlining and tracking the endocardial border of the ventricles. This approach is particularly relevant for predicting RVEF, as contouring the RV in a single plane would only yield 2D RV fractional area change that has shown significant disagreements with 3DE-derived RVEF.[70] Moreover, not relying on accurate endocardial border tracking ensures that the model's performance is less likely to degrade if some LV or RV myocardial segments are visualized poorly or fall outside the imaging sector.

Another key strength of QUEST-EF is its end-to-end design, which enables the rapid and simultaneous prediction of LVEF and RVEF from a single echocardiographic video without requiring manual intervention. QUEST-EF has a rigorous multi-stage preprocessing module that not only algorithmically cleans and crops the frames but also incorporates three thoroughly validated DL models to verify the view and orientation of the input videos and to split them into individual cardiac cycles. The use of transformers instead of spatiotemporal CNNs represents another technical advancement over our previously published single-task model,[57] as well as the employment of a state-of-the-art pre-training technique that leveraged a large, unlabeled dataset to enhance performance in the downstream task of predicting LVEF and RVEF.[71] We believe these technical innovations collectively contributed to the robust performance of QUEST-EF, which we observed across a wide range of acquired and congenital cardiac diseases at centers spanning six continents. Although we should be very cautious when comparing performance metrics calculated in different datasets, the performance of QUEST-EF falls within a similar range to that reported for other recently published DL models, further supporting its robustness.[36,59,73]

In addition to its diagnostic value, the prognostic potential of QUEST-EF was also demonstrated, as lower predicted LVEF and RVEF values were independently associated with a significantly higher risk of adverse outcomes in two separate cohorts. These findings highlight the ability of QUEST-EF to facilitate automated risk stratification and early identification of patients at increased risk.

Beyond excessively evaluating the performance of QUEST-EF, we also investigated the reasoning behind its predictions by quantifying how different regions of the input contributed to the outputs. As expected, the LV blood pool and myocardium provided the most information for LVEF prediction, as they achieved the highest importance scores. For RVEF prediction, RA was found to be the most important segment; however, the differences in importance scores across segments were less pronounced than for LVEF. This phenomenon is likely attributable to the greater complexity of the RV's geometry and contraction pattern, which makes RVEF prediction from a single echocardiographic view more challenging and forces the model to rely on contextual cues and latent information from other cardiac structures.

We foresee that QUEST-EF could be particularly valuable in clinical scenarios where 3D imaging is not feasible or available – such as point-of-care ultrasound examinations performed by non-cardiologist users, including internal medicine specialists, pulmonologists, cardiac surgeons, intensivists, and emergency physicians – by enabling fast, automated, and accurate screening for LV and RV dysfunction.

5.3 Limitations

Despite its robustness, QUEST-EF has a few limitations that should be acknowledged. First, it is currently intended for research use only and has not been approved for clinical application. Therefore, regulatory approval and further rigorous testing are required before it can be integrated into clinical decision-making. Second, QUEST-EF has higher prediction errors and a larger generalization gap for RVEF than for LVEF, most likely due to the RV's more complex geometry and contraction pattern, which make single-view assessment of its function more challenging than that of the LV.[46, 47, 72] However, in a previous work with a single-task model,[57] it was demonstrated that a similarly segmentation-free approach could still achieve higher sensitivity than expert human readers, implying that it may be particularly well suited for screening purposes by non-expert physicians. Last, we may assume that a model analyzing multiple views would achieve even better performance than our single-view model, particularly for RVEF prediction. Nevertheless, we deliberately opted for this more simplistic

approach, which requires only a single routinely acquired echocardiographic view, to facilitate QUEST-EF's future clinical adoption and integration into handheld ultrasound devices and ensure ease of use, even for physicians with limited expertise in echocardiography.

6. Conclusions

1. Based on our first study, in which we proposed a novel ROI-aware masking strategy that accounts for the characteristic sector-shaped ROI in echocardiographic videos during SSL, we drew the following conclusions:
   1.1 Pre-training with ROI-aware masking yielded superior performance compared with conventional tube masking and no pre-training in self-supervised image reconstruction and two downstream tasks, namely estimating LVEF and predicting primary diagnosis from an A4C echocardiographic video.
   1.2 The advantage of the ROI-aware masking strategy was even more pronounced as the amount of labeled training data decreased, underscoring its effectiveness in data-limited settings.
2. In our second study, we developed and validated QUEST-EF, a dual-task DL pipeline trained with SSL incorporating our novel ROI-aware masking strategy for automated quantification of 3DE-derived LVEF and RVEF from A4C echocardiographic videos. The key findings of this study can be summarized as follows:
   2.1 QUEST-EF was designed as a fully automated, segmentation-free end-to-end DL pipeline featuring a multi-stage preprocessing module that algorithmically cleans and crops frames and integrates three validated DL models for view and orientation verification and splitting videos into cardiac cycles.
   2.2 QUEST-EF exhibited robust performance in predicting LVEF and RVEF and accurately detected LV and RV dysfunction during both internal and external validation across diverse patient cohorts.
   2.3 The prognostic value of QUEST-EF was also confirmed, as lower predicted LVEF and RVEF values were independently associated with a significantly higher risk of adverse outcomes in two separate cohorts.
   2.4 Explainability analysis revealed that the LV myocardium and blood pool were the most important segments for predicting LVEF. In contrast, the RA contributed the most to RVEF prediction, and less pronounced differences were observed in importance scores across cardiac structures than in the LVEF task.

7. Summary

Accurate assessment of LV and RV function is pivotal in cardiac imaging. To support clinicians in this routine task, we developed QUEST-EF, a vendor-independent and segmentation-free DL tool capable of fully automated prediction of LVEF and RVEF from single-view echocardiographic videos.

QUEST-EF was implemented as a comprehensive end-to-end DL pipeline comprising a multi-stage preprocessing module and an EF prediction module with two video vision transformers. The transformers were first pre-trained in a self-supervised manner on a large set of unlabeled A4C videos, using a novel ROI-aware masking strategy, designed to improve the suboptimal pre-training performance of conventional tube masking. By ignoring irrelevant areas of the frames, this approach substantially improved pre-training efficiency and enhanced performance in both image reconstruction and downstream tasks. In the subsequent supervised learning phase, one of the transformers was fine-tuned for LVEF prediction using the publicly available EchoNet-Dynamic dataset and a dual-center echocardiographic dataset, while the other was trained for RVEF prediction only on the latter. QUEST-EF was externally validated in patients with acquired or congenital cardiac diseases from four international centers and healthy adults from six continents enrolled in the WASE study. Associations between QUEST-EF-predicted EF values and 10-year all-cause mortality were also analyzed in a community-based cohort.

During internal and external validation, QUEST-EF exhibited robust performance in predicting LVEF and RVEF and accurately detected LV and RV dysfunction. Among patients with available outcome data, the predicted EF values were associated with the composite endpoint of heart failure hospitalization or all-cause death. In the community-based cohort, the predictions were also associated with 10-year all-cause mortality, independent of the Framingham Risk Score and LV diastolic function.

In summary, we successfully developed QUEST-EF, a dual-task DL model that enables rapid, automated, and accurate assessment of biventricular EFs from A4C echocardiographic videos, allowing efficient screening for LV and RV dysfunction.

8. References

1. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25:
2. Dai Y, Gao Y, Liu F (2021) TransMed: Transformers Advance Multi-Modal Medical Image Classification. Diagnostics 11:1384
3. He K, Gan C, Li Z, Rekik I, Yin Z, Ji W, Gao Y, Wang Q, Zhang J, Shen D (2023) Transformers in medical image analysis. Intelligent Medicine 3:59–78
4. Lin T, Wang Y, Liu X, Qiu X (2022) A survey of transformers. AI open 3:111–132
5. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv. Neural Inf. Process. Syst. 30:
6. Graves A, Mohamed A, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. pp 6645–6649
7. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780
8. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
9. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 10684–10695
10. Dosovitskiy A, Beyer L, Kolesnikov A, et al (2021) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR
11. Liu Z, Hu H, Lin Y, et al (2022) Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 12009–12019
12. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In:

Proceedings of the IEEE/CVF international conference on computer vision. pp 10012–10022

13. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, Hu H (2022) Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 3202–3211
14. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C (2021) Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp 6836–6846
15. Azad R, Kazerouni A, Heidari M, Aghdam EK, Molaei A, Jia Y, Jose A, Roy R, Merhof D (2024) Advances in medical image analysis with vision transformers: a comprehensive review. Med Image Anal 91:103000
16. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
17. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp 1597–1607
18. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 9729–9738
19. Caron M, Touvron H, Misra I, Jégou H, Mairal J, Bojanowski P, Joulin A (2021) Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp 9650–9660
20. Grill J-B, Strub F, Altché F, et al (2020) Bootstrap your own latent-a new approach to self-supervised learning. Adv Neural Inf Process Syst 33:21271–21284
21. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R (2022) Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 16000–16009
22. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA (2016) Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 2536–2544

23. Tong Z, Song Y, Wang J, Wang L (2022) VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Adv Neural Inf Process Syst 35:10078–10093
24. Wang L, Huang B, Zhao Z, Tong Z, He Y, Wang Y, Wang Y, Qiao Y (2023) Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp 14549–14560
25. Laçi H, Sevrani K, Iqbal S (2025) Deep learning approaches for classification tasks in medical X-ray, MRI, and ultrasound images: a scoping review. BMC Med Imaging 25:156
26. Fallahpoor M, Chakraborty S, Pradhan B, Faust O, Barua PD, Chegeni H, Acharya R (2024) Deep learning techniques in PET/CT imaging: A comprehensive review from sinogram to image space. Comput Methods Programs Biomed 243:107880
27. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, Ashrafian H, Darzi A (2021) Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 4:65
28. Mazurowski MA, Buda M, Saha A, Bashir MR (2019) Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. Journal of Magnetic Resonance Imaging 49:939–954
29. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL (2018) Artificial intelligence in radiology. Nat Rev Cancer 18:500–510
30. Sermesant M, Delingette H, Cochet H, Jaïs P, Ayache N (2021) Applications of artificial intelligence in cardiovascular imaging. Nat Rev Cardiol 18:600–609
31. Azad R, Kazerouni A, Heidari M, Aghdam EK, Molaei A, Jia Y, Jose A, Roy R, Merhof D (2024) Advances in medical image analysis with vision Transformers: A comprehensive review. Med Image Anal 91:103000
32. Imagawa K, Shiomoto K (2024) Evaluation of effectiveness of pre-training method in chest X-ray imaging using vision transformer. Comput Methods Biomech Biomed Eng Imaging Vis. https://doi.org/10.1080/21681163.2024.2345823

33. Shurrab S, Duwairi R (2022) Self-supervised learning methods and applications in medical imaging analysis: a survey. PeerJ Comput Sci 8:e1045
34. Yan R, Qu L, Wei Q, Huang S-C, Shen L, Rubin DL, Xing L, Zhou Y (2023) Label-Efficient Self-Supervised Federated Learning for Tackling Data Heterogeneity in Medical Imaging. IEEE Trans Med Imaging 42:1932–1943
35. Holste G, Oikonomou EK, Tokodi M, Kovács A, Wang Z, Khera R (2024) PanEcho: Complete AI-enabled echocardiography interpretation with multi-task deep learning. medRxiv 2011–2024
36. Reynaud H, Vlontzos A, Hou B, Beqiri A, Leeson P, Kainz B (2021) Ultrasound video transformers for cardiac ejection fraction estimation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24. pp 495–505
37. Czempiel T, Paschali M, Ostler D, Kim ST, Busam B, Navab N (2021) OperA: Attention-Regularized Transformers for Surgical Phase Recognition. pp 604–614
38. Vukadinovic M, Tang X, Yuan N, Cheng P, Li D, Cheng S, He B, Ouyang D (2024) EchoPrime: A Multi-Video View-Informed Vision-Language Model for Comprehensive Echocardiography Interpretation.
39. Gillam LD, Marcoff L (2024) Echocardiography: Past, Present, and Future. Circ Cardiovasc Imaging. https://doi.org/10.1161/CIRCIMAGING.124.016517
40. Vahanian A, Beyersdorf F, Praz F, et al (2022) 2021 ESC/EACTS Guidelines for the management of valvular heart disease. Eur Heart J 43:561–632
41. Arbelo E, Protonotarios A, Gimeno JR, et al (2023) 2023 ESC Guidelines for the management of cardiomyopathies. Eur Heart J 44:3503–3626
42. McDonagh TA, Metra M, Adamo M, et al (2023) 2023 Focused Update of the 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J 44:3627–3639
43. Baumgartner H, De Backer J, Babu-Narayan S V, et al (2021) 2020 ESC Guidelines for the management of adult congenital heart disease. Eur Heart J 42:563–645

44. Lang RM, Badano LP, Mor-Avi V, et al (2015) Recommendations for Cardiac Chamber Quantification by Echocardiography in Adults: An Update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. Journal of the American Society of Echocardiography 28:1-39.e14
45. McDonagh TA, Metra M, Adamo M, et al (2023) 2023 Focused Update of the 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J 44:3627–3639
46. Kovács A, Magunia H, Nicoara A, et al (2025) Challenges and opportunities in assessing right ventricular structure and function: a Roadmap for standardization, clinical implementation and research. Nat Rev Cardiol. https://doi.org/10.1038/s41569-025-01180-9
47. Mukherjee M, Rudski LG, Addetia K, et al (2025) Guidelines for the Echocardiographic Assessment of the Right Heart in Adults and Special Considerations in Pulmonary Hypertension: Recommendations from the American Society of Echocardiography. Journal of the American Society of Echocardiography 38:141–186
48. Medvedofsky D, Maffessanti F, Weinert L, et al (2018) 2D and 3D echocardiography-derived indices of left ventricular function and shape: relationship with mortality. JACC Cardiovasc Imaging 11:1569–1579
49. Sayour AA, Tokodi M, Celeng C, Takx RAP, Fábián A, Lakatos BK, Friebel R, Surkova E, Merkely B, Kovács A (2023) Association of right ventricular functional parameters with adverse cardiopulmonary outcomes: a meta-analysis. Journal of the American Society of Echocardiography 36:624–633
50. Corbett L, O'Driscoll P, Paton M, Oxborough D, Surkova E (2024) Role and application of three-dimensional transthoracic echocardiography in the assessment of left and right ventricular volumes and ejection fraction: a UK nationwide survey. Echo Res Pract 11:8
51. Soliman-Aboumarie H, Joshi SS, Cameli M, et al (2022) EACVI survey on the multi-modality imaging assessment of the right heart. European Heart Journal-Cardiovascular Imaging 23:1417–1422

52. Zhang J, Gajjala S, Agrawal P, et al (2018) Fully Automated Echocardiogram Interpretation in Clinical Practice. Circulation 138:1623–1635
53. Ghorbani A, Ouyang D, Abid A, He B, Chen JH, Harrington RA, Liang DH, Ashley EA, Zou JY (2020) Deep learning interpretation of echocardiograms. NPJ Digit Med 3:10
54. Ouyang D, He B, Ghorbani A, et al (2020) Video-based AI for beat-to-beat assessment of cardiac function. Nature 580:252–256
55. Ouyang D, He B, Ghorbani A, Lungren MP, Ashley EA, Liang DH, Zou JY (2019) Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. NeurIPS ML4H Workshop: Vancouver, BC, Canada 5:
56. He B, Kwan AC, Cho JH, et al (2023) Blinded, randomized trial of sonographer versus AI cardiac function assessment. Nature 616:520–524
57. Tokodi M, Magyar B, Soós A, et al (2023) Deep learning-based prediction of right ventricular ejection fraction using 2D echocardiograms. Cardiovascular Imaging 16:1005–1018
58. Magyar B, Tokodi M, Soós A, Tolvaj M, Lakatos BK, Fábián A, Surkova E, Merkely B, Kovács A, Horváth A (2022) RVENet: A large echocardiographic dataset for the deep learning-based assessment of right ventricular function. In: European Conference on Computer Vision. pp 569–583
59. Muhtaseb R, Yaqub M (2022) Echocotr: Estimation of the left ventricular ejection fraction from spatiotemporal echocardiography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp 370–379
60. Goyal A, Bengio Y (2022) Inductive biases for deep learning of higher-level cognition. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. https://doi.org/10.1098/rspa.2021.0068
61. Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1492–1500
62. Asch FM, Banchs J, Price R, Rigolin V, Thomas JD, Weissman NJ, Lang RM (2019) Need for a Global Definition of Normative Echo Values—Rationale and

Design of the World Alliance of Societies of Echocardiography Normal Values Study (WASE). Journal of the American Society of Echocardiography 32:157-162.e2

63. Bagyura Z, Kiss L, Édes E, et al (2014) Cardiovascular screening programme in the Central Hungarian region. The Budakalász Study. Orv Hetil 155:1344–1352
64. He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp 630–645
65. Loshchilov I, Hutter F (2016) Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983
66. Miyoshi T, Addetia K, Citro R, et al (2020) Left ventricular diastolic function in healthy adult individuals: results of the world alliance societies of echocardiography normal values study. Journal of the American Society of Echocardiography 33:1223–1233
67. Parzen E (1962) On estimation of a probability density function and mode. The annals of mathematical statistics 33:1065–1076
68. Leem S, Seo H (2024) Attention guided CAM: Visual explanations of vision transformer guided by self-attention. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp 2956–2964
69. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB (2008) General Cardiovascular Risk Profile for Use in Primary Care. Circulation 117:743–753
70. Tolvaj M, Kovács A, Radu N, et al (2024) Significant disagreement between conventional parameters and 3D echocardiography-derived ejection fraction in the detection of right ventricular systolic dysfunction and its association with outcomes. Journal of the American Society of Echocardiography 37:677–686
71. Szijártó Á, Magyar B, Szeier TÁ, et al (2025) Masked Autoencoders for Medical Ultrasound Videos Using ROI-Aware Masking. pp 167–176
72. Kovács A, Lakatos B, Tokodi M, Merkely B (2019) Right ventricular mechanical pattern in health and disease: beyond longitudinal shortening. Heart Fail Rev 24:511–520

9. Bibliography of the candidate's publications

9.1 Bibliography related to the present thesis

1. **Ádám Szijártó**, Béla Merkely, Attila Kovács, Márton Tokodi on behalf of, the QUEST-EF Investigators
   Deep Learning-Enabled Echocardiographic Assessment of Biventricular Ejection Fractions: The Dual-Task QUEST-EF Model
   European Heart Journal-Cardiovascular Imaging 26 : 8 pp. 1402-1405. , 4 p. (2025) DOI: 10.1093/ehjci/jeaf147
   IF: **6.6**
2. **Ádám Szijártó**, Bálint Magyar, Thomas Á. Szeier, Máté Tolvaj, Alexandra Fábián, Bálint K. Lakatos, Zsuzsanna Ladányi, Zsolt Bagyura, Béla Merkely, Attila Kovács, Márton Tokodi
   Masked Autoencoders for Medical Ultrasound Videos Using ROI-Aware Masking
   Lecture Notes in Computer Science 15186 pp. 167-176. , 10 p. (2025) 65 : 2 pp. 5-16. , 12 p. (2020) DOI: 10.1007/978-3-031-73647-6_16

9.2 Bibliography not related to the present thesis

1. **Ádám Szijártó**, Péter Lehotay-Kéry, Attila Kiss
   EXPERIMENTAL STUDY OF SOME PROPERTIES OF KNOWLEDGE DISTILLATION.
   Studia Universitatis Babes-Bolyai, Informatica 65 : 2 pp. 5-16. , 12 p. (2020) DOI: 10.24193/subbi.2020.2.01
2. Máté Tolvaj, Alexandra Fábián, Márton Tokodi, Bálint Lakatos, Alexandra Assabiny, Zsuzsanna Ladányi, Kai Shiida, Andrea Ferencz, Walter Schwertner, Boglárka Veres, Annamária Kosztin, **Ádám Szijártó**, Balázs Sax, Béla Merkely, Attila Kovács
   There is more than just longitudinal strain: Prognostic significance of

biventricular circumferential mechanics

Frontiers in Cardiovascular Medicine 10 Paper: 1082725 , 9 p. (2023) DOI: 10.3389/fcvm.2023.1082725

IF: **2.8**

3. Márton Tokodi, Bálint Magyar, Ádám Szijártó, Bálint Károly Lakatos, Attila Kovács

   Reply: Can Deep Learning Improve 2D Echocardiographic RV Assessment? First Important Steps

   JACC-Cardiovascular Imaging 16 : 12 pp. 1636-1636. , 1 p. (2023) DOI: 10.1016/j.jcmg.2023.10.015

4. **Ádám Szijártó**, Alexandra Fábián, Bálint Károly Lakatos, Máté Tolvaj, Béla Merkely, Attila Kovács, and Márton Tokodi

   A machine learning framework for performing binary classification on tabular biomedical data

   Imaging 15 : 1 pp. 1-6. , 6 p. (2023), DOI: 10.1556/1647.2023.00109

   IF: **0.7**

5. **Ádám Szijártó,** Ellák Somfai, András Lőrincz

   Design of a Machine Learning System to Predict the Thickness of a Melanoma Lesion in a Non-Invasive Way from Dermoscopic Images

   Healthcare Informatics Research 29 : 2 pp. 112-119. , 8 p. (2023), DOI: 10.4258/hir.2023.29.2.112

   IF: **2.3**

6. Richard Walter Schwertner, Márton Tokodi, Boglárka Veres, Anett Behon, Eperke Dóra Merkel, Richárd Masszi, Luca Kuthi, **Ádám Szijártó**, Attila Kovács, István Osztheimer, Endre Zima, László Gellér, Máté Vámos, László Sághy, Béla Merkely, Annamária Kosztin, Dávid Becker

   Phenogrouping and risk stratification of patients undergoing cardiac resynchronization therapy upgrade using topological data analysis

   Scientific Reports 13 : 1 Paper: 20594 , 13 p. (2023) DOI: 10.1038/s41598-023-47092-x

   IF: **3.8**

7. Boglárka Veres, Walter Richard Schwertner, Márton Tokodi, **Ádám Szijártó**, Attila Kovács, Eperke Dóra Merkel, Anett Behon, Luca Kuthi, Richárd Masszi, László Gellér, Endre Zima, Levente Molnár, István Osztheimer, Dávid Becker, Annamária Kosztin, Béla Merkely
Topological data analysis to identify cardiac resynchronization therapy patients exhibiting benefit from an implantable cardioverter-defibrillator
Clinical Research in Cardiology 113 : 10 pp. 1430-1442. , 13 p. (2024) DOI: 10.1007/s00392-023-02281-6
IF: **3.7**
8. **Ádám Szijártó**, Alina Nicoara, Mihai Podgoreanu, Márton Tokodi, Alexandra Fábián, Béla Merkely, András Sárkány, Zoltán Tősér, Sergio Caravita, Claudia Baratto, Michele Tomaselli, Denisa Muraru, Luigi Paolo Badano, Bálint Lakatos, Attila Kovács
Artificial intelligence-enabled reconstruction of the right ventricular pressure curve using the peak pressure value: a proof-of-concept study
European Heart Journal - Imaging Methods and Practice 2 : 4 Paper: qyae099 , 7 p. (2024) DOI: 10.1093/ehjimp/qyae099
9. Bálint K. Lakatos, Zvonimir Rako, **Ádám Szijártó**, Bruno R. Brito da Rocha, Manuel J. Richter, Alexandra Fábián, Henning Gall, Hossein A. Ghofrani, Nils Kremer, Werner Seeger, Daniel Zedler, Selin Yildiz, Athiththan Yogeswaran, Béla Merkely, Khodr Tello, Attila Kovács
Right ventricular pressure-strain relationship-derived myocardial work reflects contractility: Validation with invasive pressure-volume analysis
The Journal of Heart and Lung Transplantation 43 : 7 pp. 1183-1187. , 5 p. (2024) DOI: 10.1016/j.healun.2024.03.007
IF: **6**
10. FJolla Zhubi Bakija, Máté Tolvaj, **Ádám Szijártó**, Márton Tokodi, Andrea Ferencz, Bálint Károly Lakatos, Zsuzsanna Ladányi, Loretta Kiss, Zsolt Szelid, Pál Soós, Béla Merkely, Zsolt Bagyura, Attila Kovács & Alexandra Fábián
Long-term prognostic value of myocardial work analysis across obesity stages: insights from a community-based study

International Journal of Obesity , 10 p. (2025) DOI: 10.1038/s41366-025-01863-w

IF: **3.8**

11. Máté Tolvaj, Fjolla Zhubi Bakija, Alexandra Fábián, Andrea Ferencz, Bálint Lakatos, Zsuzsanna Ladányi, Ádám Szijártó, Borbála Edvi, Loretta Kiss, Zsolt Szelid, Pál Soós, Béla Merkely, Zsolt Bagyura, Márton Tokodi, Attila Kovács **Integrating Left Atrial Reservoir Strain Into the First-Line Assessment of Diastolic Function: Prognostic Implications in a Community-Based Cohort With Normal Left Ventricular Systolic Function** *Journal of the American Society of Echocardiography* 38 : 7 pp. 570-582. , 13 p. (2025) DOI: 10.1016/j.echo.2025.03.012

    IF: **6**

12. Márton Tokodi, **Ádám Szijártó** **From Promise to Practice: Reducing Research Waste in Deep Learning Model Development for Cardiovascular Imaging** *JACC-Cardiovascular Imaging* 18 : 7 pp. 765-767. , 3 p. (2025) DOI: 10.1016/j.jcmg.2025.05.003

13. Timea Teszak, Timea Barcziova, Csaba Bödör, Lajos Hegyi, Luca Levay, Beata Nagy, Attila Fintha, **Adam Szijarto**, Attila Kovacs, Bela Merkely, Balazs Sax **Donor-Derived Cell-Free DNA Versus Left Ventricular Longitudinal Strain and Strain-Derived Myocardial Work Indices for Identification of Heart Transplant Injury** Biomedicines 13 : 4 Paper: 841 , 12 p. (2025) DOI: 10.3390/biomedicines13040841

    IF: **3.9**

14. **Ádám Szijártó**, Alexandra Fábián, Bálint Károly Lakatos, Máté Tolvaj, Béla Merkely, Attila Kovács, and Márton Tokodi **Addendum to: A machine learning framework for performing binary classification on tabular biomedical data** *Imaging* 17 : 1 pp. 83-83. , 1 p. (2025) DOI: 10.1556/1647.2023.11109

15. Marius Keller, Alexandra Fábián, Andrea Bandini, **Ádám Szijártó**, Zoltán Tősér, Béla Merkely, Tim Heller, Marcia-Marleen Dürr, Peter Rosenberger, Attila Kovács & Harry Magunia
Impact of the right ventricular mechanical pattern assessed by three-dimensional echocardiography on adverse outcomes following cardiac surgery
Scientific Reports 15 : 1 Paper: 5623 , 13 p. (2025) DOI: 10.1038/s41598-025-89122-w
IF: **3.9**

10. Acknowledgements

As someone coming from a purely technological background, it was particularly challenging and humbling, yet at the same time an indescribably fulfilling experience to carry out my PhD research, which would not have been possible without the help of many people.

First and foremost, I would like to express my deepest gratitude to my two supervisors, Dr. Attila Kovács, whose guidance and immense support were invaluable in learning to navigate the initially confusing world of medical research, and Dr. Márton Tokodi, whose wide knowledge and commitment to perfectionism always pushed me to be a better researcher. Thank you both for your infinite patience and countless hours of hard work you dedicated to me and my research.

I would also like to express my sincere gratitude to Prof. Béla Merkely for giving me the opportunity to carry out my research projects, providing the background required for my research.

I would like to thank all my fellow researchers, in particular Dr. Alexandra Fábián, Dr Bálint Lakatos, Dr Máté Tolvaj, Dr Zsuzsanna Ladányi, and all co-authors of my papers for their scientific contributions. Without their help, none of my achievements would have been possible.

Many thanks to every member of the Argus Cognitive team, especially Bálint Magyar and Dr Janka Hatvani. Their technical knowledge and continuous advice helped me keep up with the latest technological innovations.

Last but not least, I would like to thank my friends and family, who initially convinced me to pursue academic research. Their unwavering support has helped me immensely, and I am truly grateful to them.