Automated echocardiographic surgical guidance in mitral valve repair

Igor George Lenting

February 2026

Master thesis Technical Medicine

AUTOMATED ECHOCARDIOGRAPHIC SURGICAL GUIDANCE IN MITRAL VALVE REPAIR: A FEASIBILITY STUDY

Igor George Lenting

Student number : 5057655

16-02-2026

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University ; Delft University of Technology ; Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)

Dept. of Cardiothoracic surgery, UMC Utrecht

14-07-2025 TM30004 – 05-03-2026 TM30004

Supervisor(s):

Dr. MD, MSc Amir H. Sadeghi

Dr. ir. Ruisheng Su

Thesis committee members:

Prof. dr. ir. John van den Dobbelsteen,	TU Delft (chair)
Dr. ir. Ruisheng Su	TU Eindhoven
Dr. Bardia Arabkhani (MD, PhD)	Erasmus MC
Dr. Amir H. Sadeghi (MD, PhD)	UMC Utrecht

Universiteit Leiden

TU Delft Delft University of Technology

Erasmus Universiteit Rotterdam

Abstract

Introduction: Accurate interpretation of two-dimensional transesophageal echocardiography (2D TEE) is essential for successful mitral valve repair, yet echocardiographic assessment is highly operator dependent and subject to variability. While automation has been explored in transthoracic echocardiography, standardized interpretation of 2D TEE at the scallop level remains largely unexplored. This study investigates the feasibility of automating key components of preoperative 2D TEE interpretation to support standardization of mitral valve surgery planning.

Methods: Preoperative 2D TEE recordings from 50 patients undergoing mitral valve repair were retrospectively collected. Separate algorithmic components were developed and evaluated for (1) automatic classification of standard mid-esophageal views and visible mitral scallops, (2) anatomical segmentation of mitral valve–related cardiac structures, (3) automated extraction of clinically relevant geometric measurements combined with end-systolic frame identification, and (4) automated detection of mitral valve prolapse. Deep learning–based models were trained using patient-level data splits, and performance was assessed using standard classification, segmentation, and agreement metrics. Measurement feasibility was evaluated in a subset of 10 mid-esophageal long-axis views.

Results: View classification achieved high accuracy across standard mid-esophageal views, while scallop-level classification showed moderate performance (accuracy 0.64), mainly limited by severe class imbalance. Segmentation of clinically relevant structures was moderate, with mean Dice scores of 0.76 for the mitral valve and 0.69 across measurement-critical structures. Automated geometric measurements showed good agreement with expert references in the mid-esophageal long-axis view, with limited systematic bias (Coaptation-septal distance −0.95 mm; aortic–mitral angle −3.28°; anterior-to-posterior leaflet (AL:PL) ratio −0.27; A2 length −3.85 mm). Larger deviations occurred in parameters sensitive to segmentation quality, particularly AL:PL ratio and A2 length, while wide limits of agreement reflected the small sample size. End-systolic frame identification was robust, with most predictions within two frames of expert annotation (mean difference 1.8). Prolapse detection achieved 0.90 accuracy, with one false positive.

Conclusion: This study demonstrates the feasibility of evaluating automated methods for multiple components of 2D TEE interpretation in mitral valve repair planning. View recognition, anatomical segmentation, automated measurement extraction, and prolapse detection showed encouraging performance within the limitations of the available dataset. Scallop-level interpretation remains constrained by limited data availability and class imbalance. These findings motivate further development using larger, multi-center datasets and suggest that standardized automated analysis may ultimately support improved consistency and enable large-scale, outcome-driven research in mitral valve surgery.

Table of Contents

Introduction

Mitral valve repair is central to the management of mitral valve disease and is associated with improved survival and preservation of left ventricular function compared to valve replacement (1-3). The demand for mitral valve surgery is rising, driven by an aging population (4). At the same time, increasing life expectancy implies that more patients may require reintervention later in life, underscoring the need for durable and well-planned repairs. Achieving such durability relies heavily on accurate preoperative assessment and intraoperative guidance, placing growing demands on imaging-based decision-making (5-7). This expanding patient population further amplifies the need for efficient, reliable, and reproducible surgical workflows.

Two-dimensional transesophageal echocardiography (2D TEE) remains a cornerstones of preoperative planning for mitral valve repair (3). Despite the advent of three-dimensional (3D) echocardiography, 2D TEE continues to be essential and complements 3D imaging by providing high temporal resolution, reliable anatomical landmarks, and robust image quality in a wide range of patients. Importantly, 2D TEE is the primary imaging modality by which surgeons and imaging specialists assess mitral valve morphology, leaflet motion, and pathological features relevant for repair strategy selection.

However, successful interpretation of 2D TEE requires substantial expertise. Surgeons and echocardiographers must mentally reconstruct the three-dimensional anatomy of the mitral valve from a series of two-dimensional views, a process that is cognitively demanding and associated with a steep learning curve (8). A critical step in this process is the accurate identification of which leaflet and scallops are visualized in each standard echocardiographic view. Errors or ambiguity at this stage can propagate into surgical planning decisions. For instance, intraoperative findings such as a newly observed anterior leaflet prolapse during saline testing, previously absent on preoperative echocardiography, should not prompt additional repair maneuvers (2). This reliance on preoperative expertise underscores the need for advanced diagnostic tools that can simplify the steep learning curve and enhance the surgeon's confidence in the initial surgical plan.

These challenges are further amplified by the increasing adoption of minimally invasive and endoscopy-assisted mitral valve procedures. These approaches reduce surgical trauma and accelerate postoperative recovery, but they also place greater demands on preoperative planning and intraoperative imaging interpretation (9). Despite this, adoption of endoscopy-assisted valve procedures remains limited. Endoscopic mitral and aortic valve replacements remain limited to a small number of expert surgeons (10). This shift demands exceptionally detailed preoperative planning. Surgeons must be well informed about valve reparability, optimal access routes, and appropriate cannulation techniques. High-quality imaging-based pre- and intraoperative strategies can play a pivotal role in streamlining these processes, reducing surgical time, and improving overall outcomes.

Despite its central role, current echocardiographic assessment remains constrained by several limitations, including operator dependency, manual measurement variability, and limited standardization (8, 11, 12). While automated methods have been proposed for mitral valve segmentation, these approaches have not addressed scallop-by-scallop interpretation in 2D TEE sequences, which remains a critical unmet need. For example, Chen et al., and Pouch et al., and developed frameworks for the mitral valve using three-dimensional echocardiography, focusing on reconstruction of global valve anatomy and quantitative geometry rather than scallop-level interpretation within standard 2D TEE (13, 14). Other echocardiographic automation studies mainly

address view classification, predominantly in transthoracic echocardiography, and do not provide scallop-level anatomical interpretation in standard 2D TEE views (15, 16).

To address these challenges, this study explores the feasibility of an artificial intelligence–based surgical guide for preoperative planning and intraoperative guidance in mitral valve surgery using 2D TEE sequences. The proposed guide is designed to support consistent image interpretation, reduce operator dependency, and facilitate objective decision-making. This work is explicitly positioned as a feasibility study, focusing on technical and clinical plausibility rather than definitive validation.

Methods

Study Design and Overview

This study is designed as a single-center, observational feasibility study aimed at the development of an artificial intelligence-based surgical guide for preoperative planning and intraoperative guidance in mitral valve surgery using 2D TEE. The study focuses on algorithm and model development, including view classification, anatomical segmentation, quantitative measurement extraction, and pathology detection.

Ethical Approval and Regulatory Considerations

Ethical approval was obtained during the initial phase of the thesis period prior to the start of data collection. The study was conducted as a non-WMO study in accordance with Dutch medical research regulations, as participants are not subjected to additional procedures nor required to follow specific behavioral rules beyond standard clinical practice.

Study Population and Data Sources

The study population consisted of patients scheduled for surgical mitral valve repair. Data were collected from a random selection of eligible patients. Predefined exclusion criteria were the absence of usable transesophageal echocardiographic data or insufficient image quality for reliable interpretation and annotation. No exclusion criteria were applied with respect to age, sex, or specific mitral valve pathology, provided that the echocardiographic data met minimum quality requirements.

Data collection

A dataset of 2D TEE sequences was acquired for the development and evaluation of the proposed algorithms and models. All included patients underwent TEE as part of routine preoperative assessment for mitral valve repair and therefore predominantly present with pathological mitral valve anatomy.

For each patient, unique 2D TEE sequences corresponding to standard mid-esophageal views were extracted. Duplicate views from the same patient were excluded to prevent data redundancy and reduce bias during model training.

Study Objective and System Overview

The objective of this study was to develop a surgical guide for preoperative planning and intraoperative guidance in mitral valve surgery based on 2D TEE sequences. The proposed system aims to support consistent image interpretation, reduce operator dependency, and facilitate objective and reproducible surgical decision-making.

The surgical guide was designed to provide the following core functionalities:

1. Automatic detection and classification of standard mid-esophageal 2D TEE views.
2. Identification of the mitral valve leaflets and scallops visible in each 2D TEE view.
3. Automated extraction of quantitative measurements relevant to preoperative planning, with a specific focus on parameters associated with systolic anterior motion (SAM) of the mitral valve.
4. Automated detection of mitral valve prolapse in the mid-esophageal long-axis (LAX) view.

System Architecture and Algorithmic Components

To achieve the stated objectives, the surgical guide was implemented as a modular pipeline combining artificial intelligence models with rule-based algorithms. Each component is designed to address a specific step in the interpretation and analysis of 2D TEE data.

The system consists of the following components:

1. A neural network for the classification of 2D TEE views and the identification of visible scallops.
2. A neural network for the segmentation of anatomically relevant cardiac structures required for downstream analysis.
3. An algorithm that derives standardized quantitative measurements from the segmentation outputs.
4. An algorithm for the automatic identification of the frame closest to end-systole, which will serve as the reference frame for quantitative measurements.
5. An algorithm that uses segmentation results to detect mitral valve prolapse.

Echocardiographic View Selection and Definition

Model development and evaluation were restricted to four standard mid-esophageal 2D TEE views commonly used in clinical practice: the Mid-Esophageal Two-Chamber View, the Mid-Esophageal Commissural View, the LAX View, and the Mid-Esophageal Four-Chamber View.

View classification follows established echocardiographic conventions. However, strict adherence to the below established criteria such probe angulation ranges was not enforced. Instead, classification was based on the presence of characteristic anatomical landmarks, reflecting real-world clinical practice where not all structures are consistently visible. In cases where image appearance deviates from textbook definitions, expert judgment was used to determine the most appropriate view classification. For example, if an expert identifies an image as a four-chamber view based on overall anatomical context, this classification was accepted as the reference label. This approach ensures that the system was trained on clinically realistic data rather than idealized imaging conditions.

Scallop Visibility and Anatomical Labeling Strategy

For each echocardiographic view, the mitral valve scallops considered visible are defined according to established anatomical conventions, while allowing for variability in image quality and probe orientation. The different standard views with their variation can be seen in figures 1 to 4.

Mid-esophageal four-chamber view

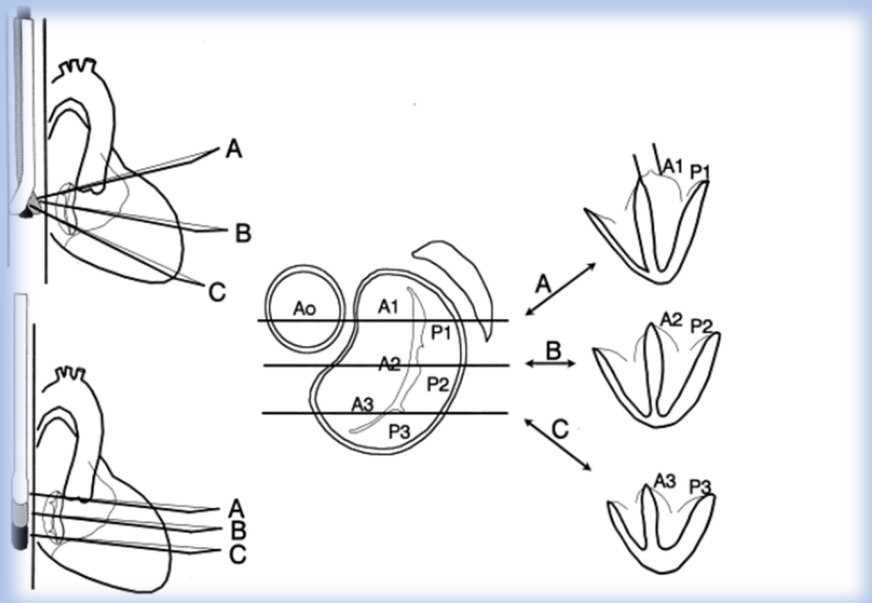


Figure 1; Different variations of the mid-esophageal four-chamber view. At probe angles of approximately 0–20°, this view typically displays all four cardiac chambers. Depending on the visible anatomical landmarks, the mitral valve scallops visualized may include A2–P2 (when neither the aorta nor the coronary sinus is visible), A1–P1 (when the aorta is visible), or A3–P3 (when the coronary sinus is visible).

Mid-esophageal commissural view

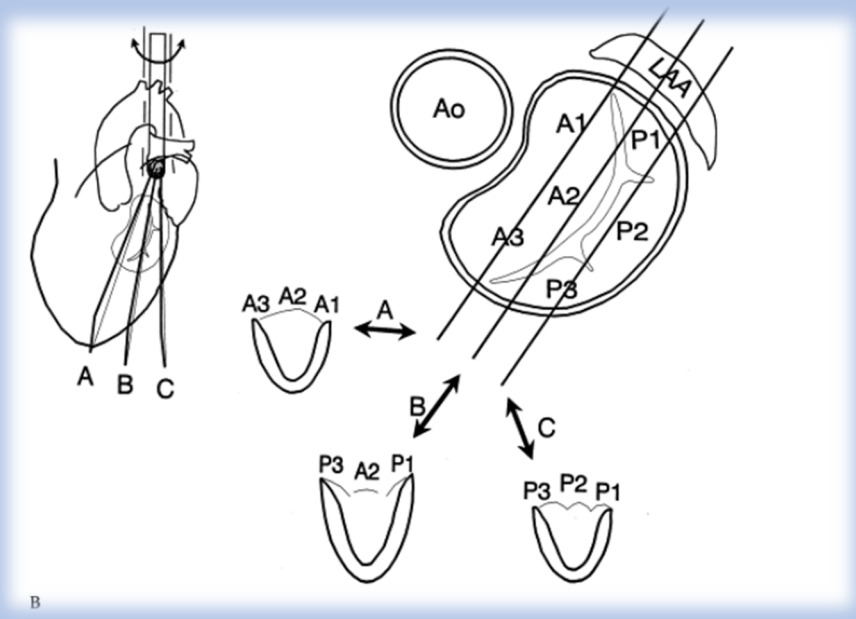


Figure 2; Different variations of the mid-esophageal commissural view. At probe angles of approximately 45–75°, this view generally displays the left atrium, left ventricle, and the mitral valve in plane, often with both papillary muscles visible. The most common scallop configuration observed in this view is P3–(A3),A2,(A1)–P1; however, continuous appearances of A3–A2–A1 or P3–P2–P1 may also occur, depending on probe orientation and anatomical variability.

Figure 3; Different variations of the mid-esophageal two-chamber view. At probe angles of approximately 80–100°, this view typically displays two cardiac chambers, with the left atrial appendage visible on the right side of the image. The mitral valve scallops visualized in this view are P3, A3, A2, and A1.

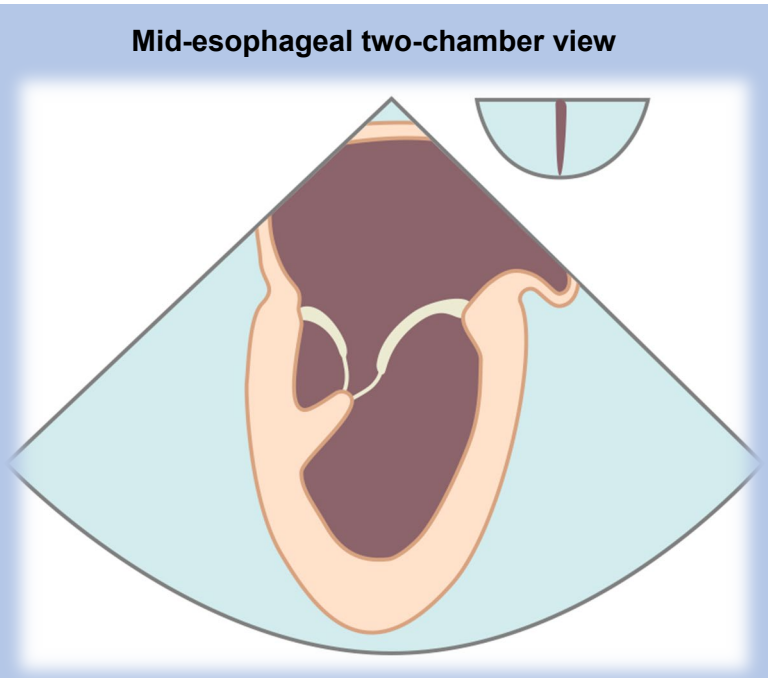
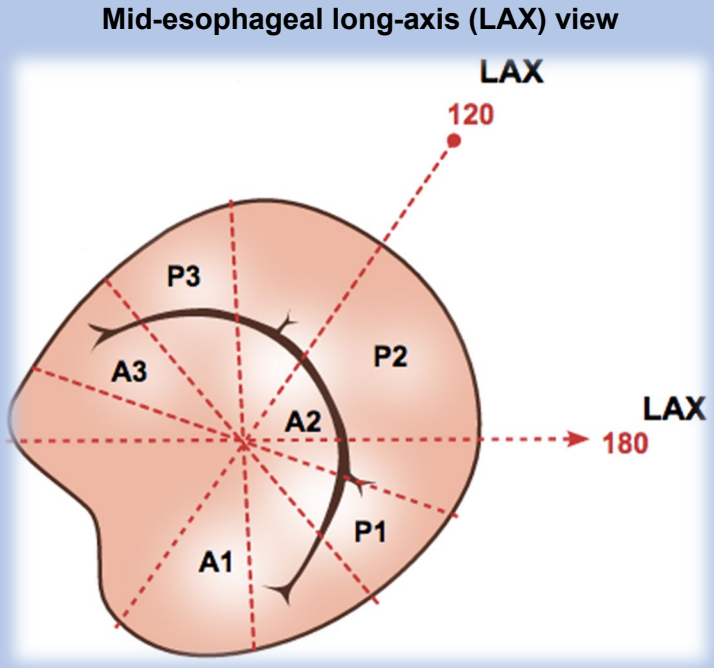


Figure 4; Different variations of the mid-esophageal long-axis view (120–180°). This view typically displays two cardiac chambers together with the aortic valve and the left ventricular outflow tract. Depending on probe orientation and visible anatomical landmarks, the mitral valve scallops visualized in this plane may include P2–A2, P3–A3, or P1–A1.



Annotation protocol

Annotation followed a structured, multi-step approach. I.G.L. from the Department of Cardiothoracic Surgery, who had been trained by an expert cardio-anesthesiologist (E.T.) and a cardiologist–radiologist (M.G.), served as the primary annotator. The experts adjudicated ambiguous cases flagged by the student.

For segmentation model development, eight frames were manually annotated per TEE sequence. These frames were selected at evenly spaced intervals across the temporal extent of each sequence to ensure representative coverage of the cardiac cycle.

Classification model development

Multiple classification models were developed to systematically assess the contribution of temporal context, probe angle information. All models aim to classify mitral valve scallops within standard mid-esophageal 2D TEE views. An overview of the classification pipeline, including input data, model configurations, outputs, and evaluation metrics, is provided in Table 1.

Dataset Splitting Strategy

The dataset was divided at patient level to prevent data leakage between training and evaluation phases. Data was split into:

- Training set: 80%
- Validation set: 10%
- Test set: 10%

In addition, five-fold cross-validation was performed on the training and validation sets to ensure robustness of model performance and reduce sensitivity to a single data split. The held-out test set remained untouched during model development and was used exclusively for final performance reporting.

Model Variants

To investigate the impact of temporal information and probe orientation on classification performance, the following model configurations were implemented and compared:

1. Frame-based classification

   Model was trained on individual 2D TEE frames only. This configuration serves as a baseline to evaluate performance achievable using static image information without temporal context.

2. Frame-based classification with limited temporal context and probe angle.

   Model was trained on individual frames augmented with a limited temporal window by including neighboring frames and probe angle as additional input. This approach assesses whether restricted temporal information improves classification while maintaining a frame-based formulation.

3. Sequence-based classification with probe angle

   The model was trained on complete 2D TEE sequences (≤40 frames), with shorter sequences padded by repeating the final frame. Probe rotation angle was appended as a normalized feature vector to the final convolutional embedding prior to classification, allowing integration of temporal dynamics and acquisition context.

Evaluation Metrics

Model performance was evaluated using standard classification metrics calculated on the held-out test set. Primary evaluation metrics included:

- Accuracy, to assess overall classification correctness
- F1-score, to account for class imbalance and provide a balanced performance measure

Confusion matrices were used to visually assess misclassification patterns between scallops in the best performing model.

segmentation model development

Segmentation of anatomically relevant structures is a critical step in enabling downstream measurement extraction and pathology detection. An overview of the individual system components, including input data, model or algorithm type, outputs, and evaluation metrics, is provided in Table 1.

Data Preprocessing and Splitting

All TEE sequences were cropped to a maximum of 40 frames per sequence to standardize input size and reduce computational requirements. The dataset splitting strategy for the segmentation networks mirrored those of the classification models, with 80% of patients assigned to the training set, 10% to the validation set, and 10% to the held-out test set, ensuring patient-level separation to prevent data leakage. In addition, five-fold cross-validation was performed on the training and validation sets to ensure robustness of model performance and reduce sensitivity to a single data split. An ensemble was made of the five-folds for testing.

2D Segmentation Network

A 2D U-Net architecture was trained on individual frames to segment the following cardiac structures:

- Mitral valve (split into 3 parts)
- Left ventricle
- Aorta
- Coronary sinus
- Left atrial appendage

Importantly, only the mitral valve regions and the left ventricle were directly used for downstream measurement extraction. All remaining structures were included to provide anatomical context and spatial orientation, but were not used in any quantitative measurements.

The network was initially trained using a subset of eight manually annotated frames per TEE sequence, selected to represent different phases of the cardiac cycle. After training, the 2D segmentation model was applied to segment the remaining frames within each sequence, resulting in full-frame annotations for all 40 frames. These automatically generated segmentations were subsequently used as input for training downstream models.

Temporal Segmentation Network

To exploit temporal continuity and inter-frame anatomical consistency, a temporally augmented 2D U-Net architecture was developed. Rather than operating on isolated frames, this network incorporates temporal context from neighboring frames within each TEE sequence, allowing it to leverage motion continuity and reduce frame-to-frame segmentation variability.

The network was trained on individual frames while incorporating information from adjacent frames, enabling it to model temporal dependencies without the data requirements of a true three-

dimensional convolutional architecture. This approach is particularly suitable given the limited size of the available dataset and the variable image quality inherent to intraoperative 2D TEE.

By integrating temporal information in this manner, the segmentation model was expected to improve robustness in anatomically ambiguous regions and in frames with reduced contrast, while remaining data-efficient and compatible with standard 2D annotation strategies.

Evaluation Metrics

Segmentation performance was evaluated on the held-out test set using the following metrics:

- Dice Similarity Coefficient: to quantify spatial overlap between predicted and reference segmentations
- Hausdorff Distance: to evaluate boundary accuracy, capturing worst-case segmentation errors

End-Systolic Frame Identification

Most quantitative measurements were referenced to the end-systolic frame, as this time point is the most appropriate for assessing systolic risk factors. Automatic identification of this frame was achieved using surrogate markers, including the positions, velocities, and accelerations of the mitral valve leaflets and aortic valve. These proxies allow the algorithm to robustly detect end-systole without manual annotation, enabling consistent measurement extraction across sequences.

Expert signals

Aortic Valve Motion Expert

The temporal derivative of the aortic valve opening signal is computed after Savitzky–Golay filtering. The absolute velocity of the aortic valve motion is inverted and normalized, such that low valve velocity (i.e., valve standstill) yields a high confidence score:

$$S_{AV}(t) = 1 - \mathrm{norm}(\mid \dot{AV}(t) \mid)$$

Mitral Leaflet Acceleration Expert

Angular trajectories of the anterior mitral leaflet edges are extracted from the segmentation masks. First- and second-order temporal derivatives are computed, and the absolute angular acceleration is used as an event detector for imminent mitral valve opening:

$$S_{event}(t) = \mathrm{norm}(\mid \ddot{\theta}_{R1}(t) \mid + \mid \ddot{\theta}_{R3}(t) \mid)$$

Mitral Valve Closure (Area) Expert

The projected mitral valve pixel area is normalized and inverted, yielding a maximal score when the valve is maximally closed:

$$S_{closure}(t) = 1 - \mathrm{norm}(A_{MV}(t))$$

Ensemble Confidence Function

The three expert signals are combined into a weighted ensemble confidence score:

$$C(t) = 2.0 \cdot S_{AV}(t) + 1.5 \cdot S_{event}(t) + 1.0 \cdot S_{closure}(t)$$

This weighting prioritizes aortic valve closure as the dominant physiological indicator, followed by mitral leaflet acceleration and valve closure state. The combined confidence signal is smoothed using a Gaussian filter to reduce noise. End-systole is identified as the most prominent peak in the sharpened confidence signal, obtained by combining the confidence curve with its second temporal derivative.

Automatic Measurement, Pathology Detection, and End-Systolic Frame Identification Algorithm Development

Accurate quantitative measurements and pathology detection are essential for guiding mitral valve repair and preoperative planning. To achieve this, a set of automated algorithms were developed to extract clinically relevant measurements from 2D TEE LAX views, detect mitral valve prolapse, and identify the end-systolic frame within each sequence. An overview of the individual system components, including input data, model or algorithm type, outputs, and evaluation metrics, is provided in Table 1.

Quantitative Measurements

The following measurements were automatically extracted from the LAX view on the end-systolic frame:

- C-sept distance – the distance between the coaptation point of the mitral valve leaflets and the interventricular septum, relevant for SAM risk.
- Anterior leaflet to posterior leaflet (AL:PL) ratio – the relative length of the anterior and posterior mitral leaflets, relevant for SAM risk.
- Aortic–mitral angle – the angle between the aortic and mitral valve planes, relevant for SAM risk.
- A2 leaflet length – the length of the central scallop of the anterior leaflet (A2), relevant for ring sizing and SAM risk is assessed on the frame automatically where it appears the longest.

Prolapse Detection

Mitral valve prolapse are detected using an algorithm that identifies leaflet segments moving above the annular plane during systole. By analyzing leaflet motion relative to the mitral annulus in the LAX view, the algorithm can determine the presence and approximate location of prolapse, providing objective information to guide repair planning.

Evaluation Metrics

Algorithm performance was evaluated on a separate clinical dataset consisting of 10 LAX views with-annotated measurements and prolapse labels. Evaluation metrics included were:

- Accuracy, sensitivity, and specificity for prolapse detection, assessing whether the algorithm correctly identifies the presence or absence of prolapse if applicable.
- Bland–Altman plots for agreement analysis of automated versus expert measurements and end-systolic frame detection.

Component	Input Data	Model / Algorithm	Output	Evaluation Metrics
View Classification & Scallop Identification	Individual 2D TEE frames with limited temporal context ± probe angle; full 2D TEE sequences with probe angle	Neural network (frame-based and sequence-based configurations)	Classified TEE view; identified visible scallops	Accuracy, F1-score, Confusion matrix
Anatomical Segmentation (2D)	Individual TEE frames (≤40 per sequence)	U-net variations 2D/temporal	Segmentation masks of: Mitral valve, Left ventricle, Aorta, Coronary sinus, Left atrial appendage	Dice Similarity Coefficient and Hausdorff Distance
Automatic Measurement Extraction	Segmented LAX frames, end-systolic frame	Rule-based algorithm	C-sept distance, AL:PL ratio, LV size, Aortic–mitral angle, A2 length	Bland–Altman analysis of temporal error between automated and expert-annotated measurements.
End-Systolic Frame Identification	LAX sequences, leaflet & aortic valve positions, velocities, accelerations	Motion analysis algorithm	Frame corresponding to end-systole	Bland–Altman analysis of temporal error between automated and expert-annotated end-systolic frames.
Prolapse Detection	Segmented LAX frames	Rule-based	Presence/location of mitral valve prolapse	Accuracy, Sensitivity, Specificity

Tabel 1. Overview of the modular components of the proposed surgical guide, including input data, model or algorithm type, generated outputs, and corresponding evaluation metrics for each processing step. Abbreviations: TEE = Transesophageal echocardiography), LAX = mid esophageal long axis view, AL:PL = anterior–posterior leaflet, C-sept = coaptation septal, LV = left ventricle

Results

Ethical approval

Ethical approval for this study was obtained prior to data collection in accordance with local institutional and national regulations. The study was classified as a non-WMO study, as no additional procedures were performed and patient care was not altered beyond standard clinical practice. The approved study protocol, informed consent form, data management plan, and related regulatory documents are provided in the Trial Master File.

Study Cohort and Dataset Characteristics

The dataset used for training, validation, and testing comprised preoperative TEE recordings from 50 patients scheduled for mitral valve repair and therefore presenting with pathological mitral valve anatomy. During random patient selection no patients were excluded. For each patient, unique TEE sequences were extracted without duplication, resulting in an average of approximately three sequences per patient and a total of 142 TEE sequences (N = 142).

The dataset was highly skewed. Notably, no A1–P1 configuration was identified in the LAX view across the 50 patients included in this study. The least represented classes were the commissural view in which only the anterior or posterior leaflet scallops were visible (COM AP, N = 7) and the LAX view with P3–A3 visible (LAX A3–P3). A histogram illustrating the distribution of view–scallop combinations is shown in Figure 5.

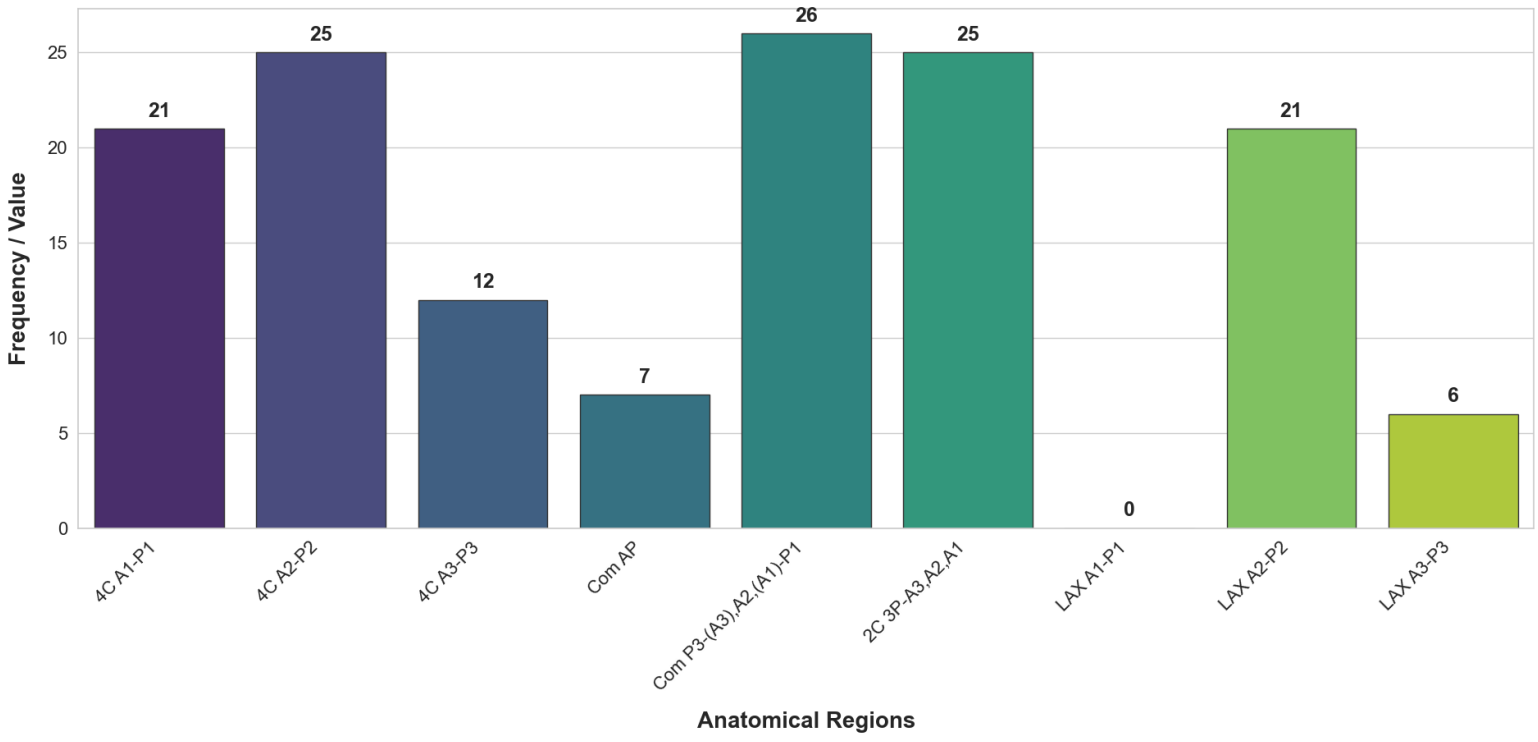


Figure 5; Histogram showing the distribution of view–scallop combinations in the dataset. Abbreviations: 4C = four-chamber view, COM = commissural view, 2C = two-chamber view, LAX = long-axis view; A1–A3 and P1–P3 denote anterior and posterior mitral valve scallops. The histogram highlights a highly imbalanced class distribution.

Each TEE sequence was cropped to a maximum of 40 frames; shorter sequences were padded by repeating the final frame. In the vast majority of cases, this frame window encompassed a complete cardiac cycle. This preprocessing step resulted in a total of 3,402 frames available for model development and evaluation. Of the 142 TEE sequences, 87 were manually annotated, with 8 frames per sequence labeled and used for training, validation and testing of the segmentation models. A flowchart illustrating the dataset composition and preprocessing steps is shown in Figure 6.

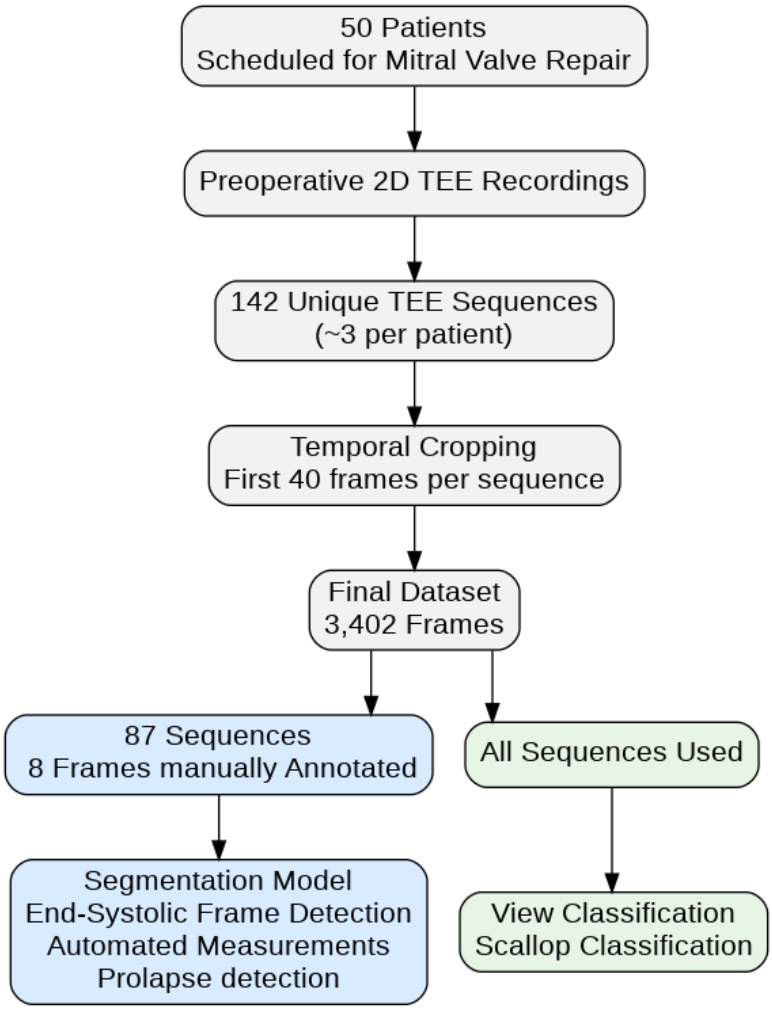


Figure 6; Overview of the dataset composition and preprocessing pipeline, showing patient inclusion, TEE sequence extraction, temporal cropping, and allocation of data to segmentation- and classification-related model components.

Annotation Reliability and Quality Control

An annotation protocol developed in collaboration with the expert MG is provided in the Appendix under “Protocol summary”. Annotation quality was assessed through expert review. Primary annotations for view classification, scallop labeling, and anatomical segmentation were performed by trained annotators under continuous expert supervision. Ambiguous classification cases were flagged during annotation and subsequently reviewed and adjudicated by experts. In total, 70 scallop classification cases were marked as ambiguous, of which 16 were revised following expert review (22%).

Classification Model

The classification models were evaluated on a held out test set consisting of 15 TEE sequences, covering multiple standard TEE views. Model training and validation was performed on a dataset of 127 TEE sequences. All results reported in this section therefore reflect generalization performance on unseen data.

Best-Performing Model and Overall Results

The best-performing classification model was a ResNet-based architecture trained on entire TEE sequences with probe angle information included as an additional input feature. This model achieved 100% accuracy in identifying the standard 2D TEE view, correctly classifying all sequences as four-chamber, two-chamber, commissural, or LAX. No misclassifications were observed at the view level, indicating robust discrimination between standard echocardiographic views.

Within correctly identified views, scallop classification performance varied. The overall scallop classification accuracy was 0.64, with a corresponding F1-score of 0.58. A confusion matrix illustrating the distribution of scallop-level predictions is shown in Figure 7.

Importantly, all observed misclassifications occurred within the same standard view. That is, scallops were occasionally misassigned to adjacent or neighboring scallop classes, but the underlying view was always correctly identified. This suggests that the model learned robust global anatomical context, while scallop discrimination remains more challenging.

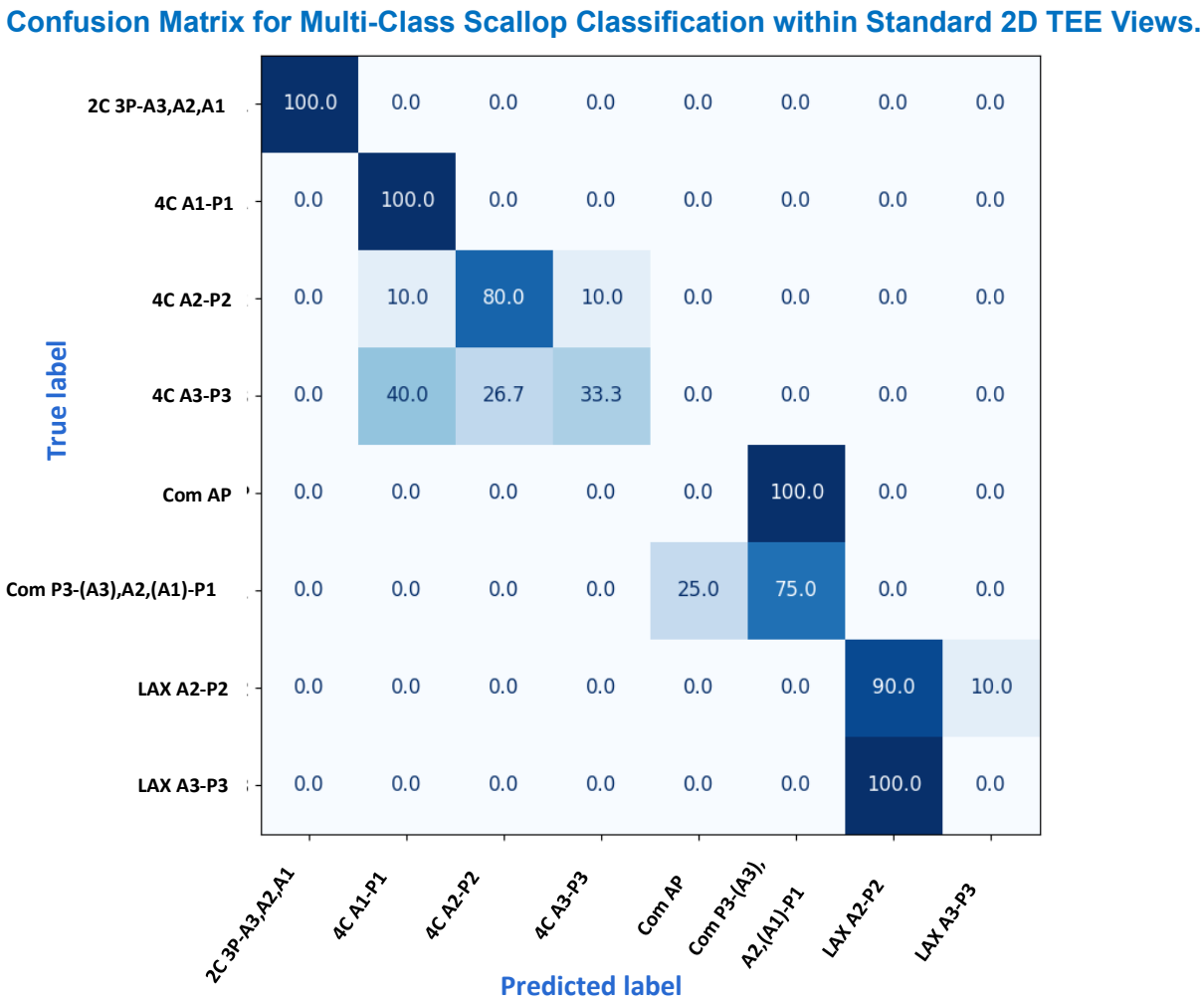


Figure 7; Confusion matrix of scallop-level classification performance. The matrix displays the performance of the ResNet model trained on entire sequences with integrated probe angle data. The y-axis represents the ground truth labels, while the x-axis represents the model's predictions. The diagonal elements indicate correct classifications, whereas off-diagonal elements highlight specific areas of anatomical ambiguity. Abbreviations: 4C = four-chamber view, COM = commissural view, 2C = two-chamber view, LAX = mid esophageal long-axis view;

Effect of Temporal Context and Model Design

When training the classifier on individual frames only, performance was poor, with an accuracy of 0.35 and an F1-score of 0.24. Although this frame-based approach substantially increased the number of training samples, from 142 sequences to 3,402 frames, it failed to provide sufficient anatomical and temporal context for accurate classification.

As an intermediate strategy, a model was trained on individual frames augmented with limited temporal context, using the five neighboring frames as additional input. This approach improved performance to an accuracy of 0.41 and an F1-score of 0.35, confirming that temporal information contributes meaningfully to classification. However, performance remained inferior to that of the fully sequence-based model as can be seen in table 2.

Model Configuration	Scallop Accuracy	F1-score	Notes
Frame-based classifier	0.35	0.24	Insufficient anatomical and temporal context
Frame-based + limited temporal context + probe angle	0.41	0.35	Temporal context improves performance but remains limited
Sequence-based + probe angle (ResNet)	**0.64**	**0.58**	**Best-performing model**

Table 2; showing accuracy and F1-score of different models

Impact of Dataset Imbalance

The dataset was highly imbalanced, with most samples originating from a small subset of view–scallop configurations. As illustrated in Figure 5, stratified train–validation–test splitting was applied to preserve representation of low-frequency classes, resulting in very small effective training sets for these views. Consequently, the model was required to learn scallop-level distinctions from a limited number of examples, likely contributing to reduced performance in underrepresented classes. Class imbalance was partially mitigated using focal loss; however, the limited absolute number of samples in rare view–scallop configurations remained a key limiting factor.

Interpretation

Taken together, these results demonstrate that automatic view classification is highly reliable, while scallop-level classification is feasible but likely constrained by data imbalance. The fact that all classification errors occurred within correctly identified views indicates that the model effectively captures global anatomical structure. Further improvements in scallop classification performance are expected to require a more balanced dataset and increased representation of under-sampled views, particularly the LAX and commissural views.

Segmentation Model

The segmentation models were evaluated on a held-out test set consisting of 11 TEE sequences, covering multiple standard TEE views. Model training and validation was performed on a dataset of 76 TEE sequences. All results reported in this section therefore reflect generalization performance on unseen data.

Quantitative Comparison of Segmentation Models

Incorporating limited temporal context through a custom nnU-Net training strategy (±2 frames) resulted in small improvements in segmentation performance compared with the standard frame-based nnU-Net (17). A quantitative comparison between the standard frame-based nnU-Net and the temporally augmented nnU-Net is presented in table 3.

Mean Dice scores differed only marginally, with differences generally ≤0.01–0.02 for measurement-critical structures, and no consistent improvement attributable to temporal augmentation. For the mitral valve subregions, Dice scores were comparable between models, and performance for the left ventricle was similar or slightly lower for the temporally augmented model. Formal statistical testing was not performed, as the small size of the held-out test set limited statistical power and would not support reliable inference.

Contextual structures showed similarly small and inconsistent differences, with no systematic benefit from temporal information. The overall mean Dice score across all regions was identical for both models (0.65).

A full 3D nnU-Net architecture was explored but training was terminated early, as performance plateaued at a Dice score of approximately 0.33. This limitation is attributed to data scarcity and insufficient sample size to support high-capacity three-dimensional models in this feasibility setting.

Taken together, these results indicate that limited temporal augmentation using three adjacent frames does not meaningfully improve segmentation performance in this dataset. While stronger temporal modeling or a larger temporal window may yield benefits, segmentation accuracy when using multiple frames may be affected by variability in temporal spacing between frames, as frame-to-frame time intervals differ across sequences and may introduce inconsistent motion information.

Quantitative Segmentation Accuracy of Temporal nnU-Net

The mitral valve was subdivided into three regions (left, middle, and right) to enable region-specific analysis and support scallop-level measurement extraction. Across all sequences, the temporally augmented model achieved high overlap for these mitral valve regions, with an overall mean Dice score of 0.76, indicating substantial agreement between predicted and reference segmentations. Hausdorff distances for the mitral regions were relatively low (12.6–31.3 pixels), reflecting good boundary alignment for the structures most critical to scallop-level interpretation.

The left ventricle achieved a mean Dice score of 0.49. While these values appear moderate in absolute terms, the accompanying Hausdorff distance of approximately 72 pixels indicates that segmentation errors are driven by localized boundary deviations rather than gross mislocalization.

Contextual anatomical structures, including the aorta, coronary sinus, and left atrial appendage, showed greater variability in overlap and boundary metrics. Dice scores for these regions ranged from 0.47 to 0.82, with higher Hausdorff distances reflecting intermittent visibility, partial volume effects, and inconsistent acoustic windows. As these structures are included solely to support anatomical orientation, their lower quantitative performance does not affect downstream measurement extraction.

Category	Anatomical Structure	Dice (Standard nnU-Net)	Mean Dice (Temporal nnU-Net)	Hausdorff-distance (Temporal nnU-Net)
Measurement-critical structures	Mitral valve – left scallop	0.79	0.80	12.6
	Mitral valve – middle scallop	0.70	0.70	49.1
	Mitral valve – right scallop	0.78	0.79	31.3
	Mitral valve (mean)	0.76	0.77	31.0
	Left ventricle	0.51	0.49	72.3
	Mean (critical structures)	0.69	0.67	41.3
Contextual structures	Aorta	0.48	0.49	73.2
	Coronary sinus	0.84	0.82	8.1
	Left atrial appendage	0.42	0.47	43.8
Overall	**Total mean (all regions)**	0.65	0.65	41.4

Table 3. Similarity coefficients per anatomical structure for the normal and temporal nnU-Net models. Hausdorff-distance are reported only for the best-performing model to avoid redundancy.

Qualitative Segmentation Results of the Temporal nnU-Net

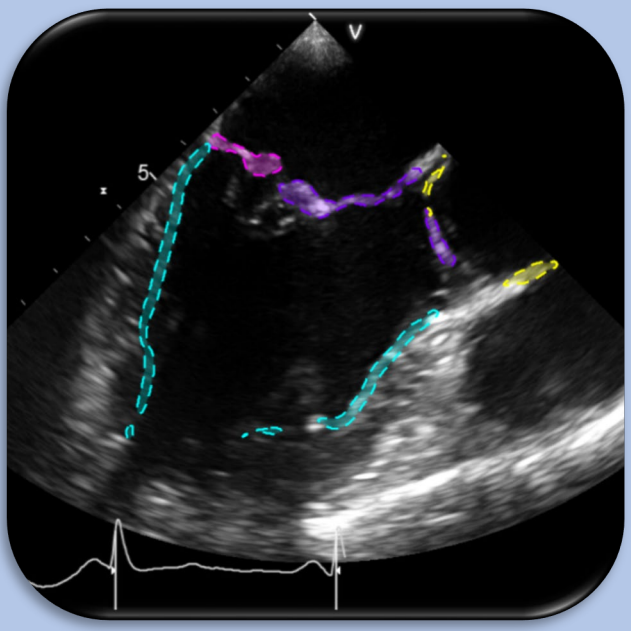
Qualitative inspection confirmed that segmentation of the mitral valve was spatially coherent and anatomically plausible across views and sequences. Leaflet bodies and annular regions were consistently identified, providing a reliable basis for geometric measurements.

For the left ventricle, segmentation accuracy was highest in the upper ventricular region, encompassing the basal and mid-ventricular segments. These regions are directly used for geometric reference and measurement extraction in the proposed pipeline. Segmentation errors were primarily observed in the lower ventricular cavity, which exhibits increased noise, motion artifacts, and partial visibility in 2D TEE. As this region is not used in any downstream analysis, its reduced segmentation accuracy does not compromise the functional applicability of the model.

Contextual structures demonstrated greater variability, particularly in sequences with limited acoustic windows or intermittent visibility. However, these structures were rarely completely mislocalized; segmentation errors predominantly occurred at boundaries rather than at the level of gross anatomical placement.

Segmentation preview of Temporal - nn-Unet

Mid-esophageal long-axis view



Representative examples of segmentation results obtained with the temporal nnU-Net. The figure 8 to 11 shows both well-segmented and suboptimal segmented 2D TEE frames across different views. In well-segmented cases, mitral valve and ventricular anatomy are accurately delineated, supporting reliable downstream automatic measurement extraction. In suboptimal cases, segmentation errors, such as inter-region misclassification of mitral valve subregions, are observed.

Legend

Mitral left | Mitral right | Left atrial appendage
Mitral middle | Left ventricle | Coronary sinus | Aorta

Figure 8; Example of suboptimal segmentation in a **mid-esophageal long-axis view**. The segmentation model incorrectly classified a portion of the aorta as part of the right mitral leaflet region. The remaining mitral leaflets are segmented accurately. The distal ventricular region, where image contrast is limited, is segmented less accurately. As this region is not used for downstream measurement extraction, the observed segmentation error has no practical impact.

Mid-esophageal two-chamber view

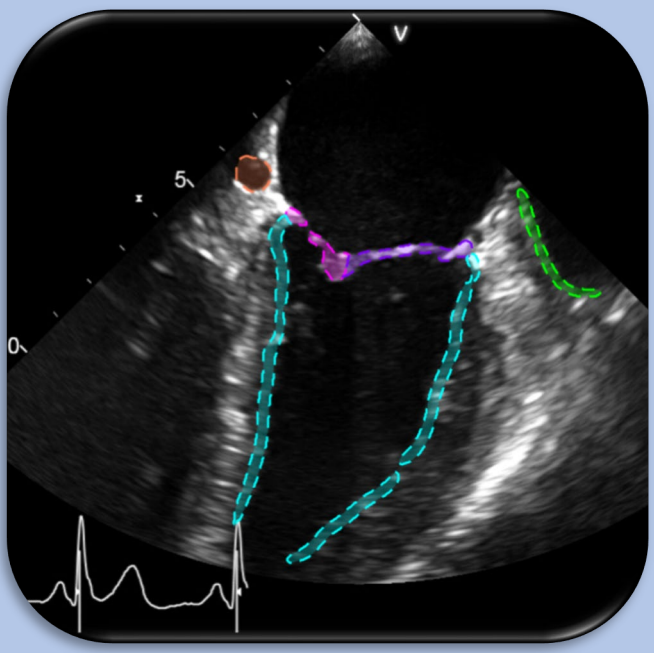


Figure 9; Example of an optimal segmentation in a **mid-esophageal two-chamber view**. The segmentation model correctly segmented all structures.

Legend

Mitral left | Mitral right | Left atrial appendage
Mitral middle | Left ventricle | Coronary sinus | Aorta

Segmentation preview of Temporal - NN-Unet

Mid-esophageal commissural view

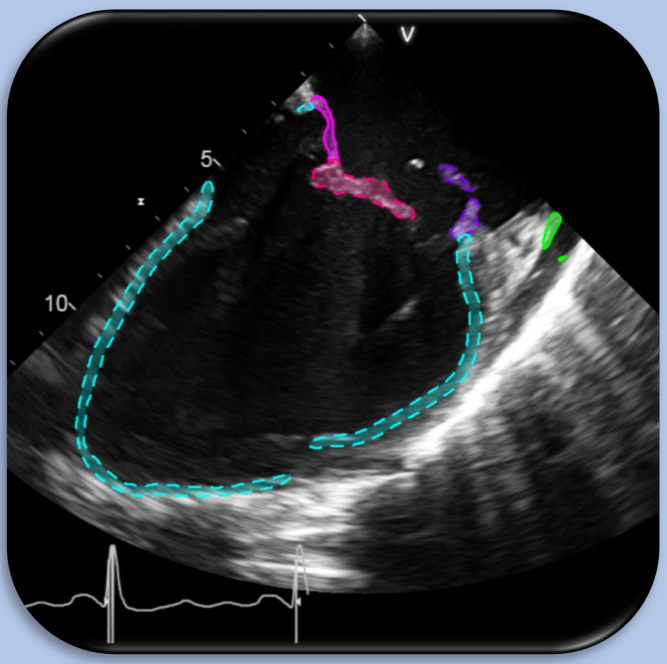


Figure 10; Example of a segmentation in a **mid-esophageal commissural view**. The segmentation model accurately delineated all relevant anatomical structures. A small discontinuity is present in the ventricular segmentation.

Legend

Mitral left | Mitral right | Left atrial appendage
Mitral middle | Left ventricle | Coronary sinus | Aorta

Mid-esophageal four-chamber view

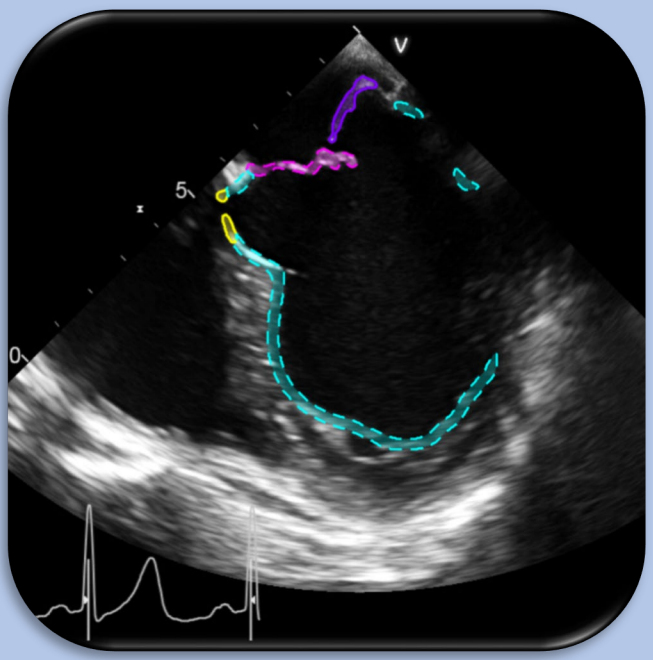


Figure 11; Example of optimal segmentation in a **mid-esophageal four-chamber view**. The segmentation model accurately delineated all relevant anatomical structures. The left ventricle was partially under-segmented in regions with low image contrast.

Legend

Mitral left | Mitral right | Left atrial appendage
Mitral middle | Left ventricle | Coronary sinus | Aorta

Automatic end-systolic frame detection

Accurate assessment of SAM–related parameters requires consistent identification of the end-systolic (ES) frame, defined as the moment immediately following aortic valve closure and preceding mitral valve opening.

To identify this frame automatically, an ensemble-based detection algorithm was developed, combining three complementary signal sources derived from segmentation and kinematic analysis of 2D TEE sequences. Each source acts as an independent “expert,” capturing a different physiological marker of end-systole. More information on the expert signals is provided in the appendix.

Measurement Agreement with Clinical Reference

For end-systolic frame identification, agreement analysis showed that automatically selected frames were generally within a small temporal offset of expert-annotated frames. Deviations were predominantly limited to two frames and did not result in visually apparent differences in leaflet configuration. In a single case, a larger discrepancy of approximately 20 frames was observed, attributable to reduced segmentation quality in that sequence. A Bland–Altman plot and the corresponding expert signals are shown in Figures 12 and 13.

Because both the clinical reference and the algorithm operate in discrete frame space, agreement was quantified in number of frames rather than seconds. End-systolic identification in routine TEE assessment is performed by scrolling frame-by-frame and selecting a specific image within the cine loop, and the algorithm similarly outputs a discrete frame index. The decision variable is therefore inherently discrete. Expressing deviations in frames directly reflects the practical selection process and aligns with how inter-observer variability is assessed clinically.

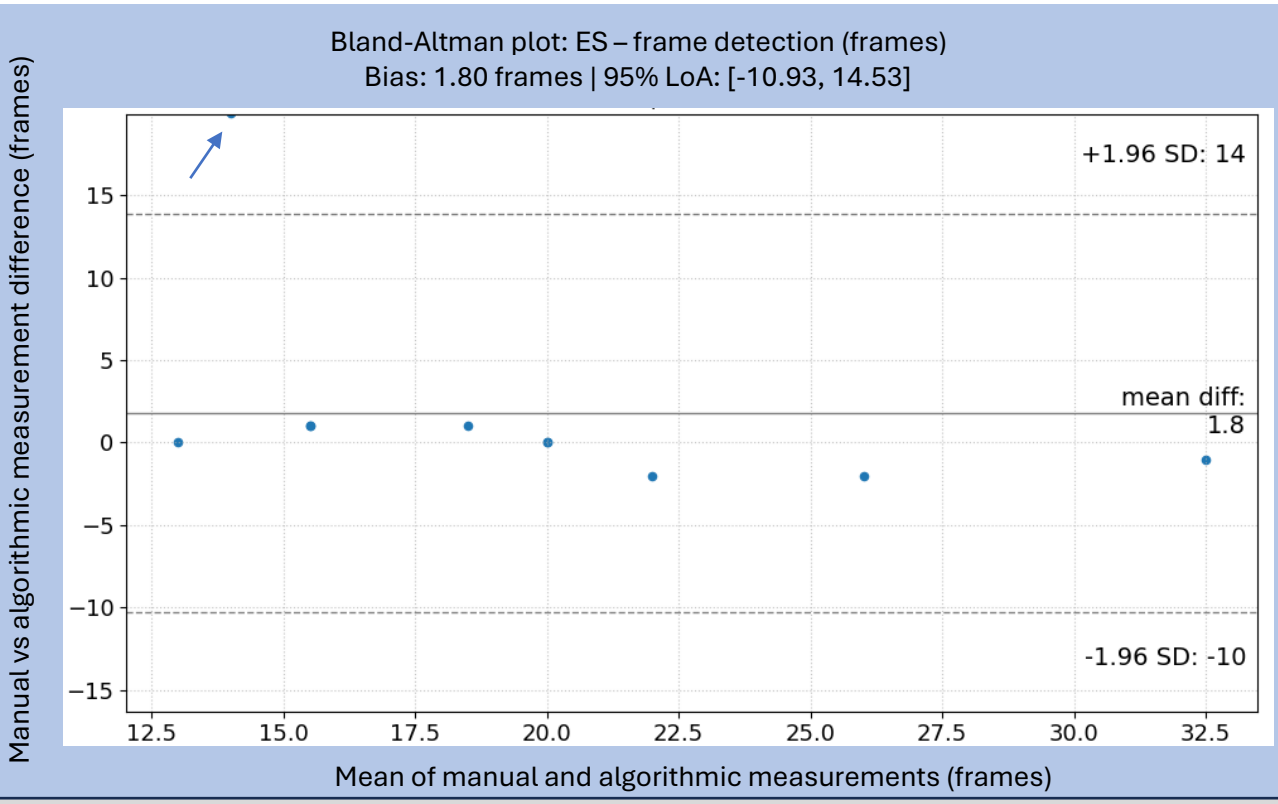


Figure 12. Bland–Altman plot showing agreement between manual and algorithmic **end-systolic (ES) frame detection**. Ten measurements are included. In most cases, the error is approximately **two frames** which on average corresponds to 30 milliseconds, with differences appearing to be normally distributed. A single large outlier, with a discrepancy of approximately 20 frames, is present and substantially increases the limits of agreement; this outlier is indicated by the arrow in the figure.

The clinical relevance of a given frame error depends on the specific TEE sequence, as frame rates vary between recordings. In sequences where consecutive frames show minimal anatomical change, a difference of one or two frames is unlikely to affect measurement outcomes. In contrast, when anatomical motion between frames is more pronounced, larger frame selection errors may result in substantial measurement differences.

Based on discussions with expert E.T., the expected inter-observer variability for end-systolic frame identification is approximately **±1 frame**. Importantly, end-systolic frame detection was relatively robust to segmentation errors, as the algorithm integrates multiple features, such as leaflet and aortic valve motion, velocities, and accelerations, rather than relying on a single segmentation-derived parameter.

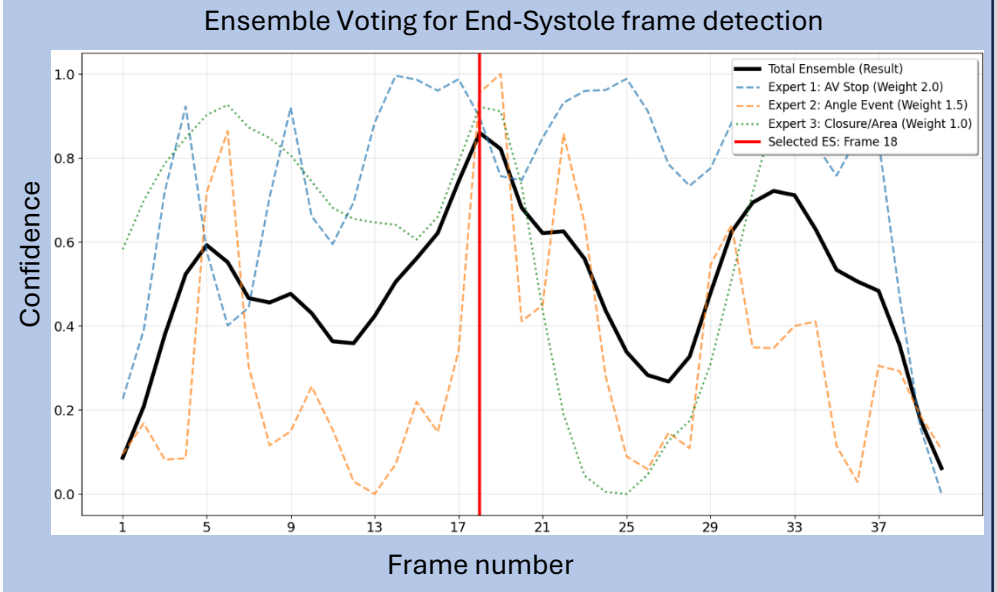


Figure 13. Example of **ensemble voting for end-systolic frame selection** in one patient. In this case, the difference between the automatically selected frame and the expert reference was **one frame**. The deviation appears to be primarily driven by the closure/area expert (blue line), which identifies the moment of maximal mitral valve closure. However, the clinically relevant end-systolic frame corresponds to the moment just before the mitral valve begins to open. This temporal offset suggests that the contribution of the closure/area expert could be improved by applying a small temporal shift of one to two frames along the time axis.

Automatic Measurements

Automatic measurement performance was evaluated on a held-out test set of 10 patients, all assessed in the LAX view. For each case, clinically relevant geometric parameters related to SAM risk were extracted automatically and compared with expert-derived reference measurements. Given the limited sample size (N=10), agreement analyses should be interpreted descriptively rather than as definitive performance estimates.

Overall, the system was able to successfully extract all intended measurements, including C-sept distance, anterior–posterior leaflet ratio, A2 leaflet length, and the aortic–mitral angle, in the majority of cases. Except for cases in which coaptation was absent; in those cases, the C-sept distance is not physiologically defined. These cases were therefore excluded from the quantitative comparison.

Measurement Agreement and with Clinical Reference

Agreement between automated and expert-derived measurements was assessed using Bland–Across parameters, agreement with clinical reference measurements was generally good, with limited systematic bias observed in Bland–Altman analyses (Figures 14 –21). A slight tendency toward underestimation by the automated measurements was noted for several parameters. Importantly, inspection of individual cases suggested that small underestimations were often associated with greater internal consistency and anatomical plausibility of the automated measurements, highlighting the inherent variability and subjectivity of manual clinical measurements.

Measurement performance varied between parameters and was strongly influenced by segmentation quality. Parameters that depend directly on precise leaflet delineation, most notably the AL:PL ratio and A2 leaflet length, showed greater variability and wider limits of agreement. In contrast, measurements such as the aortic–mitral angle and C-sept distance were less sensitive to local segmentation inaccuracies and demonstrated more stable agreement patterns.

Several observed discrepancies were attributable to known limitations of clinical measurement tools, particularly the lack of dedicated angle measurement functionality and limited zoom or cursor precision in commonly used ultrasound software. As a result, expert-derived reference measurements themselves are subject to non-negligible uncertainty, which must be considered when interpreting agreement metrics.

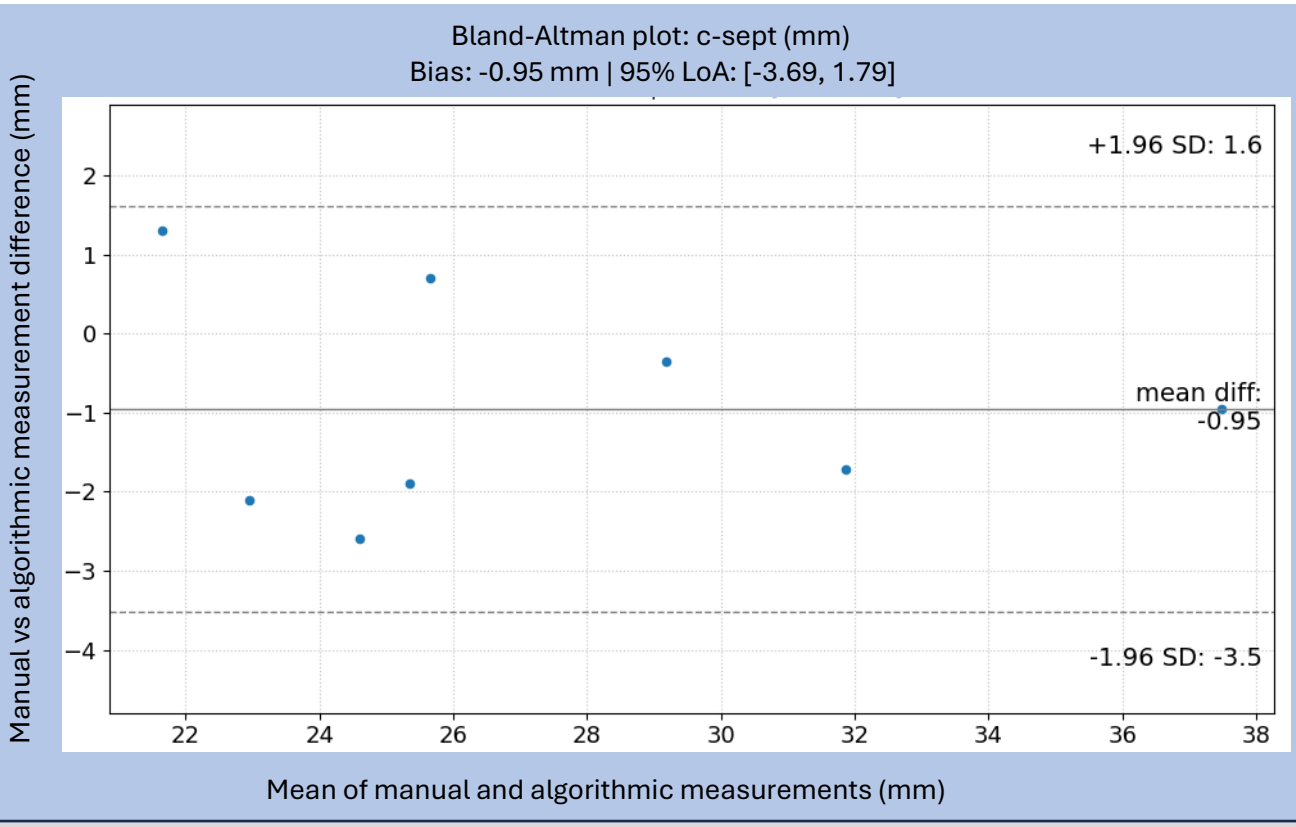


Figure 14. Bland–Altman plot showing agreement between manual and algorithmic **C-sept distance** measurements. Eight measurements are included; in two LAX cases, coaptation was absent and C-sept distance could therefore not be measured. A tendency toward underestimation by the algorithm is observed. The spread of the differences exceeds **2 mm** in two cases. An error margin of approximately 2 mm would generally be considered acceptable, as it is close to the expected **inter-observer variability**, based on discussions with expert E.T.; However, clinically, even an error of 2 mm may have a significant impact on surgical decision-making

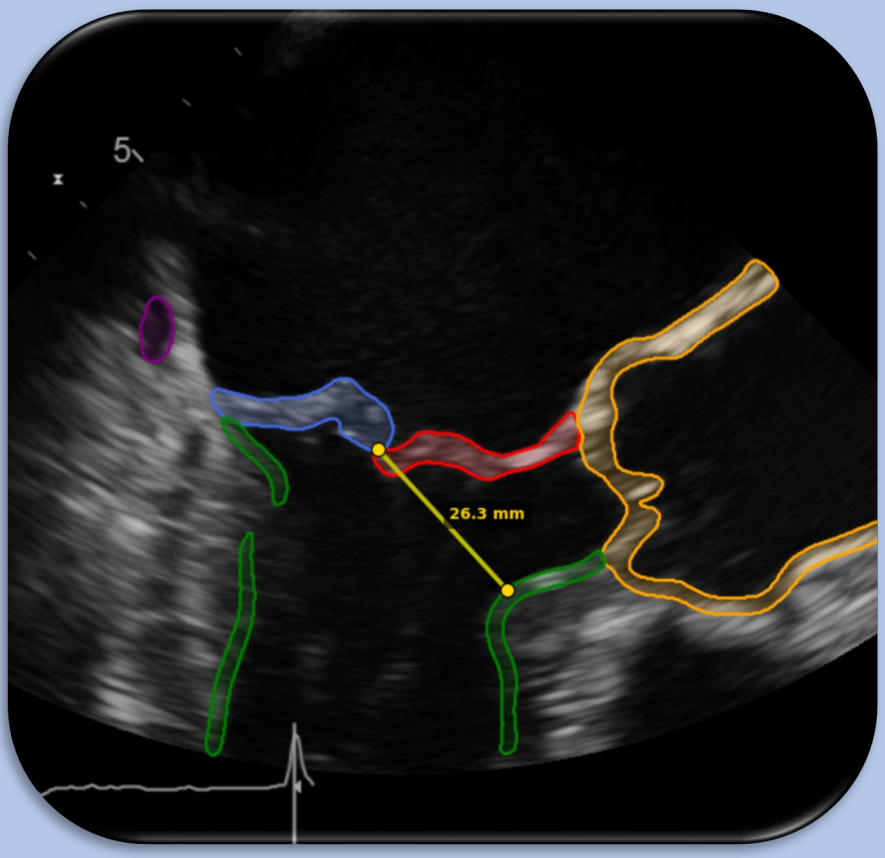


Figure 15. Example of automated **C-sept distance** measurement for one patient. In this case, the difference between the algorithmic and clinical measurement was **1.99 mm**. Despite this numerical discrepancy, the figure demonstrates accurate segmentation and anatomically consistent measurement by the algorithm. This example illustrates that small differences between manual and automated measurements may partly reflect inherent variability and uncertainty in clinical measurement rather than true algorithmic error.

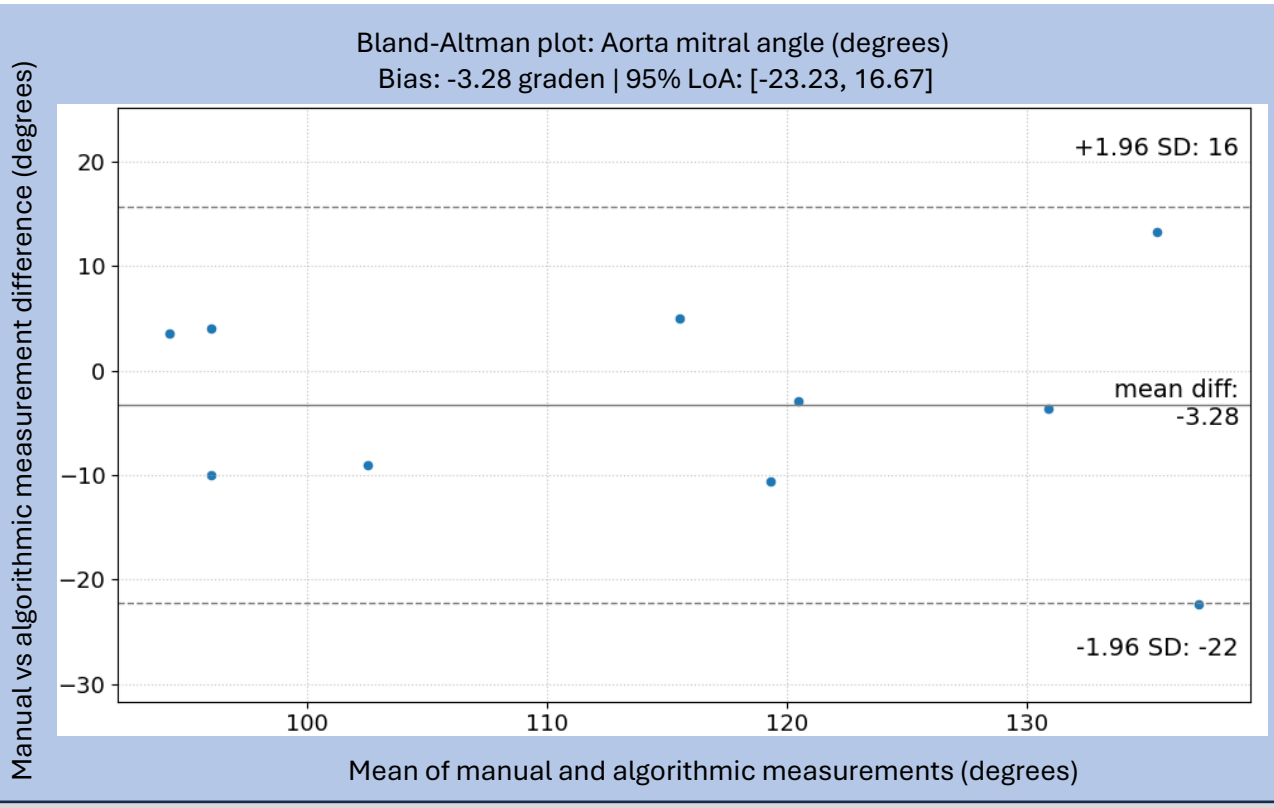


Figure 16. Bland–Altman plot showing agreement between manual and algorithmic **aortic–mitral angle** measurements. Ten measurements are included. The mean difference between automated and clinical measurements is small, indicating close agreement.

In routine clinical practice, the GE ultrasound system and IntelliSpace do not provide a dedicated tool for direct angle measurement. Instead, angles are typically estimated by manually drawing two lines and visually approximating the angle between them. For this study, the clinical reference measurements were obtained by overlaying an external angle measurement tool onto the images. As a result, the clinical reference itself is subject to measurement uncertainty.

Given these limitations, differences of up to **10 degrees** between manual and automated measurements may already be considered clinically acceptable

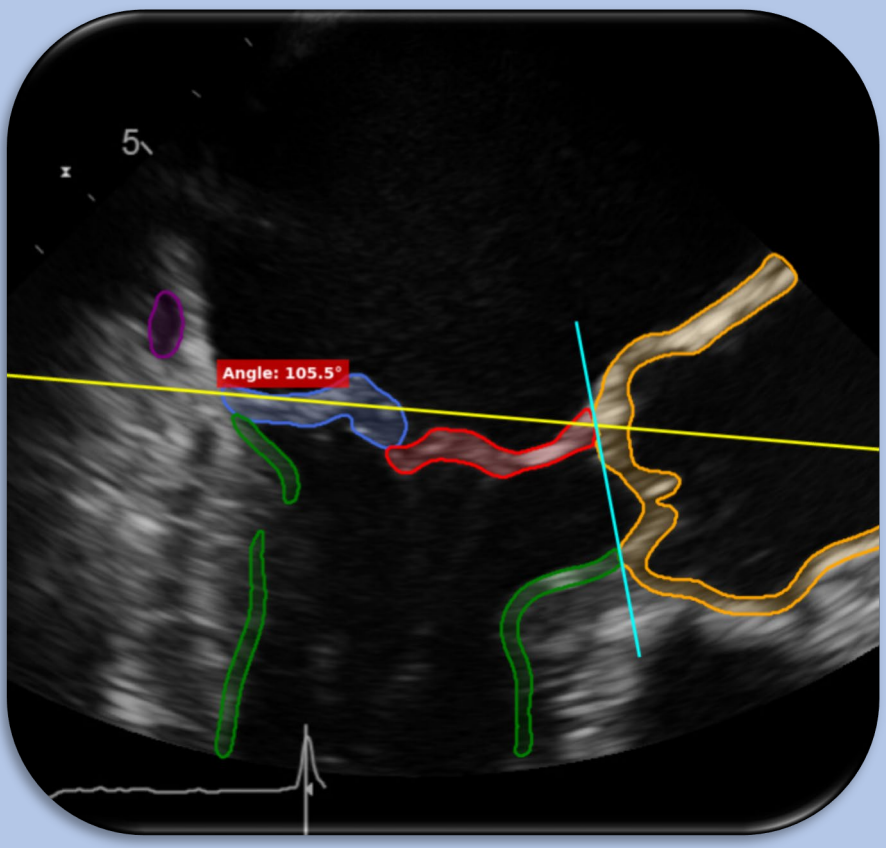


Figure 17. Example of automated **aortic–mitral angle** measurement for one patient. In this case, the difference between the algorithmic and clinical measurement was **9 degrees**. The figure shows accurate segmentation and anatomically consistent angle estimation by the algorithm. A small degree of overestimation may be present, as the mitral plane could have been positioned slightly higher on the anterior leaflet (indicated in red); however, this would not be expected to account for a discrepancy as large as 9 degrees. This example further illustrates the inherent **subjectivity and variability** associated with clinical angle measurements.

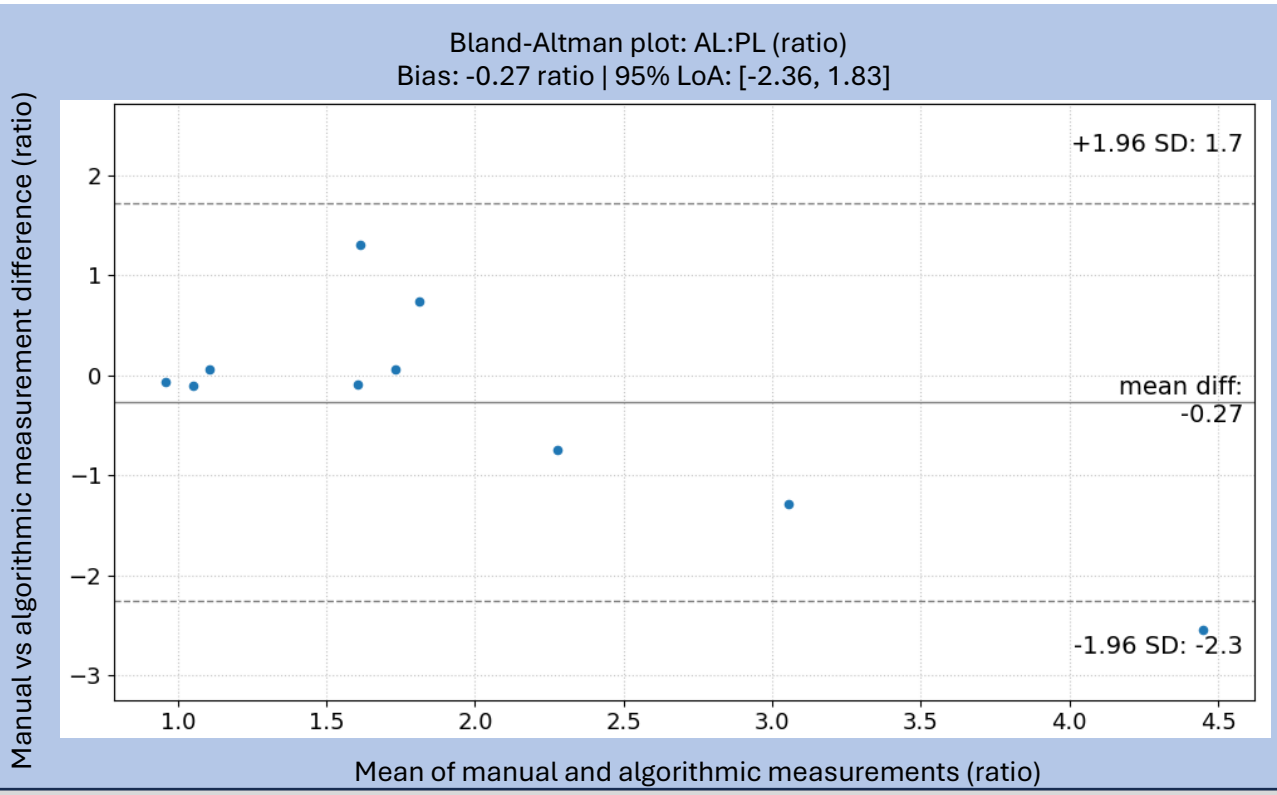


Figure 18. Bland–Altman plot showing agreement between manual and algorithmic **anterior–posterior leaflet (AL:PL) ratio** measurements. Ten measurements are included. Two distinct clusters are observed: one with minimal measurement error and one with substantially larger deviations. This pattern reflects the high sensitivity of the AL:PL ratio to segmentation quality of the mitral valve. Small inaccuracies in leaflet segmentation can result in disproportionately large errors in the AL:PL ratio.

The AL:PL ratio itself is a sensitive parameter, where a difference of approximately **0.2** may already be considered clinically relevant. This assessment is based on the authors' judgment and discussions with expert E.T. However, it is likely that the true inter-observer variability for this measurement exceeds this threshold.

Manual measurement of the AL:PL ratio is inherently challenging, as it requires estimating the lengths of the anterior and posterior leaflets within the mitral plane. This involves projecting leaflet vectors and is difficult to perform precisely using the GE ultrasound system or IntelliSpace. Additionally, accurately identifying the transition between leaflet tissue and ventricular myocardium introduces further uncertainty in manual measurements.

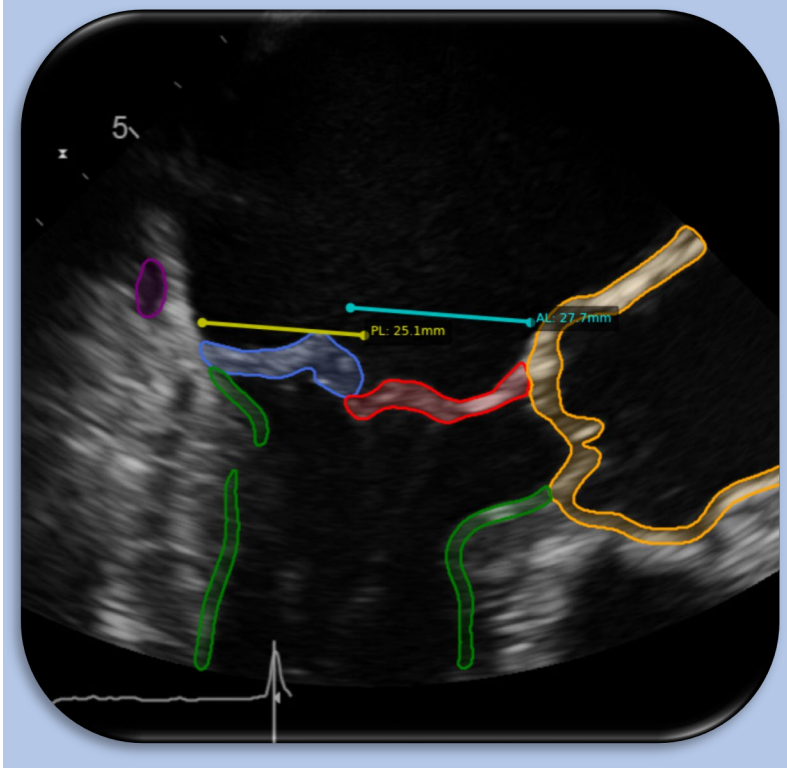


Figure 19. Example of automated **anterior–posterior leaflet ratio** measurement for one patient. In this case, the difference between the algorithmic and clinical measurement was 0.1, indicating close agreement. The figure shows the measured lengths of the anterior and posterior leaflets, calculated by projecting leaflet vectors onto the mitral valve plane (dot product), resulting in an anatomically consistent and accurate ratio estimation.

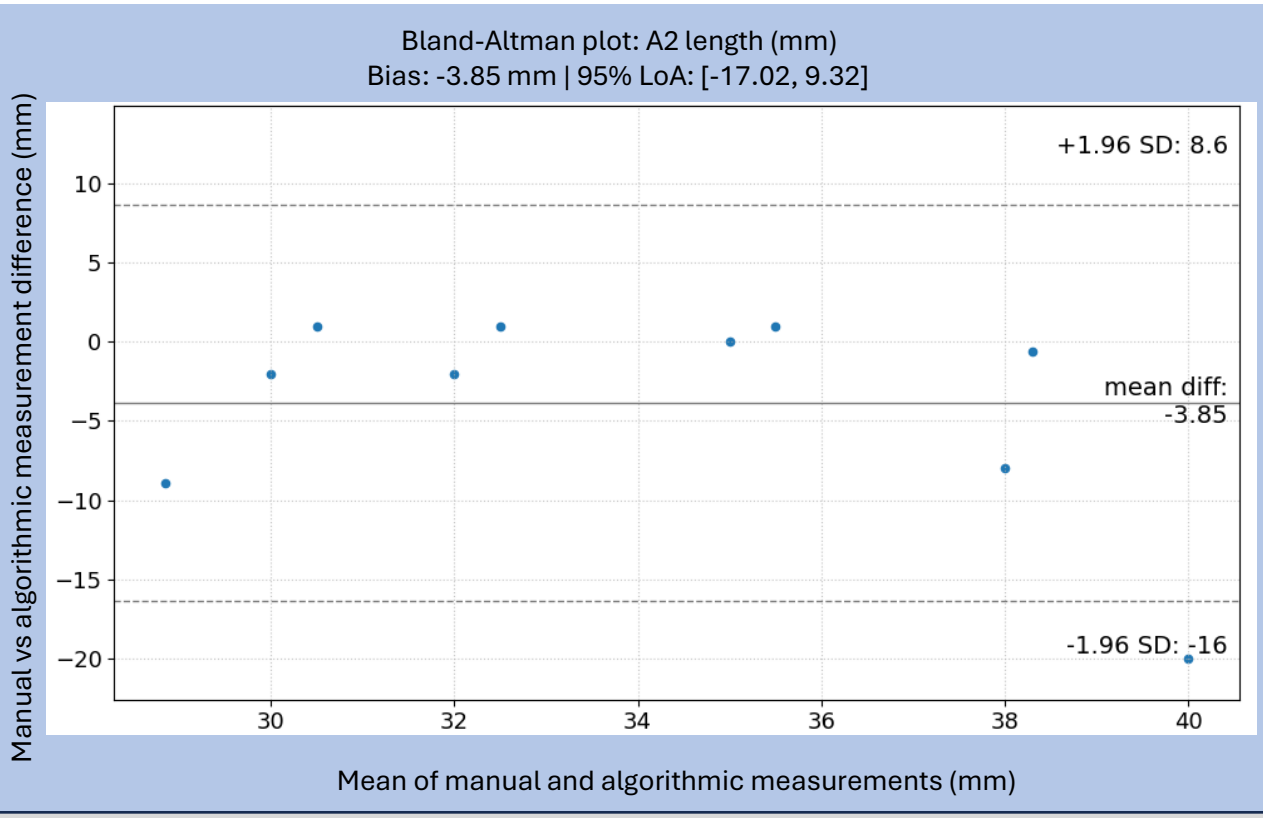


Figure 20. Bland–Altman plot showing agreement between manual and algorithmic **A2 leaflet length** measurements. Ten measurements are included. In most cases, differences between manual and automated measurements are within approximately 2.5 mm. However, similar to the anterior–posterior leaflet ratio, A2 length is highly sensitive to segmentation quality. When segmentation is suboptimal, small errors in leaflet delineation can result in disproportionately large measurement deviations.

Both the clinical reference and the algorithm determine A2 length in the frame where the A2 leaflet appears longest. Consequently, if a single frame contains a segmentation error that artificially elongates the A2 leaflet, this frame may be selected, resulting in a substantial measurement discrepancy. This mechanism explains the presence of a pronounced outlier with a difference of approximately 20 mm.

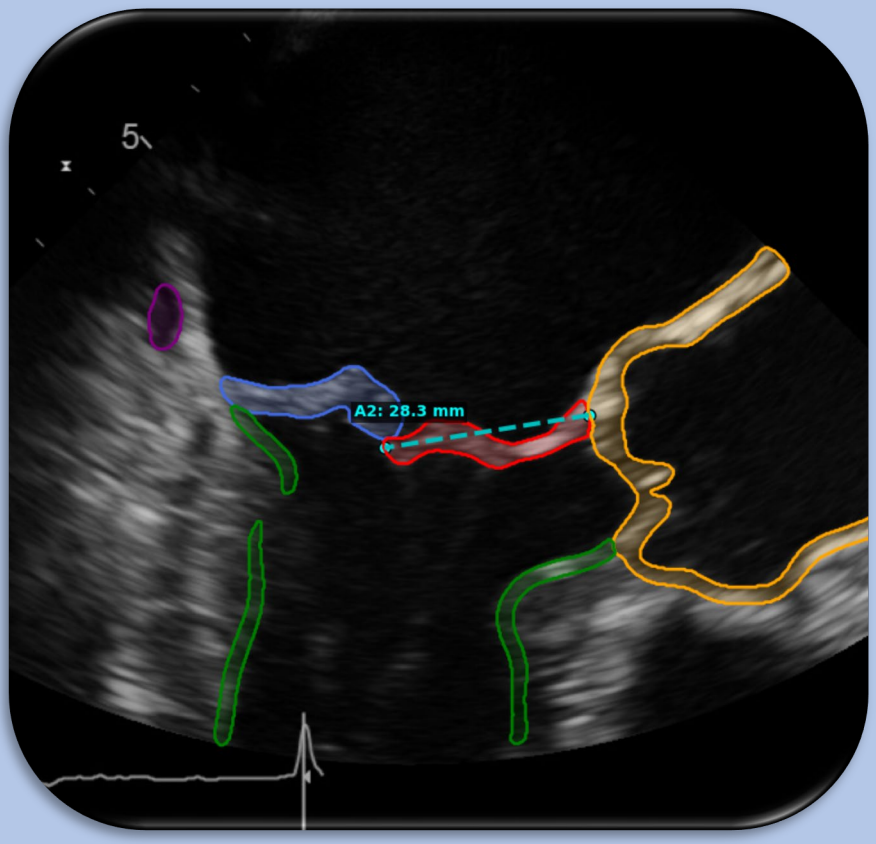


Figure 21. Example of automated **A2 leaflet length** measurement. In this case, the difference between the algorithmic and clinical measurement was **2 mm**. The figure demonstrates accurate segmentation and anatomically consistent leaflet delineation by the algorithm, suggesting that the observed discrepancy may reflect variability in clinical measurement rather than true algorithmic error.

Mitral Valve Prolapse Detection Performance

Mitral valve prolapse detection was evaluated qualitatively on the same 10 LAX cases. The operative report was used as the reference ground truth. The algorithm identified prolapse by detecting leaflet excursion above the mitral annular plane during systole. Examples of prolapse detections are presented in figure 23 and 26. Given the limited sample size (N=10), agreement analyses should be interpreted descriptively rather than as definitive performance estimates.

An overall accuracy of 90% was observed, with one false-positive case. Given the limited sample size, these results are reported descriptively and are intended to demonstrate technical feasibility rather than diagnostic performance. The false positive was due to a wrong placement of the hinge point algorithm. The hinge point placement logic follows a spatiotemporal approach designed to isolate the stationary base of the mitral valve leaflets from their mobile tips. The algorithm operates through the following steps:

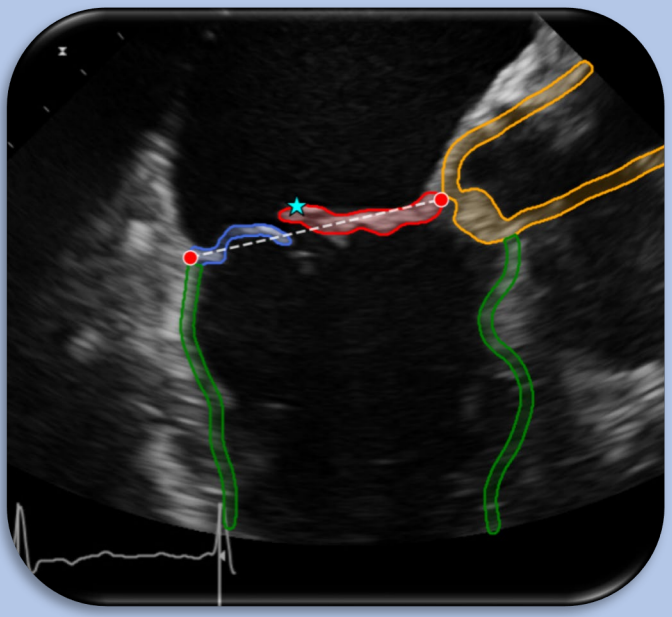
1. Diameter-based Extraction and Initial Sorting: For each frame, the algorithm identifies the two coordinates within the leaflet mask (for anterior and posterior) that are farthest apart, representing the long axis of the leaflet. To prevent the labels from flipping randomly between frames, an initial "A or B" assignment is generated: each new pair of points is sorted based on which point is closest to the hinge point identified in the previous frame. This serves as the primary prevention measure against label swapping.

2. Kinematic Anchoring: Since the hinge is physiologically fixed to the annulus, it should exhibit significantly less total displacement than the tip across the cardiac cycle. The algorithm calculates the cumulative Euclidean distance traveled by both sorted endpoint lists across all frames. The list with the minimum total displacement is prioritized as the stable hinge point, while the high-displacement list is assigned as the leaflet tip.

The observed false-positive result was caused by a failure during the initial sorting phase. In this specific case, one frame contained a very small leaflet segment due to signal dropout. Because the leaflet area was reduced to a nearly circular cluster of pixels, the distance between the two "farthest" points was negligible.

Step 1, which relies on finding the point closest to the previous frame's hinge, became unreliable because the anatomical tip was mathematically closer to the previous hinge location. This caused an incorrect assignment in that single frame, which then propagated through the kinematic anchoring phase. The resulting "flipped" hinge point tilted the annulus plane into the ventricle, causing the algorithm to incorrectly classify the leaflet tip as being above the annulus, thus triggering a false-positive prolapse detection.

Mitral valve prolapse detection preview



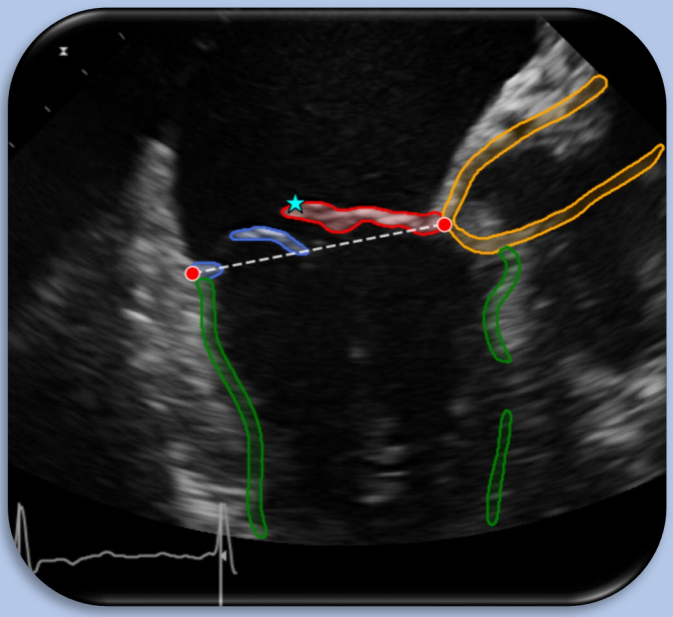
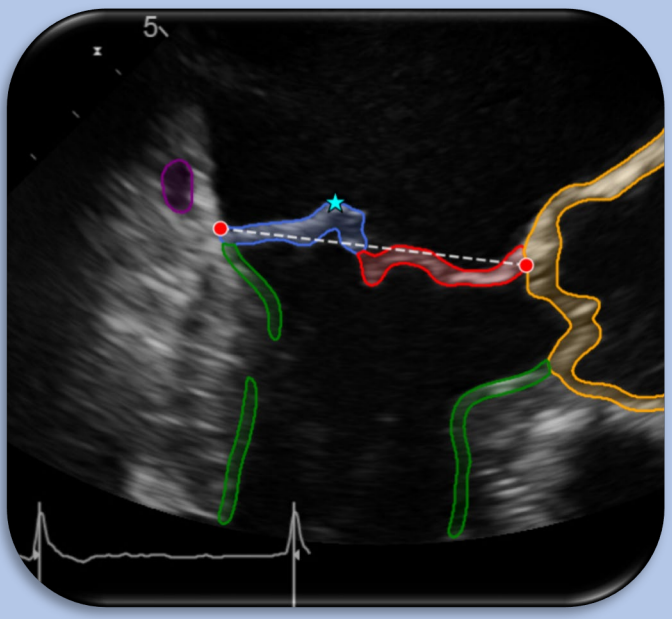


Figure 22 and 23. represent a patient with an A2 prolapse of the mitral valve. In this case, the prolapse was successfully detected by the algorithm using intraoperative echocardiography. These automated findings were consistent with the intraoperative surgical report, which described a significant prolapse of the A2 segment extending slightly into the A3 segment.



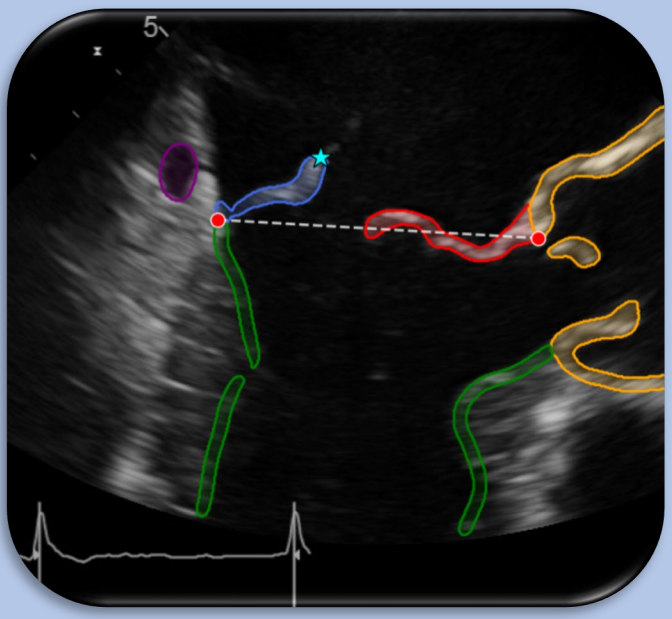


Figure 24 and 25. represent a patient with a P2 prolapse of the mitral valve. In this case, the prolapse was successfully identified by the algorithm using intraoperative echocardiography. These automated findings were consistent with the intraoperative surgical report

Discussion

This study explores the feasibility of automated, standardized interpretation of 2D TEE for mitral valve surgery by addressing multiple key components, including view classification, scallop-level anatomical interpretation, segmentation, quantitative measurement extraction, and pathology detection. Rather than presenting a fully integrated system, the focus is on evaluating these components individually under realistic clinical and data constraints. The findings provide insight into which tasks are currently well suited for automation and which remain limited by data availability, anatomical ambiguity, and reference standard variability.

In the existing literature, automated echocardiographic analysis has primarily focused often using transthoracic echocardiography. Studies addressing intraoperative 2D TEE are comparatively scarce and typically concentrate on isolated tasks such as global valve segmentation (13). In contrast, the present work specifically targets scallop-level anatomical interpretation in intraoperative 2D TEE, using real-world images characterized by variable quality and anatomical presentation. By evaluating multiple key components separately within a single clinical context, this study provides methodological insights that complement existing approaches and address a less explored but clinically relevant imaging setting.

Classification performance revealed a clear distinction between view-level and scallop-level tasks. While standard mid-esophageal view classification proved highly reliable, scallop-level classification performance was modest. This limitation appears to be driven primarily by severe class imbalance within the dataset, particularly for LAX and commissural views, rather than by architectural limitations. These results suggest that scallop-level classification is feasible in principle but currently constrained by insufficient and imbalanced training data.

Segmentation performance was consistently moderate for anatomically relevant regions. Training on individual frames augmented with limited temporal context not as effective as expected, indicating that this strategy still has limitations. This limited improvement can be contributed to the different framerate the echo sequences. Therefore it is recommended that in future experiments the a form of temporal resampling is used. Overall, the results indicate that reliable segmentation of key structures from 2D TEE is achievable using this approach.

Both the nnU-Net and ResNet architectures can benefit from explicit temporal modeling when frame timing information is available. Frame-to-frame timing can be extracted directly from DICOM metadata, enabling models to account for non-uniform temporal spacing between frames. This opens the possibility of temporal resampling, whereby sequences are normalized to a consistent temporal resolution, thereby reducing variability in frame timing and facilitating more stable and effective model training. A method for doing this is RIFE (Real-Time Intermediate Flow Estimation) (18), which uses a deep neural network to estimate intermediate motion via optical flow. By computing motion vectors between existing frames, RIFE warps pixels to synthesize anatomically plausible intermediate timestamps. This ensures a uniform temporal flow, allowing models to better capture the rapid dynamics of cardiac structures (18).

A key methodological direction for future work is the tighter coupling of segmentation and classification by explicitly encoding scallop identity within the segmentation task itself. Rather than treating scallop identification as a separate classification problem, the segmentation network could be extended to assign a distinct class label to each individual scallop, effectively combining anatomical localization and scallop identification in a single model.

Preliminary experiments exploring this approach were conducted but did not yet yield reliable performance, likely due to the limited size and strong class imbalance of the current dataset. In particular, several scallop classes were represented by too few examples to support stable learning. With larger and more balanced datasets, this unified segmentation-based strategy has the potential to improve scallop-level discrimination by leveraging spatial continuity and anatomical context directly within the model architecture.

Beyond 2D imaging, three-dimensional echocardiography represents an important future extension. Many geometric measurements, such as A2 leaflet length, are fundamentally limited in 2D by plane orientation. Even small rotations of the imaging plane relative to the true principal axis of a scallop can result in substantial over- or underestimation of length. In 3D echocardiography, the principal axis can be identified directly in three-dimensional space, eliminating this source of error. As such, measurements derived from 3D data are theoretically more robust and anatomically faithful, though they come with increased acquisition and processing complexity (19).

The automated measurement results in this study were encouraging and, in some cases, appeared at least as consistent as clinical reference measurements. Nevertheless, discrepancies remained, and these should not be interpreted solely as algorithmic error. Clinical measurements themselves are subject to high interobserver variability (20-22). To better evaluate measurement quality in future work, a different evaluation paradigm is warranted. Rather than relying exclusively on numerical agreement, imaging experts could be asked to visually assess and rate both automated and clinical measurements using a Likert scale. Such an approach would better capture perceived anatomical correctness and clinical usefulness.

A complementary approach to evaluate measurement performance would be to assess agreement relative to inter- and intra-observer variability, thereby determining whether algorithmic deviations fall within the range of human measurement uncertainty. In addition, the use of consensus-based reference measurements, derived from multiple expert annotations, could provide a more robust comparison framework than single-expert ground truth. Finally, a risk stratification analysis could be performed to assess whether automated and clinical measurements lead to consistent patient risk categorization.

A further limitation of the current measurement approach lies in the estimation of scallop principal axes. At present, the principal axis is approximated using the two most distant points on the segmented scallop, which can introduce rotational bias if these points are not aligned with the true anatomical axis. More robust alternatives include skeleton-based centerline extraction, principal component analysis of the scallop contour, or curvature-based axis estimation (23). Implementing such methods is expected to reduce systematic measurement error.

For end-systolic frame identification, the ensemble-based approach demonstrated robust behavior, though one expert component currently emphasizes maximal mitral valve closure rather than the clinically relevant moment immediately preceding valve opening. This results in systematic selection of frames one to two frames later than ideal in some cases. A simple temporal correction could compensate for this behavior and further improve accuracy.

However, beyond algorithmic performance, a more fundamental limitation concerns the conceptual definition of end-systole itself. End-systole represents a short physiological phase occurring between aortic valve closure and mitral valve opening rather than a single independently verifiable timestamp. Previous work has shown that small variations in the definition of end-systole (on the order of a few frames) can substantially affect derived measurements such as strain, underscoring that timing definitions are not universally standardized in echocardiographic analysis (24). At present,

there is no universally defined temporal window that specifies the exact boundaries within which measurements should be obtained.

Although clinicians often perceive that there is one correct frame, this likely reflects selection within a narrow interval of minimal morphological change rather than identification of a uniquely defined physiological instant. Expressing timing differences in milliseconds does not resolve this ambiguity, as no gold-standard continuous reference exists against which accuracy can be benchmarked. Moreover, the morphological consequences of small temporal shifts depend on local deformation dynamics and may differ substantially near the transition from maximal closure to early reopening.

In practice, a wide range of strategies has been described for defining end-systole. These include ECG-based surrogates such as identification of the end of the T-wave, automated analysis of segmental or global deformation signals, Doppler-derived velocity traces, and manual annotation of aortic valve closure based on grayscale or Doppler imaging (25-28). Importantly, the specific implementation of these algorithms is often not transparently documented, and their behavior must frequently be inferred from software output or clarified through communication with vendors. Consequently, it remains unclear which approaches are physiologically most valid and whether they ensure reproducible measurements across different clinical conditions (24).

Future work may therefore benefit from moving beyond rigid phase labeling toward continuous deformation analysis, enabling measurement windows to be defined based on physiologically meaningful signal characteristics rather than assumed discrete phase boundaries.

Standardization remains a broader challenge. Measurements such as the C-sept distance are performed differently across centers and even among individual clinicians, with variation in both timing (end-diastole versus end-systole) and anatomical landmarks. Currently, there is no consensus on optimal definitions for many SAM-related parameters (29, 30). Automated, standardized measurement methods offer a pathway toward resolving this ambiguity. By enabling consistent analysis of large retrospective datasets linked to surgical outcomes, such tools could help identify which definitions are most predictive and clinically meaningful.

The potential impact of standardized measurements extends beyond individual cases. Once measurements are defined and extracted consistently, large-scale outcome analyses become feasible, opening the door to new research into risk stratification, surgical planning, and patient-specific decision support.

Although the methods presented here are not intended to replace clinical decision-making, the results illustrate how automation may contribute to greater consistency and objectivity in intraoperative echocardiographic assessment. In time-critical environments such as the operating room, automated support for view recognition, anatomical orientation, and standardized measurements could be particularly valuable for less experienced operators or in lower-volume centers. Furthermore, consistent and reproducible extraction of anatomical and geometric parameters may enable large-scale retrospective analyses linking imaging features to surgical outcomes, which is currently difficult to achieve using manual measurements alone.

The component-based design of this study reflects a deliberate choice to develop and evaluate individual algorithms independently before considering integration into a unified system. In the context of artificial intelligence applications in the operating room, transparency, interpretability, and clear delineation of limitations are essential. By explicitly analyzing the performance and failure modes of individual components, this work clarifies where automated methods are reliable. This

stepwise approach aligns with current recommendations for the responsible development and clinical validation of AI-driven decision support systems.

Finally, future phases of this work will focus on expanding the dataset through multi-center collaboration. Agreements are already in place with three external centers, enabling inclusion of data from different institutions, imaging systems, and operator practices. This expansion is essential for improving generalizability, addressing class imbalance, and enabling validation across diverse clinical settings. Additional measurements and extension toward aortic imaging are also planned, further broadening the clinical scope of this work.

References

1. Vahanian A, Beyersdorf F, Praz F, Milojevic M, Baldus S, Bauersachs J, et al. 2021 ESC/EACTS Guidelines for the management of valvular heart disease. Eur Heart J. 2022;43(7):561-632.
2. Olsthoorn JR, Heuts S, Streukens SAF, Hermans SMM, Maessen JG, Sardari Nia P. Unexpected prolapse of the anterior leaflet during saline testing in mitral valve repair. Eur J Cardiothorac Surg. 2019;55(3):552-8.
3. Otterstad JE, Froeland G, St John Sutton M, Holme I. Accuracy and reproducibility of biplane two-dimensional echocardiographic measurements of left ventricular dimensions and function. Eur Heart J. 1997;18(3):507-13.
4. Chen J, Li W, Xiang M. Burden of valvular heart disease, 1990-2017: Results from the Global Burden of Disease Study 2017. J Glob Health. 2020;10(2):020404.
5. Drake DH, Zimmerman KG, Hepner AM, Nichols CD. Echo-guided mitral repair. Circ Cardiovasc Imaging. 2014;7(1):132-41.
6. Gillinov AM, Cosgrove DM, Blackstone EH, Diaz R, Arnold JH, Lytle BW, et al. Durability of mitral valve repair for degenerative disease. J Thorac Cardiovasc Surg. 1998;116(5):734-43.
7. Nicoara A, Skubas N, Ad N, Finley A, Hahn RT, Mahmood F, et al. Guidelines for the Use of Transesophageal Echocardiography to Assist with Surgical Decision-Making in the Operating Room: A Surgery-Based Approach: From the American Society of Echocardiography in Collaboration with the Society of Cardiovascular Anesthesiologists and the Society of Thoracic Surgeons. J Am Soc Echocardiogr. 2020;33(6):692-734.
8. Garbi M, Monaghan MJ. Quantitative mitral valve anatomy and pathology. Echo Res Pract. 2015;2(3):R63-72.
9. Bakhtiary F, Salamate S, Eghbalzadeh K, El-Sayed Ahmad A. Endoscopic micro-invasive cardiac surgery: State-of-the-art. Turk Gogus Kalp Damar Cerrahisi Derg. 2024;32(4):355-66.
10. Zacharias J, Pitsis A, Glauber M, Solinas M, Kempfert J, Castillo-Sang M, et al. Endoscopic cardiac surgery: the path less taken. Lancet. 2024;404(10463):1624-6.
11. Mathis MR, Yule S, Wu X, Dias RD, Janda AM, Krein SL, et al. The impact of team familiarity on intra and postoperative cardiac surgical outcomes. Surgery. 2021;170(4):1031-8.
12. American Society of A, Society of Cardiovascular Anesthesiologists Task Force on Transesophageal E. Practice guidelines for perioperative transesophageal echocardiography. An updated report by the American Society of Anesthesiologists and the Society of Cardiovascular Anesthesiologists Task Force on Transesophageal Echocardiography. Anesthesiology. 2010;112(5):1084-96.
13. Chen J, Li H, He G, Yao F, Lai L, Yao J, et al. Automatic 3D mitral valve leaflet segmentation and validation of quantitative measurement. Biomedical Signal Processing and Control. 2023;79.
14. Pouch AM, Wang H, Takabe M, Jackson BM, Gorman JH, 3rd, Gorman RC, et al. Fully automatic segmentation of the mitral leaflets in 3D transesophageal echocardiographic images using multi-atlas joint label fusion and deformable medial modeling. Med Image Anal. 2014;18(1):118-29.
15. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. NPJ Digit Med. 2018;1.
16. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, et al. Fully Automated Echocardiogram Interpretation in Clinical Practice. Circulation. 2018;138(16):1623-35.
17. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203-11.
18. Huang Z. Real-Time Intermediate Flow Estimation for Video Frame Interpolation. ECCV. 2020.
19. Lang RM, Badano LP, Tsang W, Adams DH, Agricola E, Buck T, et al. EAE/ASE recommendations for image acquisition and display using three-dimensional echocardiography. J Am Soc Echocardiogr. 2012;25(1):3-46.
20. Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American

Society of Echocardiography and the European Association of Cardiovascular Imaging. J Am Soc Echocardiogr. 2015;28(1):1-39 e14.
21. Maron MS, Olivotto I, Zenovich AG, Link MS, Pandian NG, Kuvin JT, et al. Hypertrophic cardiomyopathy is predominantly a disease of left ventricular outflow tract obstruction. Circulation. 2006;114(21):2232-9.
22. Spirito P, Maron BJ, Rosing DR. Morphologic determinants of hemodynamic state after ventricular septal myotomy-myectomy in patients with obstructive hypertrophic cardiomyopathy: M mode and two-dimensional echocardiographic assessment. Circulation. 1984;70(6):984-95.
23. Jiang W, Xu K, Cheng Z-Q, Martin RR, Dang G. Curve skeleton extraction by coupled graph contraction and surface clustering. Graphical Models. 2013;75(3):137-48.
24. Mada RO, Lysyansky P, Daraban AM, Duchenne J, Voigt JU. How to define end-diastole and end-systole?: Impact of timing on strain measurements. JACC Cardiovasc Imaging. 2015;8(2):148-57.
25. Aase SA, Torp H, Stoylen A. Aortic valve closure: relation to tissue velocities by Doppler and speckle tracking in normal subjects. Eur J Echocardiogr. 2008;9(4):555-9.
26. Aase SA, Bjork-Ingul C, Thorstensen A, Torp H, Stoylen A. Aortic valve closure: relation to tissue velocities by Doppler and speckle tracking in patients with infarction and at high heart rates. Echocardiography. 2010;27(4):363-9.
27. Lyseggen E, Vartdal T, Remme EW, Helle-Valle T, Pettersen E, Opdahl A, et al. A novel echocardiographic marker of end systole in the ischemic left ventricle: "tug of war" sign. Am J Physiol Heart Circ Physiol. 2009;296(3):H645-54.
28. Voigt JU, Pedrizzetti G, Lysyansky P, Marwick TH, Houle H, Baumann R, et al. Definitions for a common standard for 2D speckle tracking echocardiography: consensus document of the EACVI/ASE/Industry Task Force to standardize deformation imaging. Eur Heart J Cardiovasc Imaging. 2015;16(1):1-11.
29. Ommen SR, Mital S, Burke MA, Day SM, Deswal A, Elliott P, et al. 2020 AHA/ACC Guideline for the Diagnosis and Treatment of Patients With Hypertrophic Cardiomyopathy: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation. 2020;142(25):e558-e631.
30. Authors/Task Force m, Elliott PM, Anastasakis A, Borger MA, Borggrefe M, Cecchi F, et al. 2014 ESC Guidelines on diagnosis and management of hypertrophic cardiomyopathy: the Task Force for the Diagnosis and Management of Hypertrophic Cardiomyopathy of the European Society of Cardiology (ESC). Eur Heart J. 2014;35(39):2733-79.

Appendix

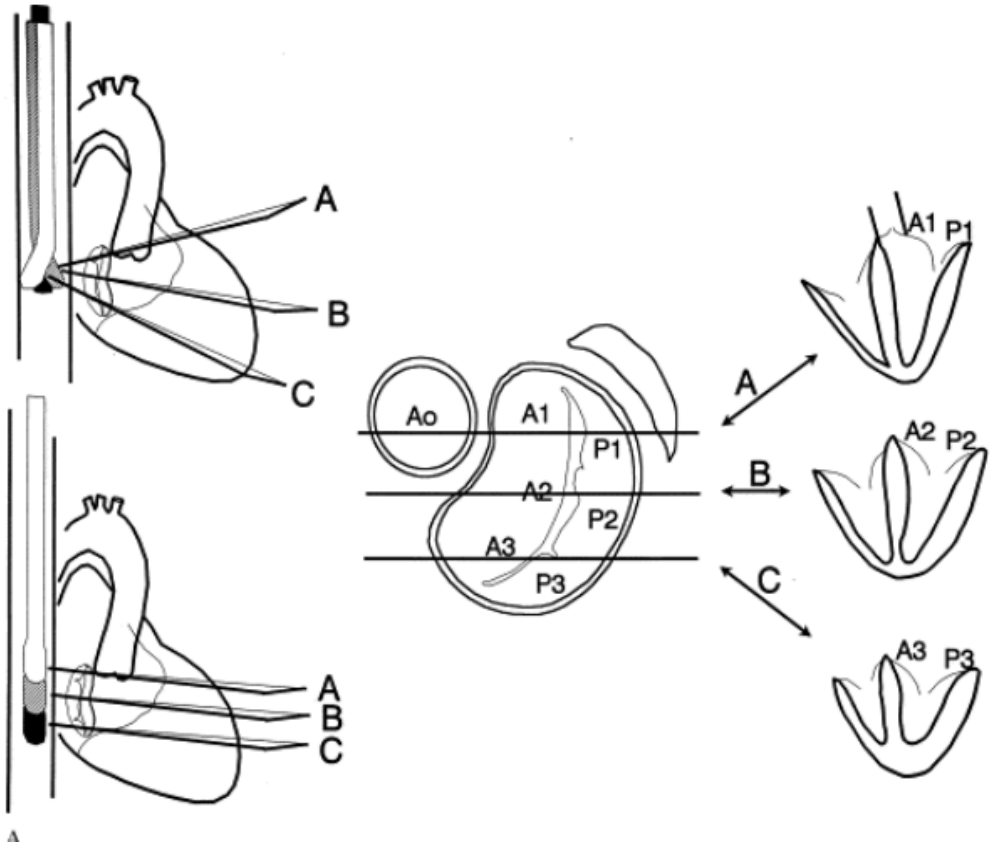
Protocol summary

Mid-Esophageal Four-Chamber View

1. **Probe angle:** Between **0–20°** → assume a four-chamber view.
2. **Confirm landmarks:** Four chambers visible.
3. **Scallops visible (if confirmed):**
    - **A2–P2** when neither the aorta nor the coronary sinus is visible,
    - **A1–P1** when the aorta is visible,
    - **A3–P3** when the coronary sinus is visible.

Structures to Label:

- Left atrium with septum
- Left ventricle with septum
- Anterior leaflet of the mitral valve
- Posterior leaflet of the mitral valve
- Tricuspid valve
- Right ventricle with no overlapping labels
- Aortic valve
- Coronary sinus

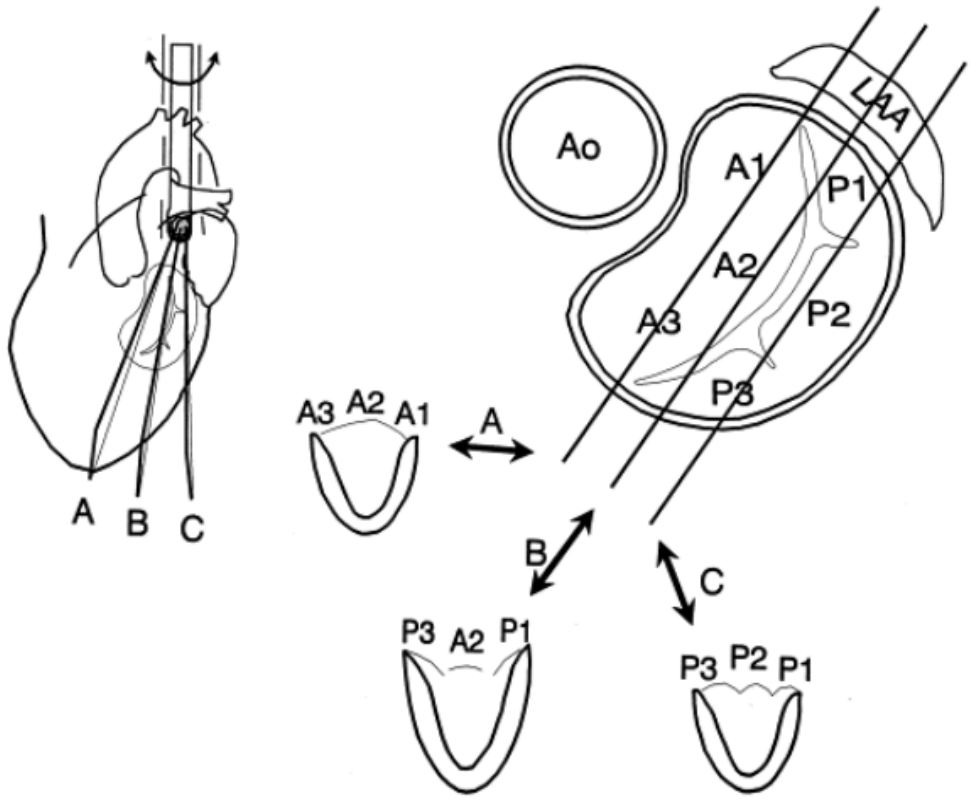


Mid-Esophageal Commissural View

1. When the probe angle is between **45–75°**, assume a commissural view of the mitral valve.
2. Confirm the view by checking for:
    - The **left atrium**, **left ventricle**, and **mitral valve** in plane,
    - Both **papillary muscles** visible (note: in end-diastole they may be difficult to visualize, and A2 may also not be visible),
    - The **coronary sinus** and **left atrial appendage** included in the image.
3. If confirmed, the scallops typically visible are:
    - **P3 – (A3) A2 (A1) – P1** when the valve appears segmented,
    - If instead the valve appears as a continuous structure, the view may be oriented differently, showing either **P3–P2–P1** or **A3–A2–A1**.

Structures to Label:

- Left atrium
- Left ventricle
- Anterior leaflet of the mitral valve
- Posterior leaflet of the mitral valve
- Coronary sinus
- Left aterial appendage



B

Mid-Esophageal Two-Chamber View

1. When the probe angle is between **80–100°**, assume a mid-esophageal two-chamber view.
2. Confirm the view by checking for:
    - Only **two chambers** visible,
    - A **large leaflet** on the right side of the image and a **smaller leaflet** on the left,
    - The **left atrial appendage** visible on the right side of the image (more prominent than in the commissural view),
    - The **coronary sinus** visible on the left side of the image.
3. If confirmed, the scallops **P3 – A3, A2, and A1** are visible.

Structures to Label:

- Left atrium
- Left ventricle
- Anterior leaflet of the mitral valve
- Posterior leaflet of the mitral valve
- Coronary sinus
- Left atrial appendage

Mid-Esophageal Long-Axis (LAX) View

1. When the probe angle is between **120–160°**, assume an LAX view.
2. Confirm the view by checking for:
    - Two chambers,
    - The aortic valve with the left ventricular outflow tract,
    - A portion of the right ventricle visible.
3. If the aortic valve is in the imaging plane and **no papillary muscles** are visible, assume the view corresponds to **P2-A2**.
4. If the **coronary sinus** or the **posteromedial commissure** is visible, assume **P3-A3**.
5. If the **anterolateral commissure** or the **left atrial appendage** is visible, assume **P1-A1**.

Structures to Label:

- Left Atrium
- Left Ventricle
- Aortic valve
- Part of the right ventricle
- mitral valve anterior
- mitral valve posterior
- coronary sinus
- left atrial appendage