Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 22 No. 7 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v22i07.60655

PAPER

Transformer-Driven Cardiac Image Segmentation: A Swin-EchoNet Approach for Two-Chamber Echocardiography

A. Shamla Beevi[1] (✉),
Amit Nagarkoti[2],
Saidalavi Kalady[1]

[1]National Institute of Technology Calicut, Calicut, Kozhikode, India

[2]Bharat Electronics Limited, Bengaluru, Karnataka, India

shamla_p180105cs@nitc.ac.in

ABSTRACT

Echocardiographic segmentation involves the automated delineation of cardiac chambers from ultrasound images to enable quantitative assessment of cardiac structure and function, which is essential for diagnosing and managing cardiovascular diseases. However, existing CNN-based segmentation models often struggle to preserve fine boundary details and global spatial context due to speckle noise, low contrast, and imaging artifacts inherent to echocardiography. To overcome these limitations, this study proposes Swin-EchoNet, a transformer-based segmentation framework that combines shifted-window self-attention with hierarchical feature extraction to capture both local texture information and long-range dependencies. The proposed approach was trained and evaluated on two distinct datasets: the CAMUS dataset comprising 1,800 echocardiographic images and a clinical Aster-MIMS dataset consisting of 1,550 frames collected from hospital studies. On the CAMUS dataset, Swin-EchoNet achieved a mean Dice score of 0.951 ± 0.227, while on the Aster-MIMS dataset, it obtained an average Dice coefficient of 0.9738 with a mean Intersection over Union (IoU) of 0.7263. Comparative evaluations demonstrate that Swin-EchoNet outperforms U-Net, ResU-Net, and Attention U-Net, highlighting its robustness and clinical applicability.

KEYWORDS

Swin transformer, deep learning, echocardiography, cardiac chamber segmentation, two-chamber (2CH) view, vision transformer, attention mechanism, medical image segmentation

1 INTRODUCTION

Cardiovascular diseases are the leading cause of illness and death worldwide, accounting for nearly one-third of all global deaths [1]. Significant advancements in cardiovascular research and clinical practice have been driven by improvements in diagnostic imaging and data analysis techniques. Accurate assessment of cardiac structure and function is essential for early diagnosis and effective treatment planning.

Beevi, A. S., Nagarkoti, A., Kalady, S. (2026). Transformer-Driven Cardiac Image Segmentation: A Swin-EchoNet Approach for Two-Chamber Echocardiography. *International Journal of Online and Biomedical Engineering (iJOE)*, 22(7), pp. 126–149. https://doi.org/10.3991/ijoe.v22i07.60655

Article submitted 2026-01-19. Revision uploaded 2026-04-09. Final acceptance 2026-04-17.

The human heart consists of four chambers, including the left atrium (LA), right atrium, right ventricle, and left ventricle (LV), which function together to maintain blood circulation [2], [3]. Structural or functional abnormalities in these chambers are indicative of severe conditions such as cardiomyopathies, myocardial infarction, and valvular diseases, emphasizing the need for precise analysis.

Modern cardiac assessment relies on imaging modalities such as echocardiography, magnetic resonance imaging (MRI), and computed tomography (CT), which enable non-invasive visualization of cardiac structures [4]. Although MRI and CT provide high-resolution images, their cost and limited accessibility restrict routine clinical use. In contrast, echocardiography is widely adopted due to its portability, real-time imaging capability, and absence of ionizing radiation [5]. Among various views, the two-chamber (2CH) view is particularly important for analyzing the left heart structures, including the LA and LV. Accurate segmentation of these chambers facilitates the estimation of clinically significant parameters such as end-systolic volume, end-diastolic volume, and ejection fraction, which are essential for diagnosis and prognosis. Cardiac image segmentation plays a crucial role in quantitative analysis by delineating anatomically meaningful regions of the heart [6]. These segmented regions enable precise computation of myocardial mass, wall thickness, and chamber volumes. Figure 1 illustrates standard segmentation tasks across MRI, CT, and ultrasound modalities, highlighting the diversity of imaging conditions and anatomical targets [7]. Traditional segmentation methods, including active shape models, active appearance models, and atlas-based techniques, provide anatomically constrained results but often require manual intervention and prior knowledge. Furthermore, their performance degrades under challenging imaging conditions such as acoustic shadowing and low signal-to-noise ratios in echocardiography.

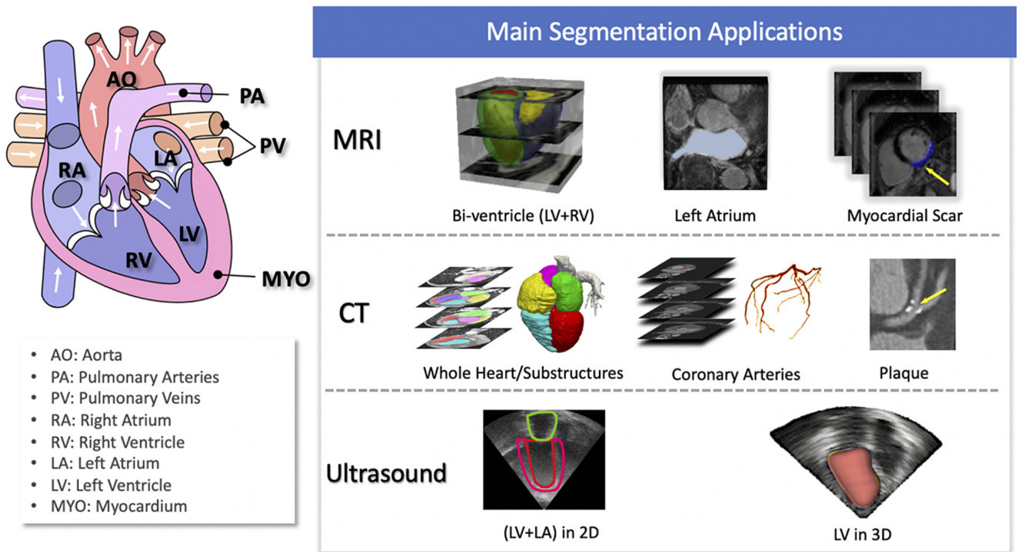

Fig. 1. Visualization of primary cardiac segmentation domains in MRI, CT, and ultrasound imaging [11]

With the advancement of deep learning (DL), particularly convolutional neural networks (CNNs) [8], automatic cardiac segmentation has significantly improved [9], [10]. CNN-based architectures, such as U-Net and its variants, have demonstrated strong performance due to their encoder–decoder structure and ability to capture spatial hierarchies. However, CNNs are inherently limited by their local receptive fields, restricting their ability to model global context and long-range dependencies, which are critical for accurate boundary delineation in complex ultrasound images. In this context, the proposed Swin-EchoNet framework employs a Swin Transformer-based architecture for 2CH segmentation in 2D echocardiography, utilizing hierarchical feature extraction and shifted-window attention to enhance spatial coherence and boundary precision under varying imaging conditions.

Despite the growing adoption of transformer-based approaches in medical image segmentation, their application to echocardiographic data remains limited due to challenges such as speckle noise, low contrast, and anatomical variability. Moreover, existing Swin-based models are primarily designed for general vision tasks and lack domain-specific adaptations required for ultrasound imaging. To address these limitations, the proposed framework incorporates ultrasound-oriented preprocessing, view-aware feature representation, and a boundary-aware composite loss function to improve segmentation performance in noisy clinical conditions. Additionally, validation on both a public benchmark dataset and a real-world clinical dataset ensures enhanced robustness and generalizability.

1.1 Research questions

To guide the development and evaluation of the proposed framework, the following research questions are formulated:

RQ1: Can transformer-based architectures improve segmentation accuracy in echocardiographic images compared to conventional CNN-based models, particularly under low-quality imaging conditions?

RQ2: How effectively does the proposed Swin-EchoNet model generalize across both benchmark datasets and real-world clinical data?

RQ3: To what extent do domain-specific enhancements, including preprocessing and boundary-aware loss functions, contribute to improved segmentation performance?

1.2 Objectives

The main objectives of the proposed model are

- To develop Swin-EchoNet, a ViT-based DL model for multi-class segmentation of 2D echocardiographic images for accurate chamber identification.
- To utilize a clinically acquired dataset comprising apical two-chamber (A2C) and four-chamber (A4C) views with manually annotated segmentation masks for model training and evaluation.
- To evaluate the effectiveness of the developed model using multiple performance metrics.

The rest of this paper is structured as follows: A review of cardiac image segmentation is given in Section 2. The methodology is shown in Section 3, and the results of the proposed model are explained in Section 4. Section 5 provides concluding observations and future research.

2 RELATED WORKS

Wu et al. [12] presented LV segmentation in TEE A4CV images using a UNeXt-based architecture on 3,000 images from 60 patients, achieving 88.60% DSC and 77.60% Intersection over Union (IoU) with 1.47M parameters and 141.73 ms run-time, though limited dataset diversity constrained generalization. Assadi et al. [13] addressed A4C segmentation in cine CMR using 814 subjects with external validation on 101 scans, reporting strong correlations (0.89–0.98) between automated and manual measurements, while slight underestimation of ventricular volumes and limited validation diversity remained concerns. Fatima et al. [14] utilized a Pix2Pix GAN framework on the CAMUS dataset, achieving Dice scores up to 0.961 (2CH) and 0.959 (4CH) with EF correlations of 0.95–0.98 but required separate training for views and showed limited real-time applicability. Wang et al. [15] integrated an attention-based GAN with transfer learning, yielding a DSC of 0.932, an IoU of 0.858, and AUC values of 0.993 (A4C) and 0.991 (A2C), though computational cost remained high. Ye et al. [16] introduced HFE-Mamba using 4,467 pediatric echocardiograms, improving boundary delineation and feature extraction, yet showing limited generalization to adult data and dependence on high-quality annotations.

Chao et al. [17] applied SAM on EchoNet-Dynamic (10,030 videos) with external validation datasets, achieving a Dice of 0.891 ± 0.040 and reducing annotation time significantly, though computational demand and protocol adaptability were limiting factors. Kang et al. [18] developed a residual-attention DNN on ICARUS TEE data (9 patients), achieving a DSC of 0.899 ± 0.017 and an IoU of 0.822 ± 0.026, with limitations due to the small sample size. Mortada et al. [19] combined YOLOv7 with U-Net on CAMUS, obtaining DSC values of 92.63% (LVepi), 85.59% (LVendo), and 87.57% (LA), though performance degraded on noisy images. Li et al. [20] introduced EchoEFNet using CAMUS and CMUEcho datasets, achieving a DSC of 0.934 and EF correlations of 0.854 and 0.916, but with complex architecture and limited pathology diversity. Cacao et al. [21] developed an unsupervised CNN model achieving Dice 0.8225 and IoU 0.7071, though reliability decreased under noisy conditions. Ding et al. [22] combined spatial and frequency-domain augmentation, achieving Dice scores of 89.48% (CAMUS) and 91.72% (EchoNet), but instability under extreme frequency changes persisted.

Naidoo et al. [23] explored self-supervised learning with the UnityLV-MultiX dataset, showing strong agreement with expert consensus using limited labeled data, though redundancy in large unlabeled datasets reduced performance. Chernyshov et al. [24] evaluated RV segmentation on 250 HUNT4 samples, achieving a bias of 0.8 ± 10.8% and –0.04 ± 0.54 cm, with challenges in thin boundary detection. Qian et al. [25] proposed a lightweight model with a DSC of 0.92714, 4472.55M FLOPs, and 28.96M parameters, though small-region segmentation remained challenging. Nemri et al. [26] applied Swin Transformer U-Net on EchoNet datasets, achieving 97% accuracy, and DSC of 0.88 (Dynamic) and 0.81 (Pediatric), with reduced performance in complex congenital anatomies.

Yun and Choi [27] compared UNet, TransUNet, and MIST on 108 cardiac CT images, where MIST achieved the best performance (Dice 0.74, HD95 5.77 mm). However, limited dataset size, low diversity, and high computational complexity restricted generalization and clinical applicability. Hatamizadeh et al. [28] developed Swin UNETR for 3D brain tumor segmentation on the BraTS 2021 dataset, integrating a hierarchical Swin Transformer encoder with a CNN decoder to capture long-range dependencies, achieving strong Dice performance, though with reduced tumor core accuracy and high computational cost. Jing Zhang et al. [29] proposed ST-Unet using Synapse and ISIC 2018 datasets, combining Transformer and CNN modules to achieve 78.86% Dice and 0.9243 recall but with increased complexity.

2.1 Research gap

Despite the remarkable progress in echocardiographic segmentation using DL, several research gaps are yet to be addressed. Most of the existing models suffer from high computational complexity, limiting their practicality for real-time clinical applications and their integration into routine cardiac imaging workflows [12], [26]. Although GAN-based and attention-based methods improved boundary precision, these models still report limited generalization across different patient data and various imaging protocols [15], [16]. Semi-supervised and self-supervised strategies [21], [23] assisted in reducing the dependence on annotations, but their performance became inconsistent when low-quality or redundant unlabeled data were used. Moreover, most studies have been focused on the segmentation of the LV; hence, research on multi-chamber analysis is at its minimal level [24]. Further, most of the models are not interpretable; they are sensitive to noise, and their adaptabilities to multi-view and pediatric datasets are restricted, which further limits clinical applications. Thus, there is a requirement for the development of efficient, interpretable, and generalizable segmentation frameworks that analyze various echocardiographic modalities while ensuring clinical reliability and scalability.

3 MATERIALS AND METHODS

The proposed model employed Swin-EchoNet, a Swin Transformer-based DL framework, to perform 2CH segmentation in 2D echocardiographic images. Figure 2 depicts the overall workflow of the proposed method, starting from data acquisition to final segmentation. The process starts with input datasets, including the publicly available CAMUS dataset and the clinically collected Aster-MIMS dataset, each with echocardiographic images along with their ground-truth masks traced by expert cardiologists. The echocardiographic images are preprocessed to maintain uniformity and quality across datasets. Further, each DICOM video sequence is converted into individual two-dimensional frames. Then, all frames were resized and normalized into a fixed intensity range. Additionally, noise reduction and contrast enhancement were performed to enhance the clarity of cardiac boundaries. The processed images are divided into a testing and training subset. The Swin-EchoNet Transformer processes the preprocessed training images and generates a segmentation mask for each frame. The predicted masks are compared with the ground-truth masks by a composite loss function that incorporates dice loss, cross-entropy loss, and a boundary consistency term. The loss value indicates a distance between predicted segmentation and true anatomical structure; hence, in case the loss value exceeds an acceptable range, weight update occurs via the backpropagation algorithm by adjusting its inner parameters as a way of minimizing segmentation error. Further, the iteration is repeated until the minimum loss condition is fulfilled.

Once the model converges, optimized weights are saved, hence resulting in the final trained Swin-EchoNet Transformer model. Finally, the segmented output is produced by the trained model on the test dataset. The segmented outputs highlight cardiac chambers, usually the LV and LA, thus enabling further quantitative analysis. The entire process guarantees that the model learns from both datasets effectively and enhances its generalization ability and clinical reliability.

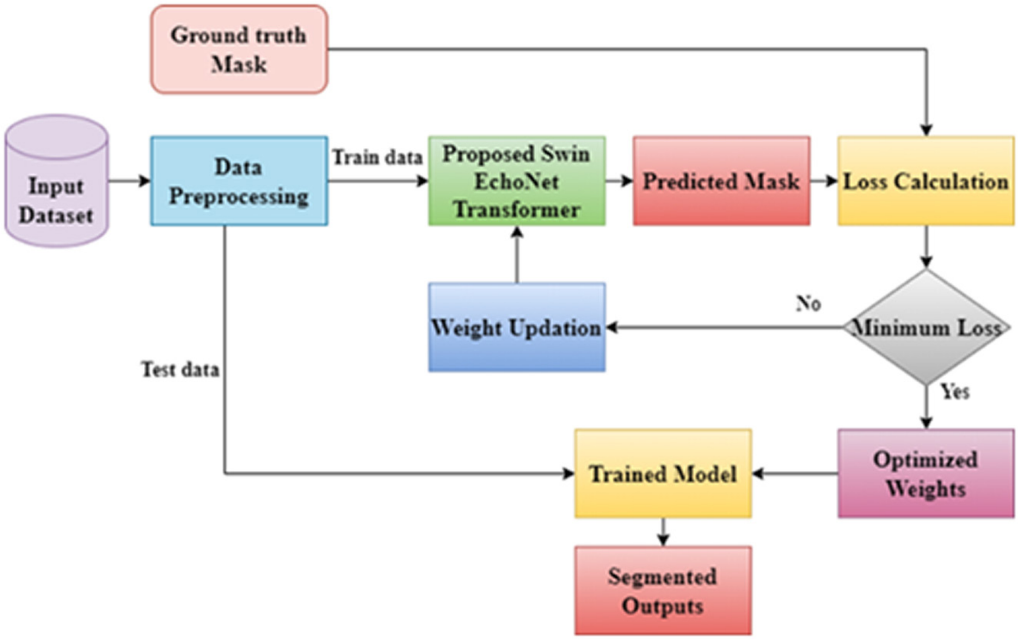

Fig. 2. Block representation of the proposed Swin EchoNet Transformer model

3.1 Dataset description

The proposed method utilized two distinct datasets, such as the CAMUS and Aster-MIMS datasets, for model training and evaluation. The CAMUS dataset was sourced from Kaggle, and Aster-MIMS is a clinical dataset. These datasets ensure both generalizability and clinical applicability of the proposed Swin-EchoNet model by considering a wide range of imaging conditions, patient demographics, and acquisition settings.

Aster-MIMS clinical dataset. The Aster-MIMS clinical dataset was developed in collaboration with Aster MIMS Hospital in Kottakkal, Kerala, India, with ethical approval from the Scientific Research Committee and the Institutional Ethics Committee. Data were collected using a Philips Epiq 7C cardiology ultrasound system, which captures high-resolution echocardiographic sequences in DICOM format. From these clinical studies, 1,550 2D frames were extracted from 120 individuals aged between 20 and 75 years, representing a wide range of physiological and pathological conditions. Each image was carefully annotated using manual segmentation methods to outline the LV, myocardium, and LA. Annotated masks were then verified by an experienced cardiac sonographer to ensure consistency and accuracy. The dataset reflects realistic variability in images due to factors like patient motion, acoustic shadowing, and speckle noise conditions that are common in routine echocardiography but not often found in publicly available data. Figure 3 shows samples from the Aster-MIMS dataset, along with the diversity in cardiac structure, with their respective manually segmented masks.

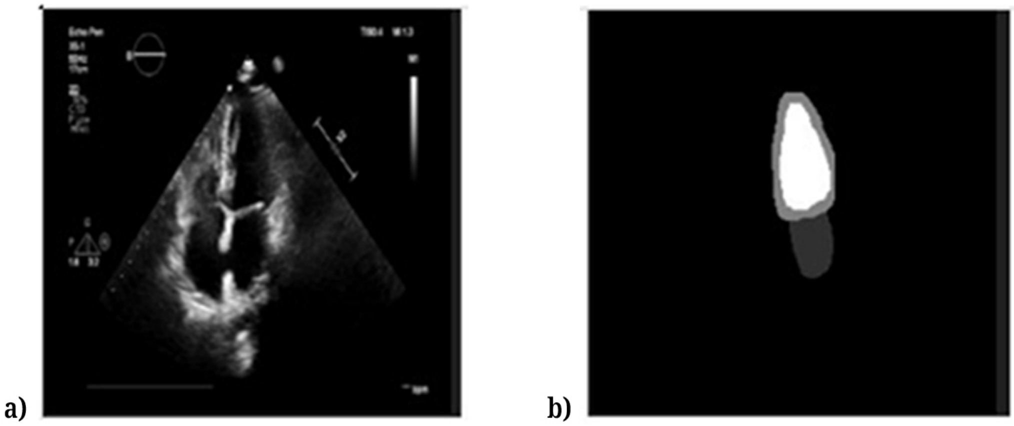

Fig. 3. Sample images from the Aster-MIMS dataset (a) Endocardiographic image (b) Mask

CAMUS dataset. The Cardiac Acquisitions for Multi-Structure Ultrasound Segmentation (CAMUS) dataset is a standardized public benchmark for echocardiographic image analysis, originally introduced by Leclerc et al. [30]. In this study, the dataset was accessed via a publicly available Kaggle repository [31] for convenience; however, all data correspond to the original CAMUS dataset described in [30]. It consists of 1,800 2D echocardiographic images from 500 patients with expert manual annotations on A2C and A4C views. The images are of varying spatial resolutions from 256×256 to 1024×1024 pixels that represent different cardiac morphologies and imaging systems. To ensure diversity in the clinical data, the CAMUS dataset divides the images into three categories of quality, such as good, comprising 916 images; medium, containing 682 images; and poor, which includes 202 images, depending on the visibility of cardiac structures and based on annotation confidence. This variability in image quality increases robustness and generalization capability during segmentation. Figure 4 illustrates the sample images from the CAMUS dataset, and the data distribution is shown in Figure 5.

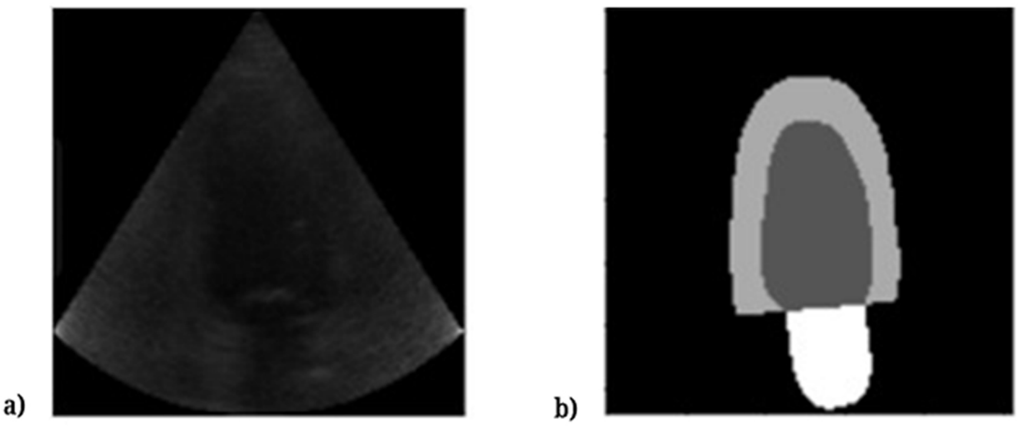

Fig. 4. Sample images from the CAMUS dataset (a) Endocardiographic image (b) Mask

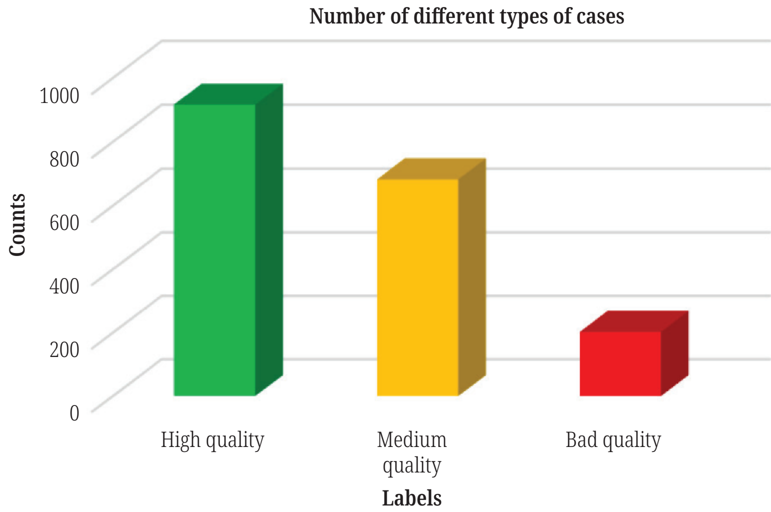

Fig. 5. Data distribution

To ensure a fair evaluation and prevent data leakage, the dataset was partitioned at the patient level rather than at the frame level, ensuring that images from the same subject do not appear in multiple subsets. The dataset was initially divided into training and testing subsets using an 80:20 ratio, and the training set was further split into training and validation subsets using an 80:20 ratio. This strategy avoids overlap between datasets and ensures that the reported performance reflects true generalization capability.

3.2 Manual segmentation

Manual segmentation plays a critical role in the development and validation of the Swin-EchoNet Transformer model. In this work, segmentation of the LA and LV was manually performed at Aster MIMS Hospital under the supervision of an experienced echocardiologist, following the standards of the American Society of Echocardiography (ASE) [32]. These expert-drawn contours represent the most accurate anatomical interpretation of each image and are treated as the ground-truth reference during training and evaluation. DL models depend on supervised learning; the accuracy of their predictions directly relies on the quality of their reference labels. The manually generated masks showing the closed contours of the LV and LA, including both endocardial and epicardial boundaries, allow the model to learn the structural patterns of the myocardium and internal cavity. During training, the Swin-EchoNet Transformer compares the predicted segmentation to these reference masks and adjusts its internal weights to minimize the error. These are also utilized for performance assessment using metrics such as DSC and IoU. Manual segmentation in this study lays the basis for supervised learning. This ensures that the Swin-EchoNet Transformer learns to replicate clinically meaningful cardiac boundaries and makes its automated segmentation results both anatomically accurate and diagnostically relevant.

Figure 6 shows a sample training mask for LV segmentation on echocardiographic images. The mask visually differentiates those key cardiac regions that are normally marked when doing manual segmentations, namely the epicardium, endocardium, and left atrium. The epicardium forms the outermost boundary of the heart, constituting the layer that provides structural support and physiological signaling. The endocardium defines the inner boundary of the left ventricle, which forms the blood-filled chamber in which volumetric and functional analyses are done. The LA is the adjoining cavity involved in the filling phase of the cardiac cycle. These were manually delineated under expert supervision using ITK-SNAP, an open-source medical imaging tool that enables precise contouring of cardiac structures. Such masks provide the anatomical ground truth required for supervised training of the Swin-EchoNet Transformer to accurately learn chamber geometry and myocardial thickness.

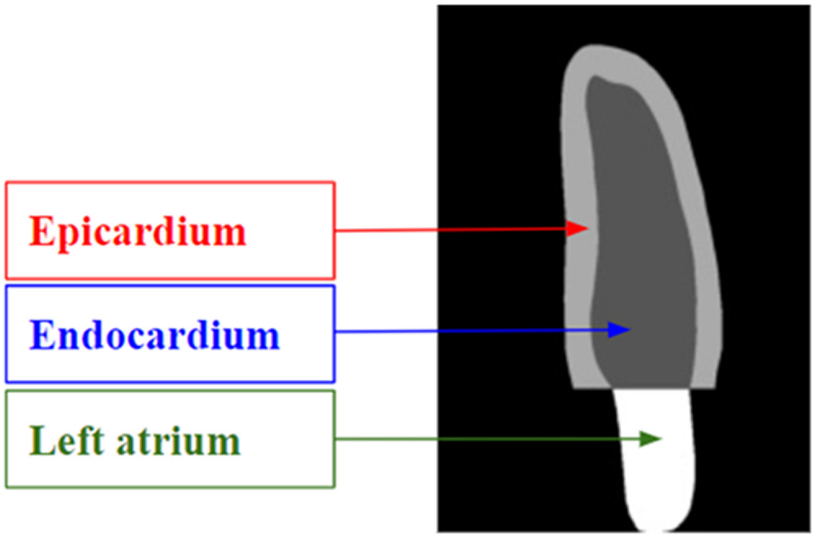

Fig. 6. Training mask for left-ventricular segmentation

3.3 Data preprocessing

All echocardiographic images were pre-processed before training the Swin-EchoNet model to enhance image quality, standardize resolution, and ensure consistency across the datasets from various sources. These steps are essential pre-processing steps to be performed on echocardiographic images for obtaining better model convergence, considering the factors of variability in echocardiographic images that include intensity variations, speckle noise, and probe orientation. Each echocardiographic frame was then resized along with its ground-truth mask to 128×128 pixels, which maintains uniform dimensions while retaining key structural details of the cardiac chambers. Then these resized images were normalized into a uniform intensity scale through min-max normalization as:

$$I_{\text{norm}} = \frac{I - I_{min}}{I_{max} - I_{min}} \tag{1}$$

where I_{norm} is the normalized pixel value in the range [0,1], I_{max} and I_{min} are the maximum and minimum intensity values in the image and I is the original pixel intensity. This normalization step removes effects of varying illumination so that the pixel intensity values contribute consistently in the process of feature learning. A Gaussian smoothing filter ($\sigma = 1$) was applied to reduce speckle noise and enhance low-contrast boundaries, smoothing high-frequency artifacts while maintaining anatomical edges. Then, CLAHE was applied to increase local contrast and improve the visualization of chamber boundaries in low-quality echocardiographic images. These preprocessing steps increase clarity of the myocardial contours and atrial-ventricular interfaces.

Data augmentation, namely random rotations of ±20°, vertical and horizontal flipping, elastic deformations, and intensity scaling, simulating realistic variations of patient position, heart orientation, and imaging conditions, was done to further enhance robustness and generalization of the proposed model. This increased the diversity within the training dataset, with less risk of overfitting, thus improving the performance on new test samples. Finally, each preprocessed image was divided into non-overlapping patches, 4×4 pixels in size, so as to be able to create token embeddings for transformer-based learning. Patch embeddings comprise the input sequence for the SwinEchoNet encoder, which allows both local textural details and global spatial relationships across the echocardiographic field to be taken into account. These preprocessing steps standardized, cleaned, and prepared the data, which is suitable for the Swin-EchoNet model.

3.4 Model development

The Swin-EchoNet Transformer was proposed for the automatic segmentation of cardiac chambers from 2D echocardiographic images. It combines convolutional preprocessing with transformer-based attention to grasp both local and global spatial dependencies. It follows an encoder-decoder structure: an encoder for hierarchical feature extraction by means of Swin Transformer blocks and a decoder to reconstruct fine structural details via patch-expanding and skip connections. The model is trained with manually annotated ground-truth masks by optimizing a composite loss function that balances Dice and cross-entropy losses. Through iterative training, Swin-EchoNet learns to outline correctly both the LA and ventricle boundaries with high accuracy, thus enhancing precision in echocardiographic diagnosis.

Swin transformer. The Swin Transformer extracts both spatial and sequential dependencies from image data, which enables the model to detect complex patterns. The shifted window mechanism enables the network to model both global and local contextual relationships with a high degree of computational efficiency [33]. This architecture leverages a preliminary convolutional layer to conduct low-level feature extraction, while a fully connected embedding layer structures the extracted features into a format that is suitable for the transformer blocks, illustrated in Figure 7. This model incorporates several Swin Transformer blocks, each striving to better capture the spatial features with more contextual understanding across image regions. Each Swin Transformer block has four main components comprising Shifted Window Multi-Head Self-Attention (SW-MSA), Multilayer Perceptron (MLP), Window-based Multi-Head Self-Attention (W-MSA), and Layer Normalization (LN). W-MSA attends to the local features within non-overlapping windows. Complementing this, SW-MSA allows for interaction between windows by enabling the transfer of information across non-overlapping windows, thus allowing the combination of global features. The nonlinear transformation of the learned representations is further consolidated through the MLP with the ReLU activation, while LN normalizes the intermediate feature distribution to use for improving convergence. The sequential operations performed in a single Swin Transformer block are mathematically defined as:

$$\hat{f}^{L} = WMSA(LN(f^{L-1})) + f^{L-1} \tag{2}$$

$$f^{L} = MLP(LN(\hat{f}^{L})) + \hat{f}^{L} \tag{3}$$

$$\hat{f}^{L+1} = SWMSA(LN(f^{L})) + f^{L} \tag{4}$$

$$f^{L+1} = MLP(LN(\hat{f}^{L+1})) + \hat{f}^{L+1} \tag{5}$$

A patch embedding module combines and down samples feature tokens, which form hierarchical feature maps and allow adaptive pooling across varying input dimensions. This attention structure enables the model to learn both fine-grained global and local spatial coherence.

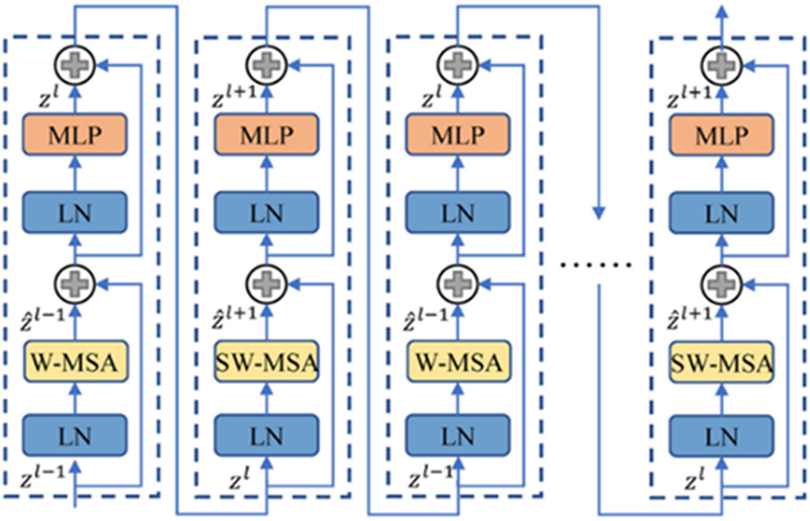

Fig. 7. Swin transformer architecture

Proposed Swin EchoNet transformer. The Swin-EchoNet Transformer is a DL model designed for the automated segmentation of cardiac chambers in 2D echocardiography. It retains several structural advantages of the Swin Transformer while introducing key modifications tailored to the challenges of echocardiographic imaging, as illustrated in Figure 8. Common issues such as speckle noise, acoustic shadowing, and low tissue contrast make accurate boundary delineation difficult. To address these challenges, Swin-EchoNet incorporates domain-specific enhancements, including view awareness, hierarchical attention, and ultrasound-oriented preprocessing within a U-Net-like encoder–decoder architecture.

The input echocardiographic images were resized to a fixed resolution of 256×256 pixels to ensure consistency across all samples. The model partitions each image into non-overlapping patches of size 4×4 pixels, which are linearly embedded and processed through hierarchical transformer blocks. These blocks combine window-based self-attention and shifted windows to effectively capture contextual relationships across the image. In the encoder path, patch-merging layers progressively reduce spatial resolution while increasing channel depth to generate multi-scale feature representations. A bottleneck module further enhances long-range feature interactions, while the decoder reconstructs the segmentation mask using patch-expanding layers and skip connections to preserve fine structural details.

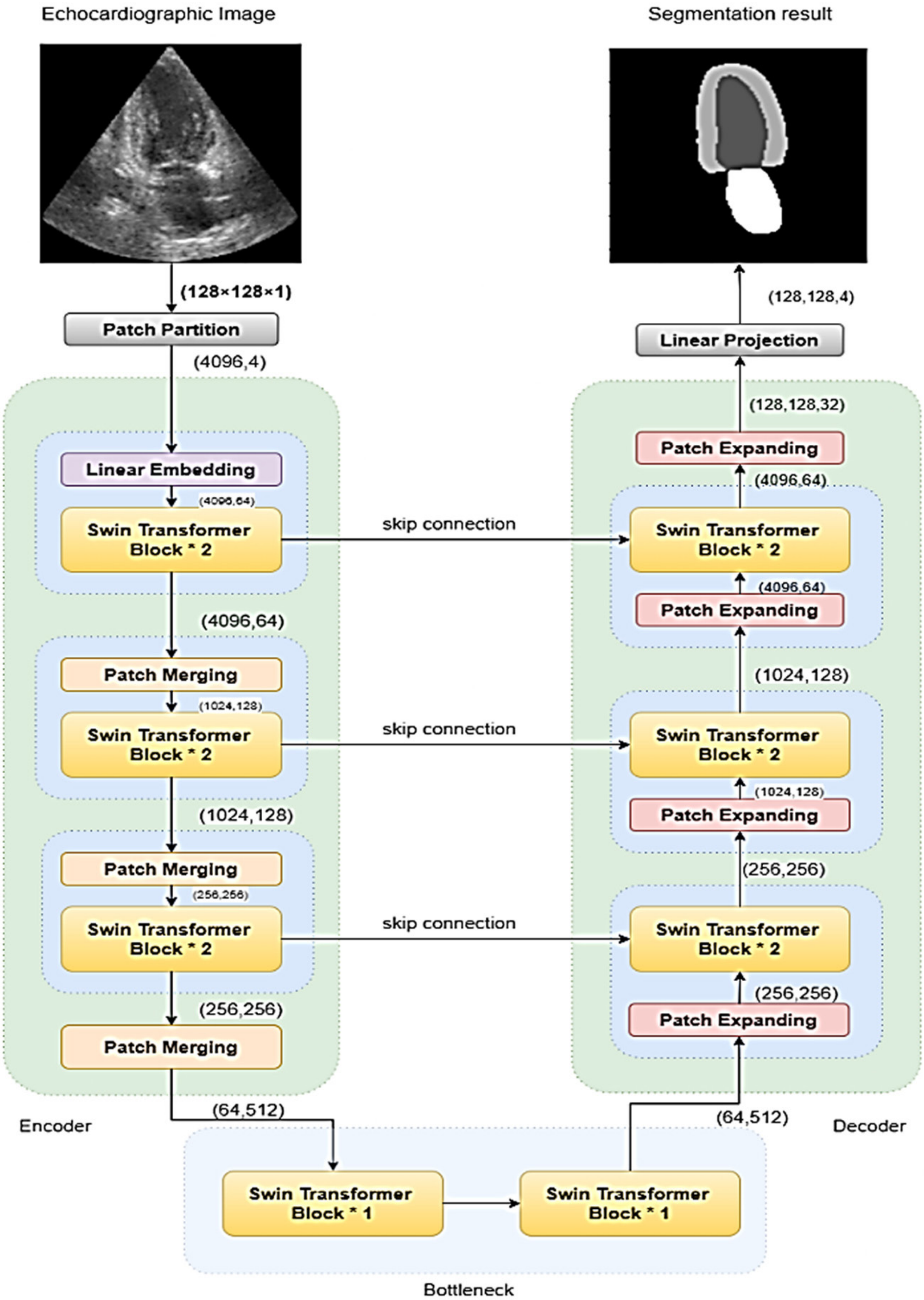

Fig. 8. Proposed Swin EchoNet transformer

A refinement layer further reduces noise and improves boundary visibility before producing the final multi-class segmentation map, highlighting the LA, LV, and surrounding myocardium. Swin-EchoNet differs from a standard Swin Transformer in three key aspects. First, it employs a fully symmetric encoder–decoder architecture with skip connections to enable precise pixel-wise segmentation instead of a classification-oriented design. Second, the framework incorporates domain-specific preprocessing tailored for echocardiographic images, improving feature quality under conditions such as speckle noise and low contrast. Third, the model utilizes a boundary-aware composite loss function combining Dice, cross-entropy, and gradient-based penalties to improve contour delineation and segmentation accuracy.

The proposed framework operates on individual 2D echocardiographic frames, and no explicit temporal modeling is incorporated. The architectural components illustrated correspond directly to the described encoder–decoder structure, ensuring clarity and reproducibility. Compared to conventional CNN-based approaches, Swin-EchoNet provides improved structural consistency in noisy imaging conditions. The hierarchical attention mechanism captures both local and global contextual relationships while maintaining computational efficiency, making the model suitable for real-time clinical applications.

3.5 Simulation setup

The system carried out the demanding calculations with an Intel Core i7-6850K processor (3.6 GHz base clock speed). The system also has 32GB of DDR4 RAM to handle large datasets and memory-intensive tasks while utilizing a 1TB SSD for faster data access and processing. Google Colaboratory was used as the DL platform to develop and train the model using the DL libraries TensorFlow and Keras, Python is utilized as the programming language. TensorFlow uses cuDNN libraries for GPU acceleration to fully utilize GPU capabilities. The model uses the Windows 10 operating system for the DL platform and is trained using carefully tuned hyperparameters, as shown in Table 1.

Table 1. Hyperparameter specifications of the proposed model

Hyperparameters	Proposed Swin EchoNet Transformer
Input size	(256,256,1)
Batch size	32
Optimizer	Adam
Output Activation	Softmax
Patch size	4×4
Learning rate	0.0001
No. of epochs	200
Loss function	Categorical crossentropy
Filter number	256

4 RESULTS AND DISCUSSION

4.1 Training performance analysis

The accuracy plot shown in Figure 9 illustrates the performance progression of the proposed model. At the beginning of training (epoch 1), the model records a training accuracy of 0.70 and a validation accuracy of 0.82, with corresponding training and validation losses of around 0.78 and 0.45, respectively. This sharp change showed that the model has quickly learned to recognize the primary structural characteristics of the LV and LA. During epoch 5, both accuracies increased to 0.85, with losses decreased to 0.33 and 0.35 for validation and training, respectively. Beyond this point, the curves begin to flatten, indicating that the model has reached a balanced state where new features are learned but overfitting is prevented. At the point of epoch 15 in both sets, accuracy remained around 0.85, with minor fluctuations, and losses stayed very constant at about 0.33-0.34. This reflects an excellent generalization on unseen data by the Swin-EchoNet Transformer and that regularization techniques such as normalization and augmentation help avoid overfitting. Finally, at epoch 25, the training accuracy approaches 0.86 while the validation accuracy remains at 0.85 for both curves, and losses for each are nearly identical (~0.34). The small difference in training versus validation performance implies excellent generalization with reliable learning. In summary, these results reflect that Swin-EchoNet Transformer converges fast and optimizes well, steadily navigating the trade-off between accuracy and loss, hence being appropriate for echocardiographic image segmentation on the CAMUS dataset.

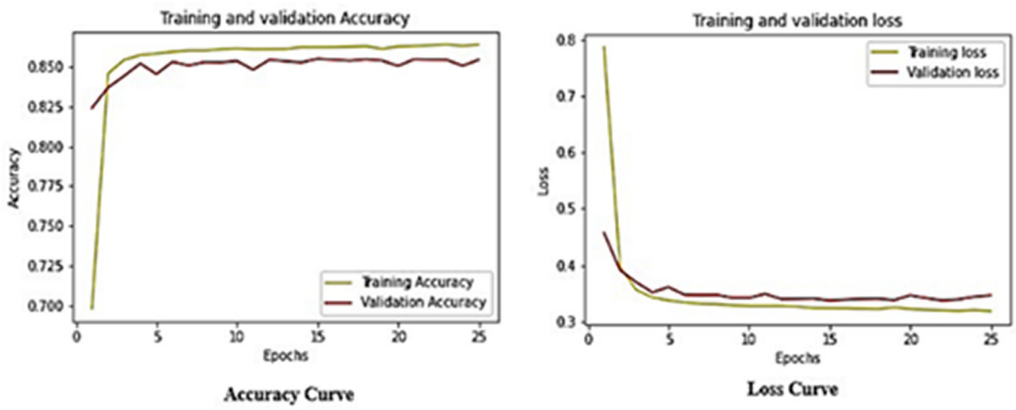

Fig. 9. Accuracy and loss plot of the proposed model for the CAMUS dataset

Figure 10 presents the accuracy and loss curves of the Swin-EchoNet Transformer model on the Aster-MIMS clinical dataset. The accuracy plot demonstrates a rapid and steady incremental increase in model performance observed at the beginning of training. Training accuracy begins at epoch 1 close to about 0.63, with validation accuracy at 0.90, indicating the model is quickly identifying the inherent structures present in the echocardiographic data. By epoch 5, training and validation accuracy exceeded values approximately over 0.95, indicating that the model accurately learned the cardiac chamber boundaries. Beyond epoch 10, the curves stabilized and were around 0.95 and 0.96 with stability maintained until epoch 25, which suggests that the model achieves convergence without overfitting or degradation in validation performance. The close overlap of both accuracy curves across every epoch exemplifies excellent generalization and consistent learning behavior between training and unseen data.

The loss curve confirms the model's stability and learning efficiency. First, training loss starts at approximately 1.0, while the validation loss starts at 0.45; both drop steeply during the initial epochs, reaching 0.15 for each at epoch 5, consistent with the steep rise in model accuracy. Beyond epoch 10, both the loss curves flattened and became just about parallel to each other; this indicates that the optimization process has arrived at convergence, and the model has stopped making active weight updates. The learning curves for the MIMS dataset indicate that the Swin-EchoNet Transformer converges rapidly, achieving high segmentation accuracy with strong generalization across diverse clinical echocardiographic images.

Although the model was configured to train for a maximum of 200 epochs, early stopping based on validation loss was employed to prevent overfitting. The training process showed stabilization within the first 25 epochs, after which no significant improvement in validation performance was observed. Therefore, Figures 9 and 10 illustrate the first 25 epochs, where the convergence behavior is most prominent.

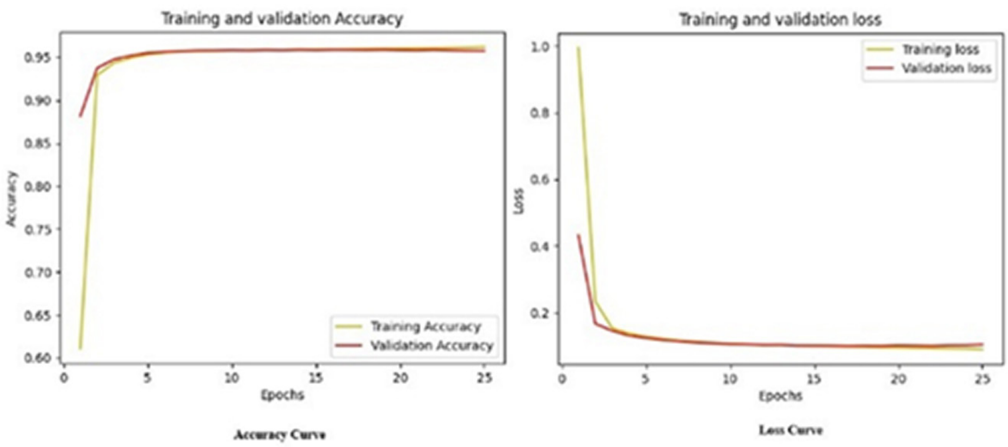

Fig. 10. Accuracy and loss plot of the proposed model for the MIMS dataset

4.2 Performance evaluation

The performance of the proposed model was evaluated using two widely used metrics, namely the DSC and the IoU, which quantify the overlap between predicted segmentation and ground truth.

DSC. DSC measures the overlap between the predicted segmentation region (A) and the ground truth region (B), expressed as:

$$Dice\ score = \frac{2|A \cap B|}{|A| + |B|} \tag{6}$$

where $|B|$ and $|A|$ denote the pixel count in the ground truth and predicted region, respectively, and $|A \cap B|$ denotes the common pixel count. The value of DSC lies in the range of 0 to 1, with 1 indicating complete overlap and 0 means no overlap.

Intersection over union. IoU calculates the similarity between two sample sets using the ratio of the intersection area and the union area, expressed as:

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{7}$$

where $|A \cup B| = |A| + |B| - |A \cap B|$ indicates the total number of unique pixels across both regions. The value ranges from 0 to 1, where 1 represents better segmentation results. Although both metrics evaluate overlap, IoU is more sensitive to boundary mismatches and penalizes false positives more strongly than DSC. In echocardiographic images, where speckle noise and low contrast lead to unclear boundaries, this difference becomes more pronounced. Consequently, higher DSC values correspond to comparatively lower IoU values, which is consistent with observations in ultrasound image segmentation studies. The reported performance metrics are computed on a per-image basis across the test dataset. The mean and standard deviation values represent the distribution of segmentation performance over all test samples. Due to variability in image quality, anatomical differences, and the presence of noise in echocardiographic data, higher standard deviation values are observed, particularly in low-quality images.

Table 2 summarizes the performance of the proposed model on the CAMUS dataset. The model obtained high DSC in the range of 0.94 to 0.98 for all cardiac structures, indicating excellent overlap between the manually annotated and the predicted regions. The IoU values ranging from 0.32 to 0.52 show excellent boundary accuracy, even under varying quality conditions. Among the segmented regions, the LV endocardium (LV endo) obtained the highest accuracy, followed by the LA. The small difference in training and testing evaluations, within 2–3%, supports good generalization ability and model stability. The Swin EchoNet Transformer demonstrates robust and reliable segmentation performance on the CAMUS dataset, producing high accuracy across all anatomical regions. Figure 11 illustrates the comparison of test images, ground truth masks, and model predictions, where the strong visual agreement is consistent with the high DSC and IoU values reported in Table 2, confirming accurate boundary delineation across different image qualities.

Table 2. Performance metrics of the Swin-EchoNet model on the CAMUS dataset

		Training Dataset	Validation Dataset	Testing Dataset
Image count		1080	180	540
2*LV_{endo}	DSC	0.9789 ± 0.2101	0.9745 ± 0.2554	0.9718 ± 0.1181
	IoU	0.3345 ± 0.1642	0.3176 ± 0.2176	0.2995 ± 0.1842
2LV_{epi}	DSC	0.9562 ± 0.1665	0.9504 ± 0.2342	0.9448 ± 0.1524
	IoU	0.3257 ± 0.1243	0.3230 ± 0.2432	0.3127 ± 0.1557
2*LA	DSC	0.9724 ± 0.2311	0.9671 ± 0.1534	0.9658 ± 0.1256
	IoU	0.5235 ± 0.2771	0.5124 ± 0.2658	0.4914 ± 0.1426

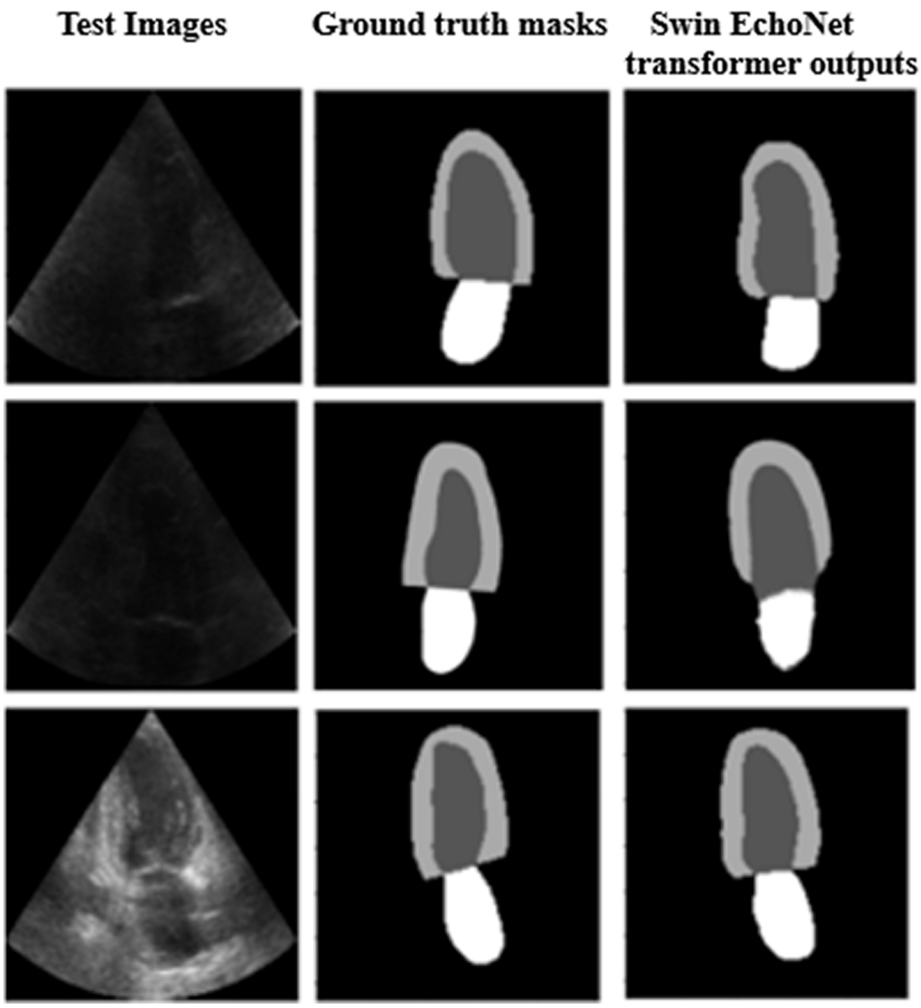

Fig. 11. Visual comparison of test image, ground truth mask, and SwinEchoNet predictions for CAMUS dataset

The per-class evaluation provides a clearer understanding of model performance across different anatomical structures, as shown in Table 3. The LV endocardium achieves the highest segmentation accuracy with a DSC of 0.9718 ± 0.1181 and an IoU of 0.2995 ± 0.1842, indicating strong overlap performance. The LA also demonstrates high accuracy with a DSC of 0.9658 ± 0.1256 and IoU of 0.4914 ± 0.1426. In contrast, slightly lower performance is observed for the LV epicardium, with a DSC of 0.9448 ± 0.1524 and IoU of 0.3127 ± 0.1557, due to its complex boundary characteristics. The variation in standard deviation across classes reflects differences in segmentation difficulty and image quality in echocardiographic data.

Table 3. Per-class segmentation performance on CAMUS dataset

Structure	DSC (Mean ± Std)	IoU (Mean ± Std)
LV Endocardium	0.9718 ± 0.1181	0.2995 ± 0.1842
LV Epicardium	0.9448 ± 0.1524	0.3127 ± 0.1557
Left Atrium	0.9658 ± 0.1256	0.4914 ± 0.1426

4.3 Comparative analysis with baseline models

To assess the effectiveness of the Swin-EchoNet Transformer, a comparative analysis was conducted against three established segmentation architectures, such as U-Net [34], Attention U-Net [35], and ResU-Net [36], using the Aster-MIMS clinical dataset. All three were used due to the emphasis of those approaches in the medical image segmentation field. U-Net, a primarily symmetric encoder-decoder architecture, shows effectiveness in pixel-wise learning but has challenges with complicated texture variation, which occurs in echocardiography. Attention U-Net, which incorporates attention gates to highlight relevant features, is still not accurate regarding fine-grained boundary details when contrast of the image is low. ResU-Net utilized residual connections to improve gradient flow and deeper feature learning but also has high computational complexity. In contrast, the Swin-EchoNet Transformer uses hierarchical attention with window-based self-attention mechanisms, which capture local spatial features and global contextual dependencies, key advantages in noisy and low-contrast ultrasound environments.

As indicated in Table 4, the Swin-EchoNet model obtained a mean DSC of 0.9738 ± 0.1978 and a mean IoU of 0.7263 ± 0.1243, outperforming all baseline models. On the other hand, the established models (ResU-Net, U-Net, and Attention U-Net) obtained dice scores between 0.92 and 0.94, while the IoU was less than 0.52, highlighting overlap accuracy. The Swin-EchoNet predictions demonstrate a high degree of agreement with the manually segmented ground truth, maintaining a sharp definition of the LV and LA across different quality images, including low, medium, and high-quality images. Even on low-quality images, the segmentation generated by the Swin-EchoNet model is structurally consistent and has minimal false borders. These accurate segmentation results underline the strength of the model and its adaptability in real clinical echocardiographic data. The Swin-EchoNet Transformer presents clearer boundaries, with smooth contours and almost human-level consistency in segmentation, which leads not just to improving accuracy in segmentation estimation but also facilitates a more efficient workflow for cardiac measurements apart from diagnosis reliability in clinical practice.

The performance improvements observed are consistent across both Dice and IoU metrics, indicating the robustness of the proposed model. The clear margin of improvement over baseline models suggests that the gains are not due to random variation but reflect the effectiveness of the proposed architecture. Figure 12 visualizes the comparison test, ground truth, and predicted images of various models on the Aster-MIMS dataset.

Table 4. Performance evaluation of different models on the Aster-MIMS dataset

Model	Mean IoU	Mean DSC
U-Net	0.4934 ± 0.1842	0.9292 ± 0.2435
Attention U-Net	0.4993 ± 0.1673	0.9328 ± 0.2164
ResU-Net	0.5156 ± 0.1911	0.9418 ± 0.2013
Swin- EchoNet	0.7263 ± 0.1243	0.9738 ± 0.1978

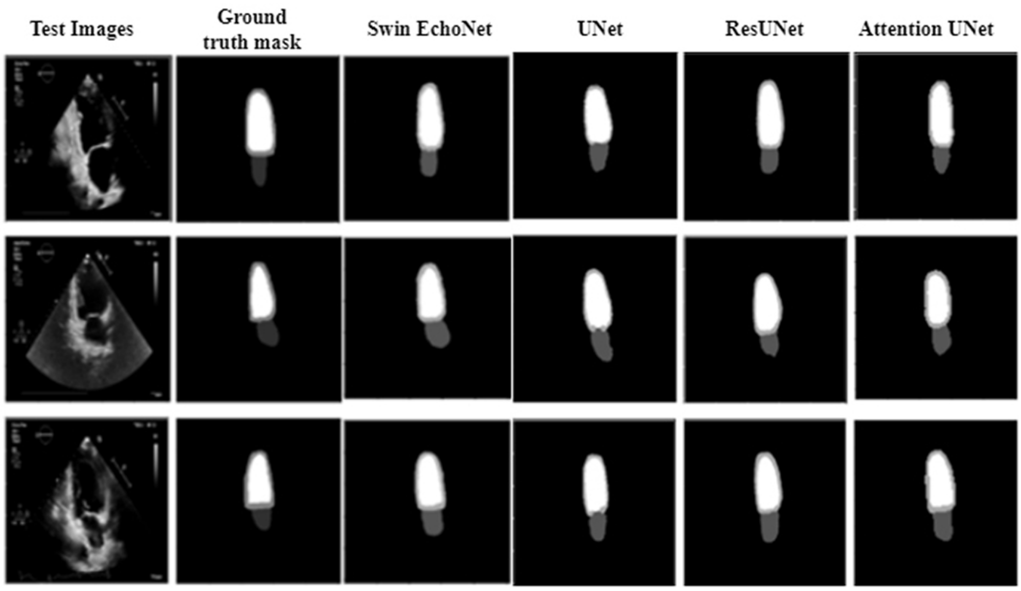

Fig. 12. Comparison of test images, ground truth masks, and predicted outputs of different models on the Aster-MIMS dataset

4.4 Computational complexity comparison

To evaluate practical applicability, a computational complexity analysis was conducted comparing Swin-EchoNet with baseline models using parameter count and inference time. As shown in Table 5, U-Net, Attention U-Net, and ResUNet have approximately 31M, 34M, and 36M parameters with inference times of 45 ms, 52 ms, and 58 ms, respectively. In contrast, Swin-EchoNet has a lower parameter count of about 28M and an inference time of 50 ms. The use of shifted window attention enables efficient computation, demonstrating that the proposed model achieves a balance between performance and computational efficiency for clinical applications.

Table 5. Computational complexity comparison of segmentation models

Model	Parameters (M)	Inference Time (ms/image)	Complexity Type
U-Net	~31 M	~45 ms	Convolution-based
Attention U-Net	~34 M	~52 ms	CNN with Attention
ResUNet	~36 M	~58 ms	Residual CNN
Swin-EchoNet	~28 M	~50 ms	Window-based Transformer

4.5 Ablation study

To evaluate the contribution of the transformer-based architecture within the proposed Swin-EchoNet framework, an ablation study was conducted comparing the Swin Transformer encoder with a CNN-based encoder represented by the ResUNet model. The analysis was performed on the Aster-MIMS clinical dataset to reflect real-world imaging conditions. The CNN-based encoder achieved a mean DSC of 0.9418 ± 0.2013 and IoU of 0.5156 ± 0.1911, while the Swin Transformer achieved higher performance with a DSC of 0.9738 ± 0.1978 and IoU of 0.7263 ± 0.1243, as shown in Table 6. The improvement is attributed to the transformer's ability to capture both local and global contextual dependencies, enabling better boundary delineation and robustness in noisy and low-contrast echocardiographic images.

Table 6. Ablation study on Aster-MIMS dataset

Model Configuration	Mean DSC	Mean IoU
CNN-based Encoder (ResUNet)	0.9418 ± 0.2013	0.5156 ± 0.1911
Swin Transformer Encoder	0.9738 ± 0.1978	0.7263 ± 0.1243

4.6 Comparison with state-of-the-art methods

Table 7 presents the comparative performance of the proposed model against existing approaches. The UNeXt model achieved a DSC of 0.886 ± 0.0292, indicating moderate segmentation performance with limitations in capturing global contextual information. The modified U-Net obtained a Dice score of 0.899 ± 0.017, showing slight improvement but still constrained by localized feature learning. Similarly, the SAM fine-tuned model achieved 0.891 ± 0.040, demonstrating effective feature focus but reduced sensitivity to subtle boundary variations in low-contrast regions. The hybrid EchoEFNet model reported a higher Dice score of 0.934 ± 0.002, reflecting strong performance with increased computational complexity. In comparison, the proposed Swin-EchoNet Transformer achieved a Dice score of 0.951 ± 0.0227, outperforming all evaluated models. This improvement is attributed to the shifted window self-attention mechanism, which effectively captures both local and global contextual dependencies. Additionally, comparisons with Swin-based architectures such as ST-Unet [29] and the Swin Transformer U-Net [26] indicate that, although these models leverage transformer-based feature learning, they lack domain-specific adaptations for echocardiographic data. The incorporation of ultrasound-oriented preprocessing, view-aware representations, and a boundary-aware loss function enables the proposed model to achieve superior boundary delineation and robustness under challenging clinical conditions.

Table 7. Performance comparison of proposed model with existing methods

Authors [Ref]	Methodology	DSC
Jing Zhang et al. [29]	ST-Unet	0.7886
Nemri et al. [26]	Swin Transformer U-Net	0.88 (Dynamic), 0.81 (Pediatric)
Wu et al. [12]	UNeXt	0.886 ± 0.0292
Chao et al. [17]	SAM-fine-tuned	0.891 ± 0.040
Kang et al. [18]	Variation of U-Net	0.899 ± 0.017
Li et al. [20]	EchoEFNet	0.934 ± 0.002
Proposed Swin-EchoNet model		**0.951 ± 0.227**

Notes: Standard deviation (±) values are reported only where provided in the original studies; for other methods, only the mean DSC is presented due to unavailability of variability metrics.

5 CONCLUSION

The proposed Swin-EchoNet Transformer has shown a powerful and effective framework for automatic segmentation of cardiac chambers from 2D echocardiography. Hierarchical feature representation combined with shifted-window self-attention effectively captures the local and global contextual information to delineate both the LV and LA more precisely. The experimental results on both the CAMUS and Aster-MIMS datasets confirm its superior performance with high Dice and IoU scores over CNN-based models, such as Attention U-Net, U-Net, and ResUNet. The model shows excellent generalization capability across datasets of varying image quality, highlighting its potential for real-world clinical applications. Furthermore, the architecture of Swin-EchoNet maintains computational efficiency. Thus, it serves as an ideal model for integration into automated diagnostic systems and point-of-care ultrasound platforms. Although the results demonstrate consistent improvements across multiple evaluation metrics, formal statistical significance analysis was not conducted due to the unavailability of per-sample prediction data. This will be addressed in future work through detailed statistical validation to further strengthen the reliability of the observed performance gains. In the future, further work will be focused on the extension of the current segmentation framework to multi-chamber and multi-view segmentation, including A4C and short-axis views for comprehensive cardiac assessment. Temporal modeling using either video-based transformers or recurrent mechanisms of attention could further enhance the consistency of segmentation across cardiac cycles. Additionally, exploring semi-supervised and federated learning strategies enables model adaptation across hospitals while preserving patient data privacy. Integration of Swin-EchoNet into real-time clinical workflows and decision-support systems will further align research innovation with clinical application, advancing automated echocardiographic analysis toward more accurate, efficient, and accessible cardiovascular diagnostics.

6 DATA AVAILABILITY

The datasets utilized in this study are not publicly accessible due to privacy and security considerations. Data were obtained from Aster MIMS Hospital, Kottakal, Kerala, India. De-identified data may be made available upon reasonable request and with approval from the Institutional Ethics Committee (EC/13/2021, dated March 25, 2021).

7 ACKNOWLEDGMENT

The authors express their sincere gratitude to Aster MIMS Hospital, Kottakal, for their support in providing data and facilities for this study. Special thanks are extended to the doctors, echocardiography technicians, and staff for their valuable assistance and cooperation throughout the study.

8 CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this manuscript.

9 REFERENCES

[1] A. Naveed, R. Atique, H. A. Saeed, J. Sharif, A. Haidar, and A. Samad, "Cardiovascular diseases: Understanding the leading cause of death worldwide," *Global Journal of Multidisciplinary Sciences and Arts*, vol. 1, no. 1, pp. 100–110, 2024. https://doi.org/10.70445/gjmdsa.1.1.2024.100-110

[2] V. Christoffels and B. Jensen, "Cardiac morphogenesis: Specification of the four-chambered heart," *Cold Spring Harbor Perspectives in Biology*, vol. 12, no. 10, p. a037143, 2020. https://doi.org/10.1101/cshperspect.a037143

[3] M. F. Buijtendijk, P. Barnett, and M. J. Van Den Hoff, "Development of the human heart," *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, vol. 184, no. 1, pp. 7–22, 2020. https://doi.org/10.1002/ajmg.c.31778

[4] C. Liguori, S. Tamburrini, G. Ferrandino, S. Leboffe, N. Rosano, and I. Marano, "Role of CT and MRI in cardiac emergencies," *Tomography*, vol. 8, no. 3, pp. 1386–1400, 2022. https://doi.org/10.3390/tomography8030112

[5] L. D. Weberling, D. Lossnitzer, N. Frey, and F. Andre, "Coronary computed tomography vs. cardiac magnetic resonance imaging in the evaluation of coronary artery disease," *Diagnostics*, vol. 13, no. 1, p. 125, 2022. https://doi.org/10.3390/diagnostics13010125

[6] M. Habijan *et al.*, "Overview of the whole heart and heart chamber segmentation methods," *Cardiovascular Engineering and Technology*, vol. 11, no. 6, pp. 725–747, 2020. https://doi.org/10.1007/s13239-020-00494-8

[7] L. Lenchik *et al.*, "Automated segmentation of tissues using CT and MRI: A systematic review," *Academic Radiology*, vol. 26, no. 12, pp. 1695–1706, 2019. https://doi.org/10.1016/j.acra.2019.07.006

[8] A. C. Kwan, G. Salto, S. Cheng, and D. Ouyang, "Artificial intelligence in computer vision: Cardiac MRI and multimodality imaging segmentation," *Current Cardiovascular Risk Reports*, vol. 15, p. 18, 2021. https://doi.org/10.1007/s12170-021-00678-4

[9] A. Amer, X. Ye, and F. Janan, "ResDUnet: A deep learning-based left ventricle segmentation method for echocardiography," *IEEE Access*, vol. 9, pp. 159755–159763, 2021. https://doi.org/10.1109/ACCESS.2021.3122256

[10] N. Azarmehr, X. Ye, S. Sacchi, J. P. Howard, D. P. Francis, and M. Zolgharni, "Segmentation of left ventricle in 2D echocardiography using deep learning," in *Annual Conference on Medical Image Understanding and Analysis*, 2019, pp. 497–504. https://doi.org/10.1007/978-3-030-39343-4_43

[11] C. Chen *et al.*, "Deep learning for cardiac image segmentation: A review," *Frontiers in Cardiovascular Medicine*, vol. 7, p. 508599, 2020. https://doi.org/10.3389/fcvm.2020.00025

[12] L. Wu *et al.*, "Automatic segmentation of the left ventricle in apical four-chamber view on transesophageal echocardiography based on UNeXt deep neural network," *Diagnostics*, vol. 14, no. 23, p. 2766, 2024. https://doi.org/10.3390/diagnostics14232766

[13] H. Assadi *et al.*, "Development and validation of AI-derived segmentation of four-chamber cine cardiac magnetic resonance," *European Radiology Experimental*, vol. 8, no. 1, p. 77, 2024. https://doi.org/10.1186/s41747-024-00477-7

[14] N. Fatima, S. Afrakhteh, G. Iacca, and L. Demi, "Automatic segmentation of 2-D echocardiography ultrasound images by means of generative adversarial network," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 71, no. 11, pp. 1552–1564, 2024. https://doi.org/10.1109/TUFFC.2024.3393026

[15] J. Wang, X. Bo, G. Li, and Y. Tan, "Ventricular segmentation algorithm for echocardiography based on transfer learning and GAN," *Systems and Soft Computing*, vol. 6, p. 200139, 2024. https://doi.org/10.1016/j.sasc.2024.200139

[16] Z. Ye, T. Chen, D. Wang, F. Wang, and L. Zhang, "HFE-mamba: High-frequency enhanced mamba network for efficient segmentation of left ventricle in pediatric echocardiograms," *IEEE Access*, vol. 12, pp. 123038–123048, 2024. https://doi.org/10.1109/ACCESS.2024.3424546

[17] C. J. Chao *et al.*, "Foundation versus domain-specific models for left ventricular segmentation on cardiac ultrasound," *npj Digital Medicine*, vol. 8, no. 1, p. 341, 2025. https://doi.org/10.1038/s41746-025-01730-y

[18] S. Kang, S. J. Kim, H. G. Ahn, K. C. Cha, and S. Yang, "Left ventricle segmentation in transesophageal echocardiography images using a deep neural network," *PLoS ONE*, vol. 18, no. 1, p. e0280485, 2023. https://doi.org/10.1371/journal.pone.0280485

[19] M. J. Mortada, S. Tomassini, H. Anbar, M. Morettini, L. Burattini, and A. Sbrollini, "Segmentation of anatomical structures of the left heart from echocardiographic images using deep learning," *Diagnostics*, vol. 13, no. 10, p. 1683, 2023. https://doi.org/10.3390/diagnostics13101683

[20] H. Li, Y. Wang, M. Qu, P. Cao, C. Feng, and J. Yang, "EchoEFNet: Multi-task deep learning network for automatic calculation of left ventricular ejection fraction in 2D echocardiography," *Computers in Biology and Medicine*, vol. 156, p. 106705, 2023. https://doi.org/10.1016/j.compbiomed.2023.106705

[21] G. F. Cacao, D. Du, and N. Nair, "Unsupervised image segmentation on 2D echocardiogram," *Algorithms*, vol. 17, no. 11, p. 515, 2024. https://doi.org/10.3390/a17110515

[22] J. Ding and X. Han, "A semi-supervised approach combining image and frequency enhancement for echocardiography segmentation," *IEEE Access*, vol. 12, pp. 92549–92559, 2024. https://doi.org/10.1109/ACCESS.2024.3408952

[23] P. Naidoo *et al.*, "Consensus-guided evaluation of self-supervised learning in echocardiographic segmentation," *Computers in Biology and Medicine*, vol. 198, p. 111148, 2025. https://doi.org/10.1016/j.compbiomed.2025.111148

[24] A. Chernyshov *et al.*, "Automated segmentation and quantification of the right ventricle in 2-D echocardiography," *Ultrasound in Medicine & Biology*, vol. 50, no. 4, pp. 540–548, 2024. https://doi.org/10.1016/j.ultrasmedbio.2023.12.018

[25] Z. Qian, T. Hu, J. Wang, and Z. Yang, "U-shape-based network for left ventricular segmentation in echocardiograms with contrastive pretraining," *Scientific Reports*, vol. 14, no. 1, p. 29689, 2024. https://doi.org/10.1038/s41598-024-81523-7

[26] S. Nemri and L. Duong, "Automatic segmentation of echocardiographic images using a shifted windows vision transformer architecture," *Biomedical Physics & Engineering Express*, vol. 10, no. 6, p. 065017, 2024. https://doi.org/10.1088/2057-1976/ad7594

[27] S. Yun and J. Choi, "Comparative evaluation of deep learning architectures, including UNet, TransUNet, and MIST, for left atrium segmentation in cardiac computed tomography of congenital heart diseases," *Ewha Medical Journal*, vol. 48, no. 2, p. e33, 2025. https://doi.org/10.12771/emj.2025.00087

[28] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *International MICCAI Brainlesion Workshop*, Cham: Springer International Publishing, 2021, pp. 272–284. https://doi.org/10.1007/978-3-031-08999-2_22

[29] J. Zhang, Q. Qin, Q. Ye, and T. Ruan, "ST-unet: Swin transformer boosted U-net with cross-layer feature enhancement for medical image segmentation," *Computers in Biology and Medicine*, vol. 153, p. 106516, 2023. https://doi.org/10.1016/j.compbiomed.2022.106516

[30] S. Leclerc *et al.*, "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2198–2210, 2019. https://doi.org/10.1109/TMI.2019.2900516

[31] S. Leclerc, "CAMUS: Cardiac acquisitions for multi-structure ultrasound segmentation dataset," *Kaggle*, 2019. [Online]. Available: https://www.kaggle.com/datasets/saeedhamidpour/camus-dataset

[32] R. M. Lang *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," *European Heart Journal-Cardiovascular Imaging*, vol. 16, no. 3, pp. 233–271, 2015. https://doi.org/10.1093/ehjci/jev014

[33] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10012–10022. https://doi.org/10.48550/arXiv.2103.14030

[34] V. Thamilarasi and R. Roselin, "U-NET: Convolution neural network for lung image segmentation and classification in chest X-ray images," *INFOCOMP: Journal of Computer Science*, vol. 20, no. 1, pp. 101–108, 2021. http://177.105.60.18/index.php/infocomp/article/view/1422

[35] G. Gaál, B. Maga, and A. Lukács, "Attention u-net based adversarial architectures for chest x-ray lung segmentation," *arXiv preprint arXiv:2003.10304*, 2020. https://doi.org/10.48550/arXiv.2003.10304

[36] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020. https://doi.org/10.1016/j.isprsjprs.2020.01.013

10 AUTHORS

A. Shamla Beevi is with the Department of Computer Science and Engineering, National Institute of Technology Calicut, Calicut, Kozhikode, India. She is a PhD researcher specializing in artificial intelligence, computer vision, and medical image analysis, with a focus on echocardiographic data for cardiovascular disease detection. Her research includes deep learning–based approaches for image denoising, view classification, left ventricle segmentation, and automated detection of regional wall motion abnormalities. She also serves as Chief Technical Officer, contributing to applied research in AI-driven solutions, cybersecurity, and enterprise systems (E-mail: shamla_p180105cs@nitc.ac.in).

Amit Nagarkoti is currently working at the Central Research Laboratory, Bharat Electronics Limited, where he is engaged in research and development in the field of artificial intelligence. He completed his M.Tech. in Computer Science and Information Security from the National Institute of Technology Calicut. His areas of interest include machine learning, deep learning, and intelligent systems, with a focus on solving real-world problems through AI-driven solutions. Through his work, he aims to contribute to the advancement of AI technologies and their practical applications in industry and research (E-mail: amitnagarkoti111@gmail.com).

Saidalavi Kalady is an Associate Professor in the Department of Computer Science and Engineering at the National Institute of Technology Calicut, India. His research interests include computational intelligence, data mining, and operating systems, with a focus on developing intelligent algorithms and systems for real-world applications (E-mail: said@nitc.ac.in).