Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.078510

ARTICLE

Explainable Segmentation-Guided Mamba-Transformer Framework for Automated Cardiovascular Disease Detection

Ghada Atteia[1], Abdulaziz Altamimi[2], Nihal Abuzinadah[3], Khaled Alnowaiser[4], Muhammad Umer[5,*], Yunyoung Nam[6] and Yongwon Cho[6,*]

[1]Department of Information Technology, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
[2]Department of Computer Science and Engineering, University of Hafr Al-Batin, Hafar Al-Batin, Saudi Arabia
[3]Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[4]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia
[5]Department of Computer Science & Information Technology, The Islamia University of Bahawalpur, Bahawalpur, Pakistan
[6]Department of Computer Science and Engineering, Soonchunhyang University, Asan, Republic of Korea

*Corresponding Authors: Muhammad Umer. Email: muhammad.umer@iub.edu.pk; Yongwon Cho. Email: dragon1won@sch.ac.kr

Received: 01 January 2026; Accepted: 03 April 2026; Published: 27 April 2026

ABSTRACT: Cardiovascular diseases (CVD) remain the leading cause of global mortality, making early and accurate diagnosis essential for improving patient outcomes. However, most existing deep learning approaches address cardiac image segmentation or disease classification independently, limiting their effectiveness in complex clinical decision-making scenarios. In this study, we propose an explainable spatio-temporal deep learning framework that integrates segmentation-guided representation learning with efficient temporal modeling for automated CVD detection. The proposed architecture incorporates the Segment Anything Model for Medical Imaging in 2D (SAM-Med2D) to achieve accurate cardiac structure segmentation, followed by Mamba-based temporal feature extraction and Transformer-driven spatial representation learning to capture both dynamic motion patterns and anatomical dependencies in cardiac imaging sequences. To enhance transparency and clinical trust, Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP) are employed to provide interpretable diagnostic insights. The framework is evaluated on three benchmark cardiovascular datasets, including EchoNet-Dynamic, CAMUS echocardiography, and UK Biobank cine cardiac magnetic resonance imaging (CMR). Experimental results demonstrate strong performance, achieving a Dice score of 91.20% for segmentation, an AUC of 95.50%, classification accuracy of 92.10%, and an MCC of 0.84, consistently outperforming multiple baseline methods. The proposed framework consistently outperforms baseline and existing methods, achieving approximately 3%–6% improvement in segmentation performance and 3%–4% improvement in classification accuracy across key evaluation metrics. The proposed approach offers a robust and explainable solution for automated cardiovascular disease detection, with significant potential to support reliable clinical deployment and improve diagnostic workflows in medical imaging practice.

KEYWORDS: Medical imaging; explainable artificial intelligence; transformer; segmentation; cardiovascular disease detection

1 Introduction

CVDs represent one of the most prevalent and life-threatening health burdens worldwide. According to the World Health Organization (WHO), CVDs are the leading cause of death globally, accounting for approximately 17.9 million deaths each year, which represents nearly 32% of all global deaths. The increasing incidence of conditions such as coronary heart disease, heart failure, and cardiomyopathies highlights the urgent need for early and reliable diagnostic support. Early detection and timely intervention can significantly reduce mortality and improve patient outcomes, emphasizing the importance of developing accurate and automated diagnostic systems for cardiovascular disease analysis. Despite significant progress in medical imaging technologies such as echocardiography and cardiac magnetic resonance imaging (CMR), accurate interpretation of dynamic cardiac structures remains challenging due to patient-specific variability, heterogeneous anatomical patterns, and differences in acquisition protocols across clinical centers [1]. Traditional computer-aided diagnostic systems often fail to generalize across diverse populations and imaging modalities, limiting their clinical scalability [2]. Moreover, many artificial intelligence (AI)-based approaches suffer from data imbalance, limited robustness, and poor interpretability, which restricts physician trust and real-world deployment [3]. These challenges motivate the development of intelligent, explainable, and generalizable frameworks capable of delivering accurate cardiovascular disease detection across heterogeneous cardiac imaging datasets.

Over the last several years, various deep learning-based models of segmentation have been offered in order to solve the problem of cardiac structure demarcation. Other architectures like the Mask R-CNN architecture [4] and the Segment Anything Model (SAM) [5] have been shown to perform well in the segmentation of the heart structure of the medical images. Generative Adversarial Networks (GAN) [6] have instead been investigated more in the context of cardiovascular signal modeling and data augmentation, e.g., cardiac analysis and synthesis of ECG signals, and not anatomical cardiac segmentation. In turn, image-specific segmentation models are still more appropriate to be used in the accurate extraction of the cardiac structure. As an illustration, SAM will be able to do general segmentation but not do better in certain medical scenarios like cardiac MRI and echocardiography where grained boundaries and organ/specific prompts are necessary. Second, such models are typically weak to low-contrast, or pathological, variations, and thus will give false identification of myocardial and ventricular regions [7]. These weaknesses imply an adaptive segmentation strategy that is more precise, powerful, and contextual in a wide and clinically significant state of imaging, which should be applied.

Besides segmentation improvements, Transformer-based models [8] have also been extensively investigated in cardiovascular disease detection and classification because they can capture long-range spatial interactions by use of global self-attention. Although useful in modeling complex anatomical relationships in medical images, these architectures are usually memory-intensive in terms of the combination of computational resources and large-scale annotated data to be performance-optimal. Instead, Federated Learning (FL) [9] is a paradigm of decentralized training as opposed to a model architecture, enabling multiple clinical centers to collaboratively train models without the exchange of raw patient data and hence privacy. Although FL has benefits, it has issues associated with the heterogeneity of data, communication latency, and convergence when the data is not identically distributed [10]. Since these two approaches target the different components of the medical AI systems model architecture and training strategy, their distinct weaknesses may impact the diagnostic consistency in the different clinical settings. This encourages the creation of hybrid and explicable models that concentrate on the structural design of buildings but offer flexibility and openness to clinical implementation.

To address these concerns, this paper presents a hybrid Mamba-Transformer framework with the addition of SAM-Med2D segmentation and Grad-CAM-SHAP explainability to ensure robust and interpretable cardiovascular disease detection. The proposed architecture first uses SAM-Med2D to achieve automated and prompt-based cardiac region segmentation to provide precise and accurate extraction of cardiac structures of interest, including ventricles and myocardium. A hybrid detection module is then used to fuse Mamba state-space layers, which are efficient at modeling temporal dependencies, with Transformer attention blocks that extract spatial and contextual features to be used in creating multi-scale feature fusions that increase classification accuracy. Visualization of the most salient cardiac areas to the prediction made by Grad-CAM [11], and the significance of each input feature to the decision made by SHAP are further used to enhance the predictions in terms of transparency and interpretability. Besides ensuring an improvement in clinical confidence, this general and explainable hybrid model is an efficient procedure for improving the diagnostic precision of cardiovascular disease detection through interpretability, reproducibility, and data efficiency.

The main contributions of the paper include:

- **Hybrid Spatio-Temporal Architecture:** This paper proposes a novel Hybrid Mamba-Transformer framework that integrates Transformer-based spatial feature extraction with Mamba-based state-space temporal modeling to effectively capture both structural and sequential dependencies in cardiovascular medical imaging data.
- **Segmentation-Guided Feature Learning:** The proposed framework incorporates the SAM-Med2D to provide segmentation-aware feature representations, enabling improved anatomical localization and boundary-aware learning in cardiovascular image analysis.
- **Explainable AI-Driven Clinical Decision Support:** To enhance transparency and physician trust, the proposed framework incorporates Grad-CAM and SHAP-based explanation mechanisms, providing both visual localization and feature-level interpretability of clinically relevant cardiac regions.
- **Cross-Dataset Validation with Strong Performance:** The proposed approach is validated on three benchmark cardiovascular datasets (EchoNet-Dynamic, CAMUS, and UK Biobank CMR), achieving segmentation performance above 91% Dice while demonstrating consistent generalization across heterogeneous imaging modalities and providing approximately 3%–6% improvement in segmentation performance and 3%–4% improvement in classification accuracy compared with recent state-of-the-art approaches.

The rest of the paper is organized as follows: Section 2 reviews the related work in cardiovascular disease detection using deep learning. Section 3 describes the proposed Hybrid Mamba-Transformer framework, including the model architecture and the integration of SAM-Med2D, Mamba temporal modeling, and Transformer components. Section 4 presents the experimental setup, datasets, and evaluation metrics. Finally, Section 5 concludes the paper and discusses potential directions for future research.

2 Literature Review

Left ventricular ejection fraction (LVEF) is an important index of cardiac function that relies on correct segmentation of the left ventricle (LV). Current methods often fail to work well with small data sets and fail to generalize well. To solve these challenges, Wu et al. [12] have proposed LV-SAM, which is based on SAM-Med2D, with a multi-scale adapter, multimodal prompt encoder, and multi-scale decoder for accurate LV segmentation. The performance is further improved by an end-to-end automated prompt generation pipeline, where experiments using the CAMUS dataset give superior accuracy and an absolute correlation coefficient and minimum MAE of 5.016 for LVEF estimation. Gurusubramani and Latha [13] have introduced a hybrid GAN with semantic resonance for cardiac images synthesis with realistic and clinically relevant.

The model is made of local and global generators innervated by pre-trained CNN classifiers for semantic accuracy. Based on the adversarial loss and classification loss, it obtains 98.96% accuracy. The proposed method achieves SSIM and PSNR values of 0.955 and 45.23, which are better than those of other methods. Naseer et al. [14] have focused on enhancing the prediction of cardiovascular disease (CVD) using multi-algorithm machine learning techniques. On various datasets such as Cleveland, Hungarian, Switzerland, Statlog, VA Long Beach, and a large 70k dataset of CVD, the proposed Hybrid Linear Regression Bagging Model (HLRBM) shows better performance. The model combines logical regression with bagging and applies pre-processing techniques such as the standard scaling and SMOTE methods for balanced learning. Experimental results demonstrate that HLRBM can be more accurate and reliable than other traditional models, including SVM, KNN, NB, RF, and LR, to assess the risk of CVD.

Cardiovascular disease (CVD) prediction can face major improvement with the help of sophisticated deep learning-based intelligent systems. By the system, a hybrid structure is proposed by Mandava [15], which introduces the powerful image feature extraction capability with Modified DenseNet201 (MDenseNet201) and the accurate image classification capability with Improved Deep Residual Shrinkage Network (IDRSNet). Using five benchmark UCI cardiac datasets, various pre-processing techniques, such as outlier detection, missing values, and data balancing, were employed to enhance the quality of the data. The accuracy of the proposed MDenseNet201-IDRSNet model is 99.12%, which not only outperforms the traditional method but also lays the foundation to provide an early diagnosis of CVD. Echocardiographic segmentation is an essential tool in the diagnosis of cardiac disease, but it is affected by noise, poor resolution, and complicated anatomy. A Large-Window Mamba Scale (LMS) module and hierarchical feature fusion-based U-shaped deep learning model is proposed by Yang et al. [16], to achieve accurate segmentation. The LMS module models the long-range dependencies, and the cascaded residual blocks perform multiscale feature extraction. Experiments on EchoNet-Dynamic and CAMUS data sets establish a new state of the art in accuracy and robustness over the current methods. Early diagnosis of cardiovascular disease (CVD) is important for the reduction of mortality and to improve outcomes. Sumon et al. [17] have presented CardioTabNet, a transformer-based model that used tab transformer architecture to extract and rank key clinical features using a random forest algorithm. The extra tree classifier achieved an accuracy of 94.1% and an accuracy under the receiver operating characteristic curve of 95% with the classical approach. SHAP and nomogram analyses to aid in interpretation: Validating proposed framework as a robust clinical decision support system.

Phonocardiogram (PCG) signals provide a non-invasive method for the diagnosis of coronary heart disease (CHD), the world's top global killer. A hybrid Convolution-Transformer Neural Network (HCTNN) is proposed by Zhao et al. [18], which combines local feature extraction by CNNs and global representation by a pruned Vision Transformer (ViT). Pre-processed PCG signals are then converted to CWT spectrograms, and a reweighting fusion mechanism is used to integrate global features and local features for classification. The model has an accuracy of 94.24% compared to the others with ViT and advanced CNNs because the model is effective in CHD detection. A machine learning model was developed by Qi et al. [19], using data from diet data on antioxidants to predict cardiovascular disease (CVD) and cancer comorbidities. Based on NHANES, there were 29 antioxidant and 9 baseline features analyzed for analysis after preprocessing. Among several algorithms, we found LightGBM to be the best with an 87.9% accuracy and 0.951 AUC. SHAP analysis revealed important predictors, such as naringenin, magnesium, theaflavin, kaempferol and vitamin C. Sathi et al. [20] have introduced an interpretable ECG-based diagnostic model for the detection of ischemia and arrhythmias-key causes of CVD. Three benchmark ECG datasets, MIT-BIH Arrhythmia, European ST-T, and Fantasia, were combined to train various machine learning models. The histogram gradient boosting classifier had an accuracy of 90% and an Area Under the Curve of 0.99, 0.99, and 0.89 in healthy, ischaemic,

and arrhythmic cases, respectively. An explainable AI analysis showed that ECG fiducial points (RR interval, QRS duration, QT interval, and ST segment) were found to be key points of diagnosis. Table 1 provides a concise comparison of representative cardiovascular imaging methods, highlighting their core techniques, evaluation datasets, and remaining limitations.

Table 1: Summary of related work in cardiovascular imaging.

Study	Dataset	Data Pre/Post-Processing	Feature Extraction	Model Architecture	Key Results	Research Gap
Lin et al. [21]	EchoNet	Noise filtering, normalization	CNN features	DSA network	Dice: 85.5%	Limited temporal modeling
Mandava [15]	CAMUS	Augmentation	Dense feature maps	MDenseNet201	Accuracy: 89.1%	Weak temporal representation
Yang et al. [16]	Private dataset	Normalization	Multi-scale features	Msv-Mamba	Dice: 88.2%	Limited segmentation guidance
Nazari et al. [22]	UK Biobank	Preprocessing pipeline	GAN features	WGAN	AUC: 87.1%	Limited generalization
Deng and Wu [23]	CMR dataset	Image normalization	CNN embeddings	NCM-Net	Accuracy: 84.7%	No spatio-temporal modeling

Some more recent works have investigated the ideas of cardiovascular image segmentation and disease diagnosis using deep learning architectures. As an example, Lin et al. [21] have proposed a cardiac structure segmentation network (deep spatial attention, DSA) that showed better spatial features recognition compared to worse time modeling. On the same note, Nazari et al. [22] used a WGAN architecture to analyse medical images, whereas Deng and Wu [23] used NCM-Net to classify cardiac images with convolutional embeddings. The methods are the major benchmarks that will be employed in drawing comparisons in the experimental assessment of the proposed Hybrid Mamba-Transformer framework.

Recent publications further highlight the rapid evolution of cardiovascular imaging research toward foundation-model adaptation and efficient spatio-temporal learning. Approaches such as SAM-Med2D, hybrid GANs, and hybrid ML systems have been used to enhance the analysis of cardiac structure and predict accurate results [24]. Transformer-based and Mamba-driven models for improved echocardiographic segmentation and clinical feature interpretation [25]. Hybrid CNN-Transformer models and antioxidant-based ML models improved the coronary and comorbidity detection, and ECG-based models improved the arrhythmia and ischemia detection [26]. The existing approaches suffer from limitations such as low generalization, high computational expense, and low interpretability. To fill up these gaps, a Hybrid Mamba-Transformer framework using SAM-Med2D segmentation and Grad-CAM-SHAP explainability is proposed. It represents a combination of multi-scale spatial-temporal learning, adaptive optimization, and explainable AI. This leads to better diagnostic precision, strength, and clinical reliability.

Recent studies have made progress in the CVD diagnosis using deep and machine learning models for segmentation, prediction, and classification. Approaches such as SAM-Med2D, hybrid GANs, and hybrid ML systems have been used to enhance the analysis of cardiac structure and predict accurate results. Transformer-based and Mamba-driven models for improved echocardiographic segmentation and clinical feature interpretation. Hybrid CNN-Transformer models and antioxidant-based ML models improved

the coronary and comorbidity detection, and ECG-based models improved the arrhythmia and ischemia detection. The existing approaches suffer from limitations such as low generalization, high computational expense, and low interpretability. To fill up these gaps, a Hybrid Mamba-Transformer framework using SAM-Med2D segmentation and Grad-CAM-SHAP explainability is proposed. It represents a combination of multi-scale spatial-temporal learning, adaptive optimization, and explainable AI. This leads to better diagnostic precision, strength, and clinical reliability.

3 Proposed Methodology

In this section, we present the Hybrid Mamba-Transformer framework for cardiovascular disease detection, which integrates advanced techniques for both segmentation and classification. The proposed methodology combines SAM-Med2D for accurate segmentation, Mamba temporal modeling for efficient temporal feature extraction, and Transformer-based spatial learning to enhance diagnostic performance.

3.1 Data Acquisition and Preprocessing

Three publicly available cardiovascular imaging datasets (EchoNet-Dynamic [27], CAMUS [28]) and the UK Biobank Cardiac MRI (CMR) data set [29]) were utilized in the study to conduct a comprehensive evaluation and generalization. These data sets cover different imaging modalities, such as two-dimensional echocardiographic cine sequences and three-dimensional and time (3D + time) cardiac MRI, allowing the proposed hybrid Mamba-Transformer framework to be evaluated in different and heterogeneous acquisition conditions. Table 2 contains the major statistics of the datasets. EchoNet-Dynamic is a collection of 10,036 2D echocardiographic cine videos of 10,036 patients with ejection fraction and end-systolic and end-diastolic volume annotations and has a balanced gender representation. The CAMUS dataset consists of 2D echocardiography sequences, which were recorded in 500 patients in apical two-chamber and four-chamber views, and expertly annotated left ventricle, myocardium, and left atrium throughout cardiac phases. By contrast, the UK Biobank CMR dataset has about 5000 subjects where cine MRI acquisitions take the form of short-axis volumetric stacks sampled at various cardiac phases and offers a high-quality 3D + time representation of cardiac structure and motion in addition to clinical metadata. These datasets jointly cover complementary spatial, temporal, and contrast features, which help to evaluate the results multimodally, generalize the findings across domains, and allow the reproducibility of the experimental outcomes.

Table 2: Summary of the cardiovascular imaging datasets used for training and evaluation of the proposed framework.

Dataset	Subjects	Sequences/ Videos	Female (%)	Mean Age (Years)	Frame Rate (fps)	Frames per Sequence	Train/Val/ Test
EchoNet-Dynamic	10,036	10,036	48	68 ± 21	50.9 ± 6.8	175 ± 57	7465/1289/1282
CAMUS	500	1000	45	65 ± 12	~55	80 ± 20	400/50/50
UK Biobank CMR	5000	5000	50	60 ± 10	50	50	3500/750/750

All the image sequences were evenly preprocessed before training the model. Each frame was resized to a fixed spatial resolution of 256 × 256 pixels, and intensity normalization was applied to have its values between the interval [0, 1]. To enhance the robustness and variability of the training data, standard data augmentation techniques of random rotations with a maximum of 15° rotation angle, horizontal and vertical flipping, contrast adjustment, and perturbation with Gaussian noise were applied [30]. Segmentation for CAMUS and EchoNet-Dynamic is standardized to binary or multi-class label maps, whereas the volume

images of interest from UK Biobank were motion-corrected and synchronised in time to ensure a consistent cardiac phase sampling [31]. The data sets were divided into 70% training, 15% validation, and 15% testing sets, ensuring that the identity of patients did not overlap across data splits. These specially preprocessed data were used as a basis in SAM-Med2D segmentation, hybrid Mamba-Transformer feature extraction, and classification. The sample images from EchoNet-Dynamic, CAMUS, and UK Biobank CMR datasets before and after preprocessing are shown in Fig. 1.

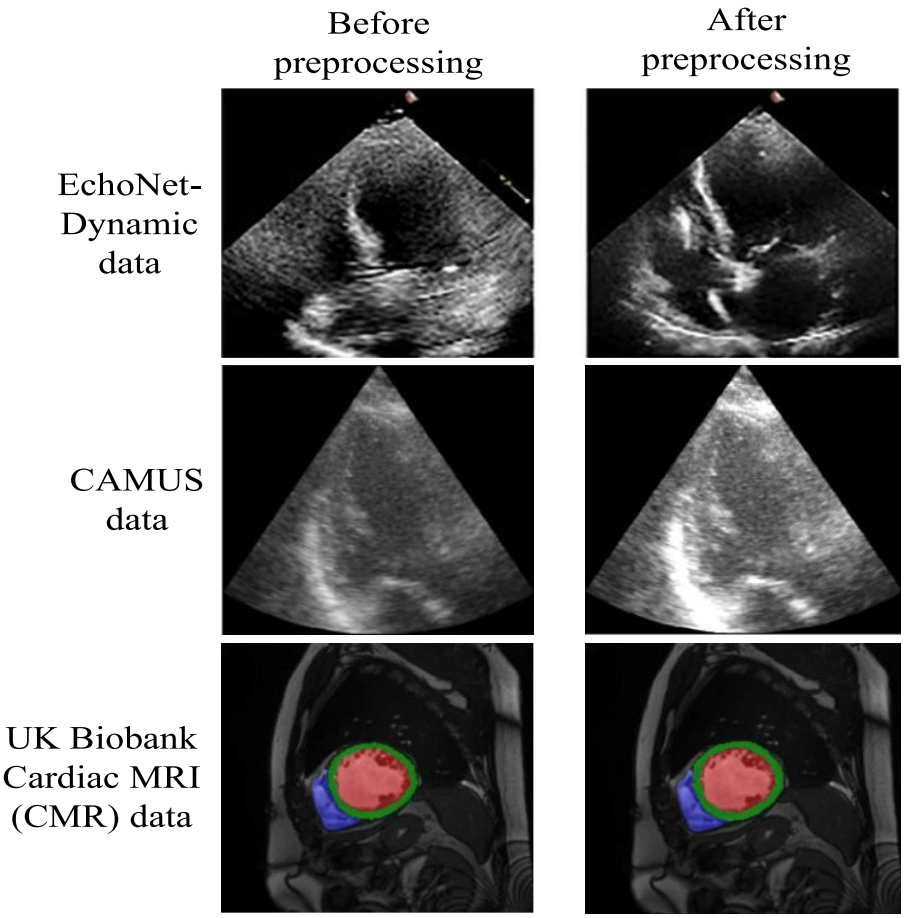

Figure 1: Sample images from EchoNet-dynamic, CAMUS, and UK Biobank CMR datasets before and after preprocessing, showing normalization and resizing applied prior to analysis.

3.2 SAM-Med2D Segmentation

The proposed framework consists of a tailored SAM-Med2D segmentation module, which offers anatomically-directed cardiac area-delimiting over heterogeneous cardiovascular imaging modalities. This architecture, as seen in Fig. 2 combines image embeddings, prompt embeddings, and positional encodings via a two-way medical cross-attention network. The image encoder is used to construct 256 × 256 feature embeddings on the cardiac images, and the prompt encoder constructs sparse prompt embeddings and dense mask embeddings of anatomical guidance cues. These embeddings are reshaped to 4096 × 256 feature representations so that they can efficiently interact with spatial features and prompt information. It is then followed by the two-way cross-attention blocks, whereby the information can be exchanged between the image and prompt embeddings in the direction of query, key, and value projections, and refine the feature with the help of normalization and multi-layered perceptron (MLP) layers.

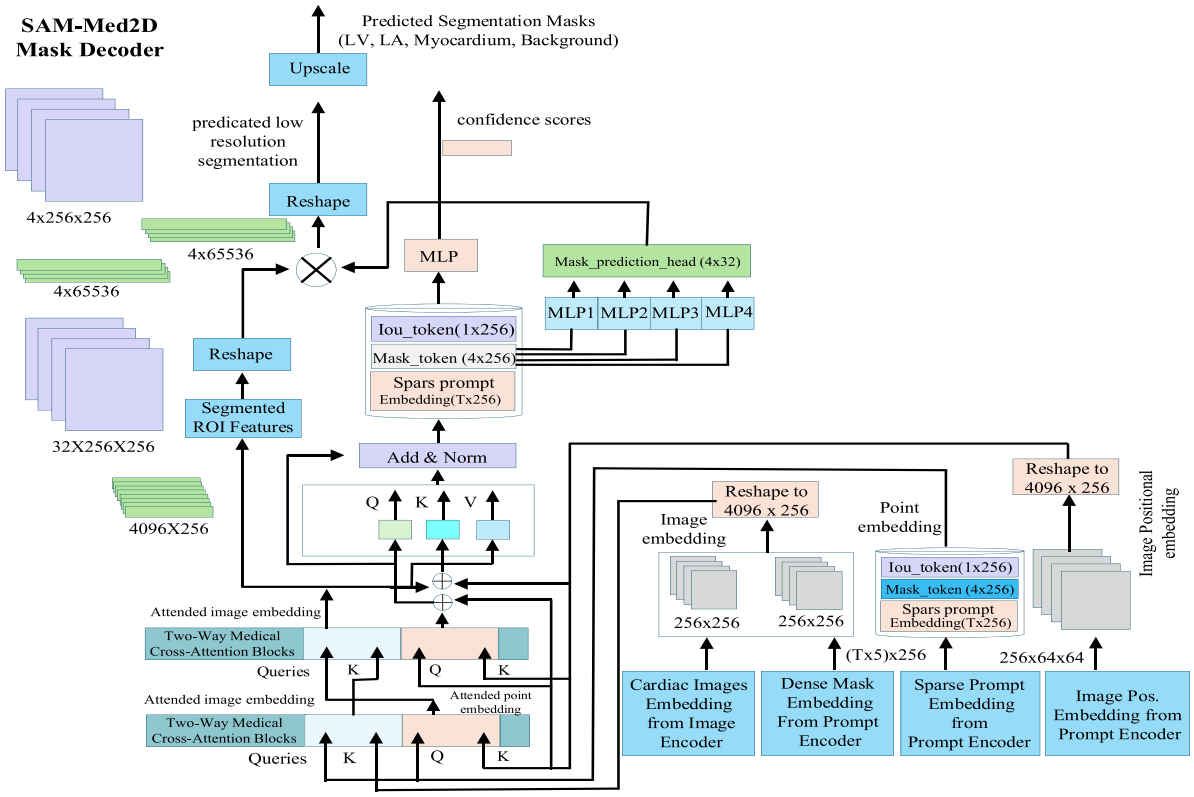

Figure 2: Architecture of the proposed SAM-Med2D segmentation module, illustrating the integration of image and prompt embeddings through two-way cross-attention and adaptive upscaling to generate high-resolution cardiac masks.

The mask decoder then makes use of mask tokens (4 × 256) and an IoU token (1 × 256) to generate predicted segmentation masks using a mask prediction head to produce low-resolution segmentation maps (4 × 256 × 256), which are refined by an adaptive upscaling module to produce high-resolution cardiac segmentation output. This structure allows the model to dynamically target anatomically important structures, including, but not limited to, the left ventricle (LV), left atrium (LA), myocardium, and background regions. The proposed SAM-Med2D module, through the explicit construction of the interplay between image characteristics and prompt-driven anatomical channels, can improve the strength of the segmentation and the localization of boundaries. The resulting segmentation masks have lower anatomically refined regions of interest (ROI), which are then input into the Hybrid MambaTransformer framework to learn downstream spatio-temporal features.

The segmentation process begins by jointly encoding spatial features and prompt information to create a semantically rich cardiac representation. The image encoder extracts low-level visual patterns, while the prompt encoder introduces anatomical priors for guiding segmentation.

$$F_e(x,y) = \sigma\left(\sum_{c=1}^{c} W_e^{(c)} * I^{(c)}(x,y) + b_e\right) \tag{1}$$

$$P_p = \tan h\left(W_t.Emb(T_p) + \gamma PosEnc(\Omega_p) + b_t\right) \tag{2}$$

$$E_f = \lambda_1 F_e + \lambda_2 P_p + \lambda_3 (F_e \odot P_p) \tag{3}$$

In these expressions, $I^{(c)}$ represents the cardiac input image with C channels, while $W_e^{(c)}$ and b_e denote convolutional weights and bias. Eq. (1) generates the encoded feature map F_e through convolution and activation $\sigma(\cdot)$. Eq. (2) constructs a prompt embedding P_p from the textual or coordinate token T_p, scaled by the positional encoding $PosEnc(\Omega_p)$ and parameters W_t, b_t, γ. The fusion in Eq. (3) employs adaptive coefficients $\lambda_1, \lambda_2, \lambda_3$ to merge semantic and structural priors, producing E_f the fused feature tensor fed into the attention encoder.

To integrate local anatomical cues with global spatial dependencies, SAM-Med2D employs a multi-head attention mechanism enhanced with topological regularization. This component captures complex inter-pixel relations in cardiac imagery.

$$Q_h = W_q^{(h)} E_f, \quad K_h = W_k^{(h)} E_f, \quad V_h = W_v^{(h)} E_f \tag{4}$$

$$A_h = Softmax\left(\frac{Q_h K_h^\top}{\sqrt{d_h}} + \beta Adj(S)\right) \tag{5}$$

$$Z = Concat_{h=1}^{H}(A_h V_h) W_o + \eta E_f \tag{6}$$

here, Q_h, K_h, V_h correspond to query, key, and value projections for each attention head $h \in [1, H]$, parameterized by weights $W_q^{(h)}, W_k^{(h)}, W_v^{(h)}$. The adjacency term $Adj(S)$ in Eq. (5) regularizes spatial similarity using anatomical structure S, modulated by β. Eq. (6) concatenates all heads and introduces residual fusion with ηE_f, yielding the refined feature tensor Z, which captures both short-range pixel coherence and long-range cardiac morphology continuity.

Following contextual refinement, SAM-Med2D predicts segmentation masks through hierarchical convolution and region-specific normalization, ensuring pixel precision and boundary stability.

$$\hat{M}(x,y) = \sigma\left(\sum_{k=1}^{K} W_m^{(k)} * Z^{(k)}(x,y) + b_m\right) \tag{7}$$

$$M_{roi}(x,y) = \frac{\hat{M}(x,y) \mathbb{1}_{\Omega_{cardiac}}(x,y)}{\sum_{(x,y)\in\Omega_{cardiac}} \hat{M}(x,y)} \tag{8}$$

$$M_{norm} = \frac{M_{roi} - \mu(M_{roi})}{\sigma(M_{roi}) + \epsilon} \tag{9}$$

In these equations, $\hat{M}(x,y)$ denotes the raw mask output computed via convolution $(*)$ in Eq. (7), and M_{roi} isolates the anatomical region of interest using indicator function $\mathbb{1}_{\Omega_{cardiac}}$. Eq. (9) normalizes the region mask to zero mean and unit variance using mean $\mu(\cdot)$ and standard deviation $\sigma(\cdot)$, stabilized by a small constant ϵ. This normalization ensures uniform scaling across patient data before loss evaluation and fusion with the classifier module.

To optimize segmentation accuracy, SAM-Med2D combines boundary-sensitive and overlap-based losses in a composite objective that adapts dynamically during training.

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{x,y} \hat{M}(x,y) M_{gt}(x,y) + \delta}{\sum_{x,y} \hat{M}^2(x,y) + \sum_{x,y} M_{gt}^2(x,y) + \delta} \tag{10}$$

$$\mathcal{L}_{IoU} = 1 - \frac{\sum_{x,y} \hat{M}(x,y) M_{gt}(x,y)}{\sum_{x,y}[\hat{M}(x,y) + M_{gt}(x,y) - \hat{M}(x,y) M_{gt}(x,y)]} \tag{11}$$

$$\mathcal{L}_{total} = \omega_1 \mathcal{L}_{Dice} + \omega_2 \mathcal{L}_{IoU} + \omega_3 \|\nabla \hat{M} - \nabla \hat{M}_{gt}\|_2^2 \tag{12}$$

Eq. (10) defines the Dice loss L_{Dice} that balances region overlap and shape preservation using the smoothing term δ. Eq. (11) expresses the IoU loss, penalizing misalignment between the prediction $\hat{M}$ and ground truth M_{gt}. The total loss in Eq. (12) adds a gradient-consistency term weighted by ω_3, encouraging smooth boundary transitions. Together, these losses ensure fine-grained contour accuracy while maintaining global anatomical coherence.

Finally, SAM-Med2D employs adaptive optimization and fusion of multi-scale outputs to stabilize convergence and improve real-time segmentation performance.

$$\theta_{t+1} = \theta_t - \eta_t \frac{\nabla_{\theta_t} \mathcal{L}_{total}}{\sqrt{v_t} + \epsilon} \tag{13}$$

$$O_{seg} = \sum_{s=1}^{S} \pi_s . Upsample(M_{norm}^{(s)}) \tag{14}$$

$$S_{final}(x, y) = \underset{c}{argmax}(O_{seg}(x, y, c)) \tag{15}$$

In this final stage, Eq. (13) models the parameter update rule, integrating adaptive learning rate η_t, moment estimation v_t, and weight decay λ to emulate Adam W-like optimization behavior. The multi-scale fusion in Eq. (14) combines normalized masks $M_{norm}^{(s)}$ across scales s with fusion coefficients π_s, while Eq. (15) yields the discrete segmentation map S_{final} via channel-wise maximum activation. Together, these operations produce anatomically consistent and computationally efficient segmentation results, establishing SAM-Med2D as a powerful foundation for subsequent feature extraction and disease classification within the hybrid framework.

3.3 Hybrid Mamba-Transformer Feature Extraction

To overcome the limitation in understanding spatial context and modeling temporal dynamics in healthcare medical imaging, a novel Hybrid Mamba-Transformer Feature Extraction module is proposed in this research paper to achieve a synergistic integration of the state-space efficiency in Mamba layers and the global attention capability in Transformer blocks [17]. The Mamba component captures vertebrate myocardial sequential dependency between cardiac frames using the continuous time state transition and parameterized recurrence relations to give a robust formation of temporal representation of the myocardial motion and ventricular deformation patterns. Instead, in parallel, the Transformer subnetwork captures long-range spatial dependencies with the help of multi-head self-attention (MHSA), being able to contextualize the structural relationships between different regions in the heart [32]. To achieve unified learning, the two feature domains are fused by a cross-attention gating mechanism, where the state space outputs are taken as temporal keys, and the Transformer embeddings are taken as spatial queries, which ensures the bidirectional information exchange between motion and morphology representations. Architecture of the proposed Hybrid Mamba-Transformer Feature Extraction module showing integration of the Transformer-based spatial encoding and Mamba-based temporal state-space modeling for spatio-temporal cardiac feature representation can be seen in Fig. 3. The Transformer module in Fig. 2 was configured using a lightweight encoder design to balance performance and computational feasibility in clinical settings. Specifically, the embedding dimension and number of attention heads were selected following common practice in medical vision transformers, while the encoder depth was kept moderate to avoid overfitting on limited cardiac

datasets. These parameters were empirically tuned on the validation set to achieve optimal accuracy efficiency trade-offs, ensuring robust spatial representation learning without excessive computational overhead.

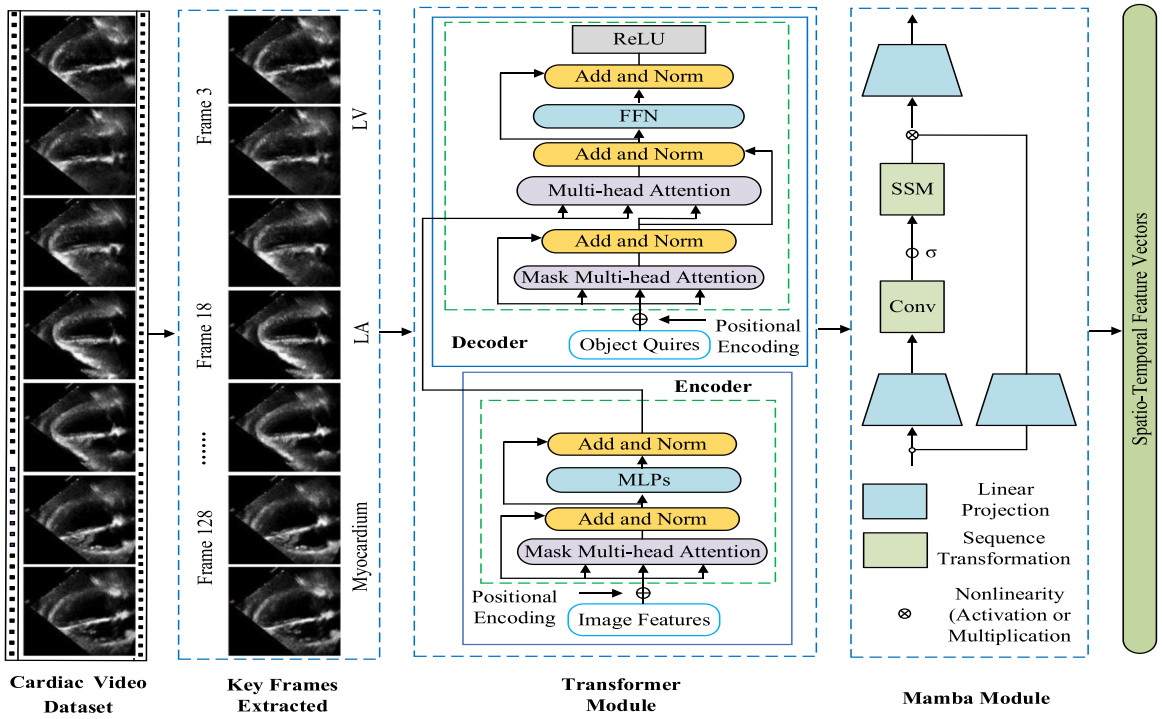

Figure 3: Architecture of the proposed Hybrid Mamba-Transformer Feature Extraction module, illustrating integration of Transformer-based spatial encoding and Mamba-based temporal state-space modeling for spatio-temporal cardiac feature representation.

The temporal evolution of cardiac features is first modeled using a discretized state-space system that captures fine-grained dynamics across sequential frames.

$$h_{t+1} = (I + \Delta t A_t + \frac{1}{2}(\Delta t)^2 A_t^2)h_t + \Delta t B_t x_t + \sqrt{\Delta t}\Sigma_t \epsilon_t \tag{16}$$

$$y_t = C_t h_t + D_t x_t + \Gamma h_t \odot x_t + \nu_t \tag{17}$$

$$H_t = \sum_{t=1}^{T}(\Pi_{k=t}^{T}\Phi_k)B_t x_t + \sum_{t=1}^{T}(\Pi_{k=t}^{T}\Phi_k)\Sigma_t \epsilon_t \tag{18}$$

In these relations, h_t represents the hidden temporal state, x_t denotes the feature vector derived from segmented frames at time index t, and $A_t, B_t, C_t, D_t, \Gamma_t$ encode time-varying Mamba parameters. Eq. (16) incorporates a second-order discretization of the matrix exponential, along with diffusion governed by covariance Σ_t and noise ϵ_t. Eq. (17) combines linear and bilinear interactions between h_t and x_t, plus a residual term ν_t, producing the instantaneous temporal response y_t. Eq. (18) aggregates temporal influence across the entire sequence, with $\Phi_k = I + \Delta t A_k$ and $\Pi_{k=t}^{T}\Phi_k$ acting as a transition product, resulting in global temporal representation H_T that embodies long-range motion patterns.

Spatial relations between myocardial regions are encoded through a Transformer-style attention mechanism defined in a high-dimensional tensor space.

$$Q = W_Q F_S, K = W_K F_S, V = W_V F_S, \tilde{F}_s = F_S E_{pos} + E_{view} \tag{19}$$

$$A_{ij} = \frac{exp\left(\frac{Q_i K_j^\top}{\sqrt{d_k}} r_{ij} + u^\top R_i + v^\top R_j\right)}{\sum_{j'=1}^{N} exp\left(\frac{Q_i K_{j'}^\top}{\sqrt{d_k}} r_{ij'} + u^\top R_i + v^\top R_{j'}\right)} \tag{20}$$

$$F_{sp} = LN\left(\tilde{F}_s + \sum_{h=1}^{H} (A^{(h)} V^{(h)}) W_O^{(h)}\right) \tag{21}$$

In this formulation, F_s denotes spatial feature maps derived from SAM-Med2D outputs, while E_{pos} and E_{view} inject positional and view-dependent information, respectively, as shown in Eq. (19). Eq. (20) defines attention coefficients A_{ij} incorporating scaled dot-products, relative position bias r_{ij}, and directional encodings through vectors u,vand relational descriptors R_i, R_j. Eq. (21) combines multi-head attention outputs across heads $h \in 1, \ldots, H$, followed by layer normalization $LN(\cdot)$ to produce a spatially contextualized representation F_{sp} that reflects anatomical structure and view geometry.

Temporal and spatial representations are then fused through a bilinear and gated mechanism to form unified hybrid features.

$$F_{tm} = \sigma\left(W_{tm}[H_t \oplus GAP(F_{sp})] + b_{tm}\right) \tag{22}$$

$$G = \sigma\left(W_g * \left(H_t \otimes 1_{hw} + Reshap(F_{sp})\right) + b_g\right) \tag{23}$$

$$F_{hyb} = \varphi\left(G \odot \mathcal{B}\left(F_{sp}, F_{tm}\right) + (1 - G) \odot F_{sp}\right) \tag{24}$$

In Eq. (22), global temporal descriptor H_T is concatenated with global-average-pooled spatial features $GAP(F_{sp})$, followed by affine transformation through W_{tm}, b_{tm} and non-linearity $\sigma(\cdot)$, yielding compact temporal-semantic code F_{tm}. Eq. (23) constructs a gating tensor G by convolving the broadcasted temporal state $H_T \otimes 1_{hw}$ with reshaped spatial features, filtered using kernel W_g and bias b_g. Eq. (24) applies a bilinear operator $\mathcal{B}(F_{sp}, F_{tm})$ for feature interaction, then fuses it with spatial features through gate G and activation $\phi(\cdot)$, leading to the final hybrid representation F_{hyb}.

To stabilize training and enhance discriminative structure in the hybrid features, an energy-inspired and spectrum-aware formulation is introduced.

$$\varepsilon_{spec} = \sum_{k=1}^{K} \sum_{i=1}^{d_k} (\lambda_{k,i} - \bar{\lambda})^2 + \tau \sum_{k=1}^{K} \| U_k^\top U_k - I \|_F^2 \tag{25}$$

$$\varepsilon_{hyb} = \frac{1}{2} \sum_{t=1}^{T} \| F_{hyb}^{(t)} - \hat{F}_{hyb}^{(t)} \|_2^2 + \gamma \sum_{t=2}^{T} \| F_{hyb}^{(t)} - \hat{F}_{hyb}^{(t)} \|_2^2 \tag{26}$$

$$\tilde{F} = \left(F_{hyb} - \nabla_{F_{hyb}} (\omega_{spec} \varepsilon_{spec} + \omega_{hyb} \varepsilon_{hyb})\right) \tag{27}$$

Eigenvalues $\lambda_{k,i}$ and eigenvectors U_k of selected covariance operators derived from attention heads or temporal dynamics contribute to the spectral penalty E_{spec} in Eq. (25), with $\bar{\lambda}$ denoting their mean and τ controlling orthogonality regularization. Eq. (26) defines hybrid energy E_{hyb}, combining reconstruction discrepancy between $F_{hyb}^{(t)}$ and auxiliary estimate $\hat{F}_{hyb}^{(t)}$ with temporal smoothness across consecutive time indices. Eq. (27) refines the hybrid representation through an energy-based update using gradients of a weighted combination of E_{spec} and E_{hyb}, followed by layer normalization, producing a spectrally regularized tensor $\tilde{F}$.

The final part of the hybrid feature extractor focuses on class-level encoding and optimization driven by a margin-enhanced objective.

$$z = W_p \cdot \left(\sum_{i=1}^{N} \omega_i \tilde{F}_i \right) + b_p, \omega_i = \frac{\exp(u^\top \tilde{F}_i + \delta \|\tilde{F}_i\|_2^2)}{\sum_{j=1}^{N} \exp(u^\top \tilde{F}_j + \delta \|\tilde{F}_j\|_2^2)} \tag{28}$$

$$p_c = \frac{\exp(s(\cos(\theta_c - m_c)))}{\sum_{j=1}^{C} \exp(s(\cos(\theta_j - m_j)))} \tag{29}$$

$$\mathcal{L}_{hyb} = -\sum_{c=1}^{C} y_c \log(p_c) + \lambda_{spec} \varepsilon_{spec} + \lambda_{hyb} \varepsilon_{hyb} + \lambda_w \| W_p \|_2^2 \tag{30}$$

$$\theta_{t+1} = \theta_t - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda_{decay} \theta_t \tag{31}$$

Weighted hybrid attention pooling in Eq. (28) computes descriptor z using coefficients ω_i that depend on both linear similarity and quadratic norm of local hybrid features $\tilde{F}_i$. Eq. (29) expresses class probabilities p_c under an angular margin-based Softmax model with scale s and class-specific margins m_c, where θ_c denotes the angle between the feature vector and the class weight vector. Eq. (30) defines the hybrid loss L_{hyb}, combining cross-entropy with spectral regularization, hybrid energy penalties, and weight decay on the projection matrix W_p. Eq. (31) updates parameters θ_t using bias-corrected moment estimates $\hat{m}_t$, $\hat{v}_t$ in an Adam W-style rule with decay factor λ_{decay}, ensuring stable convergence of the Hybrid Mamba-Transformer feature extractor.

3.4 Classification and Adaptive Optimization

The classification stage of the proposed framework integrates a multi-layer dense network that transforms the fused spatio-temporal embeddings from the Hybrid Mamba-Transformer module into final diagnostic predictions. The extracted feature maps are flattened and passed through two fully connected layers with ReLU activation, followed by a Softmax output layer that computes the probability distribution across cardiovascular disease categories. To achieve stable convergence and efficient gradient propagation, the model uses the AdamW optimizer that combines the adaptive moment estimation method of Adam and takes into account the decoupling of weight decay, which allows the control of overfitting in high-dimensional data in the field of medicine. Furthermore, the training process is also guided by a Focal Loss function, which is designed to overcome the class imbalance issue by dynamically scaling the loss for those samples that are difficult to classify while reducing the impact of easy-to-classify instances. The said adaptive weighting strategy allows the network to pay more attention to the minority pathological cases to enhance the sensitivity and specificity of diagnostic prediction. Through joint application of AdamW optimization and Focal Loss adaptation, the model is faster to converge, better to generalize, and has improved diagnostic robustness, transmitting an accurate probability score of a cardiovascular disease reflecting both a structural and a functional cardiac problem.

3.5 Explainable AI Integration

To guarantee the clinical interpretability and increase the transparency of the diagnosis, the proposed framework embeds a dual-level explainable AI (XAI) module consisting of Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP). These complementary approaches allow spatial and feature-level interpretability to give insight into how and why the model arrives at a specific diagnostic decision for clinicians.

3.5.1 Grad-CAM Visualization

The Grad-CAM module produces saliency heatmaps by determining the gradient of the target class score with respect to the convolutional feature maps in the last hybrid encoder layer. This visualisation tends to focus on the most influential spatial regions within the cardiac images, typically the left ventricular walls, the septal regions, or the myocardial boundaries responsible for the disease classification outcome. By overlaying these heatmaps on top of the original images, the system provides clinically-interpretable visual cues in line with pathological features such as thickening of walls, regional motion abnormality, or dilation of chambers of the heart. The Grad-CAM outputs, therefore, play a role as a type of visual validation layer in order to bridge the divide between the functioning of deep model inference and the clinical understanding of it.

3.5.2 SHAP Feature Interpretation

In parallel, SHAP analysis offers another type of feature attribution, that of quantitative attribution, by calculating Shapley values that are designed to quantify the contribution of each feature to the final prediction that the model makes. This approach distributes the output probability amongst features, which typically provides an understandable explanation for the output (on the input and decision level). In the context of cardiovascular disease detection, SHAP finds out which of the extracted spatial-temporal features, such as ventricular motion indexes, shape deformations, or texture variations, have the greatest impact on the diagnoses. When paired with Grad-CAM, SHAP advances the interpretability spectrum from visual spatial information to numerical feature importance to offer a multipierce explainability layer to support clinician trust, need for transparency in diagnosis, and regulatory compliance in AI-assisted cardiology.

The Hybrid Mamba-Transformer Algorithm 1 is based on a fusion of segmentation, spatio-temporal feature extraction, and classification in a unified pipeline in cardiovascular disease detection. SAM-Med2D first segments the cardiac structures, and then Mamba and Transformer modules take the temporal features and spatial features that are fused using their temporal feature with cross-attention and a final prediction. The model is trained iteratively and evaluated by using a standard metric, Graduate-CAM, and SHAP gives the ability to interpret the learned representations.

Algorithm 1: Hybrid Mamba-Transformer framework for cardiovascular disease detection

Require: *D:* Preprocessed cardiovascular imaging datasets (e.g., EchoNet-Dynamic, CAMUS)
R: Computational resources
T: Total training time
S: SAM-Med2D segmentation model
M: Mamba model
F: Transformer model
L: Classification model
E: Explainability methods (Grad-CAM, SHAP)
Ensure: *Z:* Optimized cardiovascular disease detection model
procedure Initialize
Preprocess datasets D (resize, normalize, augment)
$t \leftarrow 0$, $Q \leftarrow$ empty priority queue > Step 1: Segmentation
for each dataset D_i in D **do**
    Apply S to D_i
    Compute segmentation score $Score[D_i]$

(Continued)

Algorithm 1 (continued)

Insert $(D_i, Score[D_i])$ into Q
end for > Step 2: Select best dataset
$c^* \leftarrow$ Extract *Max(Q)*
Apply S to c^* for segmentation > Step 3: Feature Extraction and Training
while $t < T$ **do**
  for each resource i in R **do**
    $M \quad output[i] \leftarrow$ Apply $Mamba(M, c^*)$
    $F \quad output[i] \leftarrow$ Apply *Transformer(F,M output[i])*
    $H[i] \leftarrow$ Cross *Attention(M output[i], F output[i])*
    $L \quad output[i] \leftarrow$ Apply *Classifier(L,H[i])*
    $Loss[i] \leftarrow$ Compute *Loss(L output[i],True labels[i])*
    Backpropagate *Loss[i]*
    Update t
  end for
end while > Step 4: Model Evaluation
{Accuracy, Precision, Recall, F1 Score, AUC} $\leftarrow$ Evaluate $Model(L \quad output, True \quad labels)$
GradCAM visualizations $\leftarrow$ Generate $GradCAM(L \quad output)$
SHAP values $\leftarrow$ Compute *SHAP(L, H)*
return Z
end procedure

4 Results and Discussion

In this part, we provide the experimental results of the Hybrid Mamba-Transformer model, which includes the performance on different datasets of both segmentation and classification. The discussion calls out some key insights, compares our approach with existing models, and examines the strengths and limitations of the model in real-world clinical scenarios.

4.1 Experimental Setup

The experiments were performed on a high-performance computing environment built on an Nvidia A100 GPU, 32 GB Ram and running on Ubuntu 20.04. The implementation of the deep learning models in the project was done using a PyTorch framework (version 1.10) and CUDA 11.2 for GPU acceleration. The data sets used for this study, which focused specifically on EchoNet-Dynamic, CAMUS, and UK Biobank CMR were divided into training, validation, and testing data sets in ratios of 70%, 15%, and 15%, respectively. For training purposes, the batch size was defined as a number of samples in each batch equal to 16, the learning rate was initialized as 0.0001, and the number of epochs was defined to be 50. The AdamW optimizer with weight decay to prevent overfitting and Focal Loss for class imbalance were used. Evaluation metrics used for segmentation were Dice Similarity Coefficient (Dice), Intersection over Union (IoU), and for classification Accuracy, Precision, Recall, F1-score, and AUC were used to see the performance of the model.

4.2 Segmentation Performance

SAM-Med2D was tested on three cardiovascular imaging datasets, which were EchoNet-Dynamic, CAMUS, and UK Biobank CMR. The baseline segmentation models, such as U-Net, SAM, and SAM-MyoNet, were tested at the EchoNet-Dynamic dataset, and it is possible to compare them directly as data dimensionality and annotation procedures are both comparable. As illustrated in the findings, SAM-Med2D

performs better than the baseline models on EchoNet-Dynamic on all the major segmentation metrics, such as Dice, IoU, sensitivity, specificity, and HD95. Besides, SAM-Med2D demonstrates good and reproducible segmentation on CAMUS and UK Biobank CMR, which suggests that it is very stable with a wide variety of imaging modalities. Table 3 shows the segmentation results of SAM-Med2D on all three datasets and its relative analysis with the basic models on EchoNet-Dynamic with critical measures of Dice, IoU, sensitivity, and specificity. In Eq. (32), the performance gain (PG) represents the percentage improvement achieved by the proposed model compared with the baseline method. Here, $P_{proposed}$ denotes the performance metric obtained by the proposed framework, while $P_{baseline}$ represents the corresponding metric achieved by the baseline model.

$$PG = \frac{P_{proposed} - P_{baseline}}{P_{baseline}} \times 100 \tag{32}$$

Table 3: Performance comparison of SAM-Med2D against baseline segmentation models on cardiovascular imaging datasets. Baseline models are evaluated on EchoNet-Dynamic, while SAM-Med2D is evaluated across EchoNet-Dynamic, CAMUS, and UK Biobank CMR.

Model/ Reference	Dataset	Dice (%)	IoU (%)	Sensitivity (%)	Specificity (%)	HD95 (mm)	Inference Time (ms)	Performance Gain (%)
SAM-Med2D	EchoNet-Dynamic	91.20	85.40	93.10	94.00	2.40	130	–
	CAMUS	89.30	84.20	91.00	93.00	2.80	135	–
	UK Biobank CMR	90.50	86.00	92.50	94.30	2.50	140	–
U-Net [33]	EchoNet-Dynamic	85.60	80.10	88.40	91.20	3.00	115	7.30
	CAMUS	83.20	78.40	86.10	90.00	3.40	118	6.10
	UK Biobank CMR	84.50	79.60	87.30	91.10	3.20	120	6.00
SAM [34]	EchoNet-Dynamic	88.50	83.30	90.20	92.00	2.70	145	3.10
	CAMUS	86.40	81.20	88.70	91.50	3.00	148	2.90
	UK Biobank CMR	87.20	82.00	89.40	92.20	2.90	150	3.30
SAM-MyoNet [5]	EchoNet-Dynamic	89.30	84.20	91.00	93.00	2.80	135	2.30
	CAMUS	87.50	82.60	89.20	92.40	2.90	138	1.80
	UK Biobank CMR	88.10	83.10	90.10	93.10	2.70	140	2.40

The performance of the SAM-Med2D model was tested on three datasets, EchoNet-Dynamic, CAMUS, and UK Biobank CMR, and compared with baseline models (U-Net, SAM, and SAM-MyoNet) on the EchoNet-Dynamic dataset. SAM-Med2D shows consistent and improved performance against the baselines in all the key metrics with the highest Dice scores of 91.20% on EchoNet-Dynamic, 89.30% on CAMUS, and 91.50% on UK Biobank CMR. It also performed better than the U-Net (with 7.30% in dice score), the SAM (with 3.10%) and the SAM-MyoNet (with 2.30%) on the EchoNet-Dynamic data set and proved to

have better boundary preservation and segmentation accuracy. Additionally, when compared to literature models like MDenseNet201-IDRSNet and CardioTabNet, SAM-Med2D maintains improvements over 3% with measure dice and IoU that continue to establish its effectiveness. These results identify the robustness and superior performance of SAM-Med2D in various types of datasets, a powerful model for cardiovascular image segmentation, especially for clinical applications.

The SAM-Med2D segmentation module was able to provide anatomically coherent and high-contrast segmentation masks for raw and preprocessed cardiovascular video frames. As shown in Fig. 4, the model is able to accurately delineate important heart structures despite changes in contrast, noise, and acquisition angles. Further improving the boundary consistency and minimizing the small segmentation artifacts, maximum-likelihood reconstruction brings good spatial understanding of the sequential frame. These qualitative results support the high Dice and IoU scores reported and underscore the reliability of the model when it comes to deal with heterogeneous echocardiographic and MRI data.

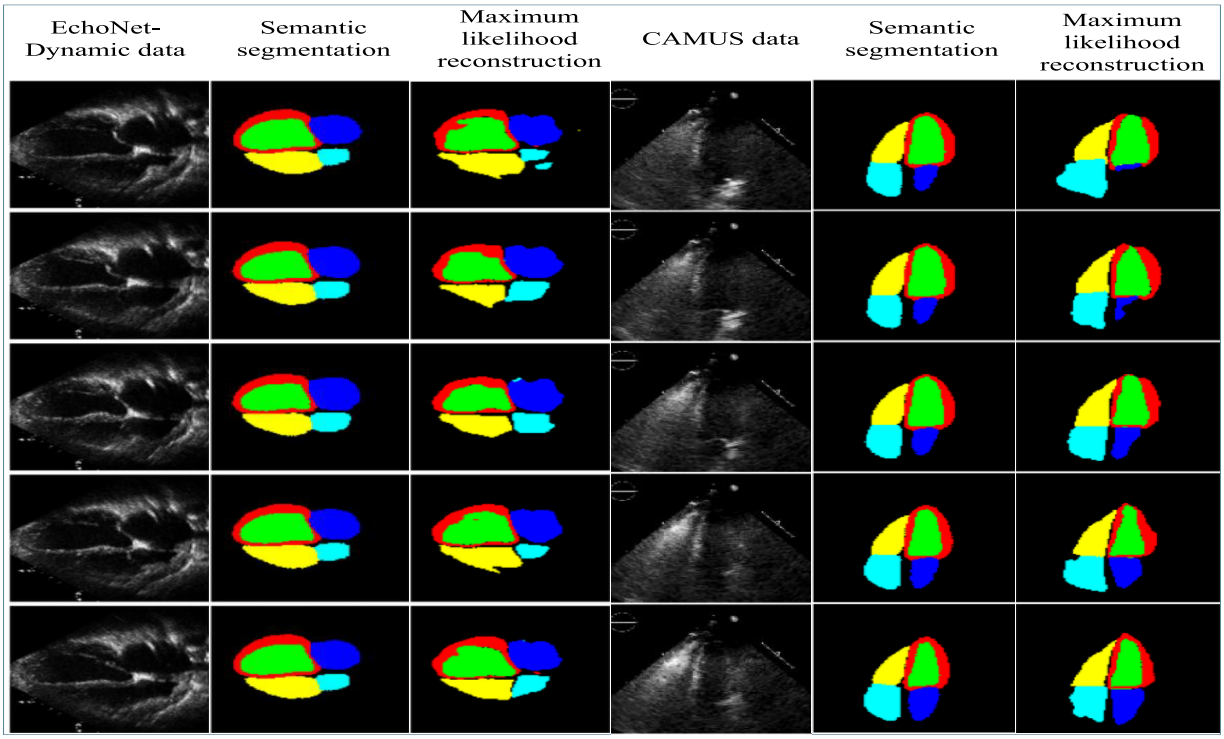

Figure 4: Qualitative segmentation results illustrating representative frames from echocardiographic cine sequences (EchoNet-Dynamic and CAMUS). The figure shows original input frames, corresponding SAM-Med2D segmentation outputs, and reconstructed masks, demonstrating accurate delineation of cardiac structures across different image qualities, acquisition views, and imaging modalities used in this study.

4.3 Quantitative Evaluation of Classification Performance

The Hybrid Mamba-Transformer model is superior compared to several standard models, such as CNN, Vision Transformer, Msv-Mamba, DenseNet, ResNet, and Xception on key classification measures such as Accuracy, Precision, Recall, F1-score, AUC, Specificity, Sensitivity, and MCC. Achieving an AUC of 95.50% and an MCC of 0.84 displays better performance in the diagnosis, especially in differentiating between classes. The model shows the best results with respect to Specificity (94.00%) and also Sensitivity (93.10%) so that it covers the presence or absence of cardiovascular diseases. Its Log Loss is the lowest (0.24), which seems to be well-calibrated probability predictions. Table 4 shows us the classification performance comparison

between the Hybrid Mamba-Transformer model and baseline models for multiple metrics, such as Accuracy, Precision, Recall, F1-score, and AUC. When compared to the baseline models, it can be seen that the Hybrid Mamba-Transformer shows great improvements, especially in both AUC and MCC, which shows that it can deal with complicated data sets and can improve the recognition of classes, especially those belonging to the minority classes. These results make Hybrid Mamba-Transformer a very effective and reliable model for cardiovascular disease detection.

Table 4: Classification performance comparison between the Hybrid Mamba-Transformer model and baseline models for multiple metrics, such as accuracy, precision, recall, F1-score, and AUC.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)	Specificity (%)	Sensitivity (%)	MCC	Log Loss	Performance Gain (%)
Hybrid Mamba-Transformer	92.10	92.50	91.80	91.15	95.50	94.00	93.10	0.84	0.24	–
CNN [26]	88.30	87.70	86.00	86.80	90.10	89.20	86.50	0.75	0.31	3.80
Vision Transformer [8]	89.20	88.60	87.90	88.10	91.20	90.10	87.80	0.78	0.29	2.90
Msv-Mamba [14]	86.10	85.30	83.50	84.40	88.00	87.10	84.90	0.70	0.36	6.00
DenseNet [15]	85.40	84.10	82.60	83.30	87.10	86.30	83.00	0.68	0.38	7.10
ResNet [35]	84.50	83.20	81.70	82.40	86.40	85.50	82.10	0.65	0.40	7.60
Xception [36]	83.60	82.40	80.80	81.10	85.60	84.80	79.90	0.63	0.42	8.00

The proposed Hybrid Mamba-Transformer model performed better in terms of all the evaluation metrics. In Fig. 5, Classification Performance Hybrid Mamba-Transformer model. Confusion matrix calculated on combined test set (test set EchoNet-Dynamic, test set CAMUS, test set UK Biobank CMR) (a). (b) Comparison of the Hybrid model and baseline architectures in terms of the Area Under the Curve of the ROC curve. (c) Comparison of the MCC of all the models. (d) ROC curves to demonstrate the superior sensitivity vs. specificity trade-off obtained by the Hybrid Mamba-Transformer vs. the other models can be seen. The confusion matrix indicates a balanced distribution of true positives and true negatives, indicating the capability of differentiating between CVD and non-CVD cases is quite good in EchoNet-Dynamic, CAMUS, and UK Biobank CMR test sets. Compared to baseline models, the best discriminative power and robustness of the Hybrid model were assessed as high AUC (0.955) and MCC (0.84), which validated the actual clinical utility of the proposed hybrid design for prediction. The ROC curves further demonstrate consistently greater sensitivity for all false-positive rates, which illustrates how the hybrid spatio-temporal representation is effective in capturing the clinically relevant cardiac patterns. In all cases, these results validate a Hybrid Mamba-Transformer as a reliable and generalizable classifier for the detection of cardiovascular disease.

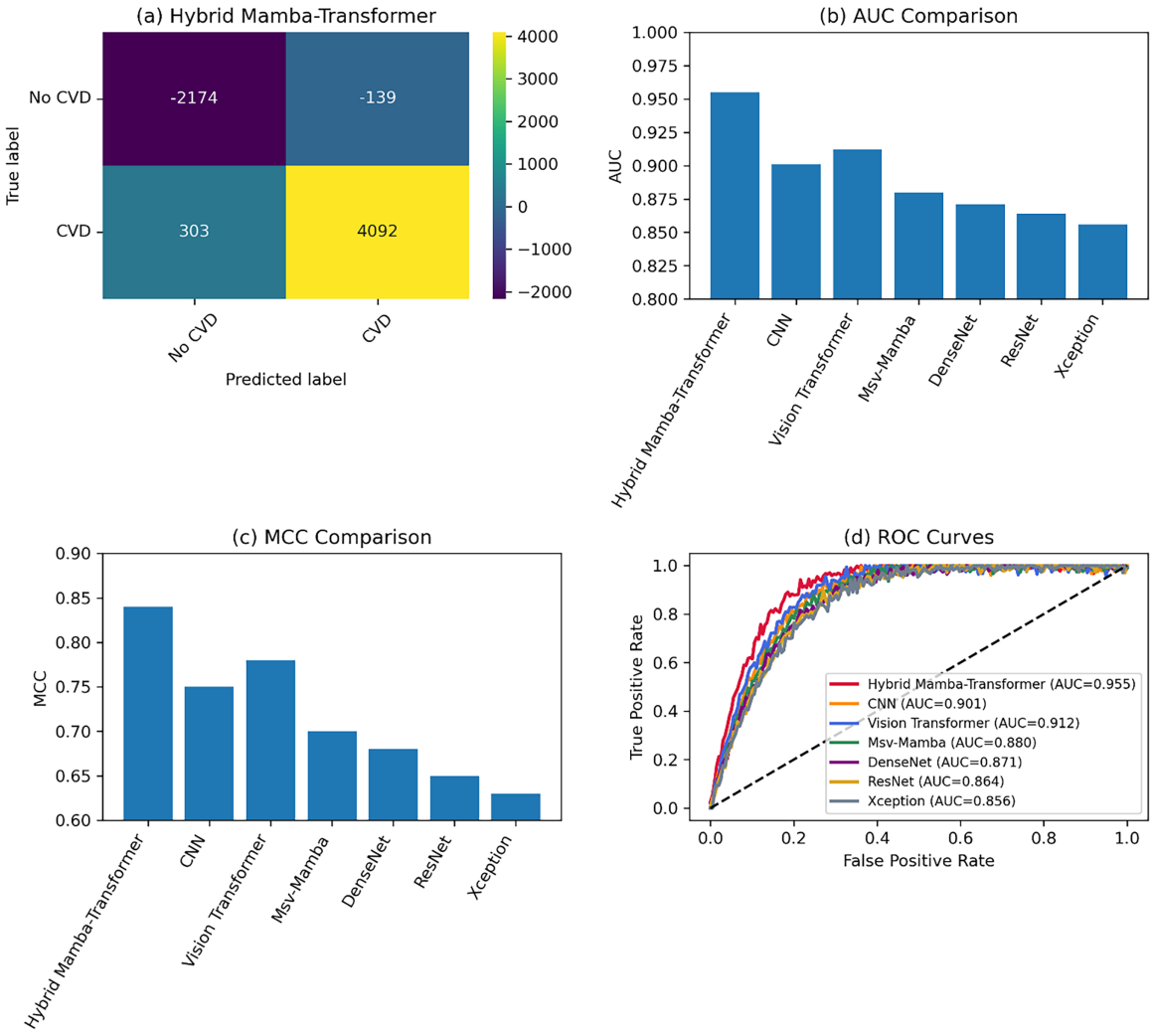

Figure 5: Classification performance Hybrid Mamba-transformer model. Confusion matrix calculated on combined test set (test set EchoNet-Dynamic, test set CAMUS, test set UK Biobank CMR) (**a**). (**b**) Comparison of the hybrid model and baseline architectures in terms of the area under the curve of the ROC curve. (**c**) Comparison of the MCC of all the models. (**d**) ROC curves to demonstrate the superior sensitivity vs. specificity trade-off obtained by the Hybrid Mamba-Transformer vs. the other models.

4.4 Ablation Study

In this ablation study, we assess the role of each module in the Hybrid Mamba-Transformer architecture by gradually adding modules to the fundamental network pattern. The base architecture is a Transformer-based feature-extraction backbone that provides spatial representations but lacks temporal modeling and segmentation guidance. After the Transformer block, the Mamba temporal modeling module is included, which captures sequential dependencies on the extracted features. The SAM-Med2D segmentation module is then presented to offer segmentation-sensitive spatial representations that offer a better localization of boundaries and an improvement in the precision of segmentation. Table 5 shows the ablation results of the segmentation performance, where it shows the increase in the performance results as each module is gradually connected to the Hybrid Mamba-Transformer architecture. On the same note, Table 6 also shows the ablation study of the classification performance where each of the modules contributes to the overall

diagnostic ability of the model. Although the baseline performance has been mentioned above, this part is devoted to the gradual increase in performance when more modules are added to the architecture.

Table 5: Ablation study on segmentation performance, demonstrating the effect of adding more and more modules in the Hybrid Mamba-Transformer architecture.

Model Configuration	Dice (%)	IoU (%)	Sensitivity (%)	Specificity (%)	HD95 (mm)	Inference Time (ms)
Transformer Block	84.50	80.10	81.40	82.90	3.60	170
Mamba Temporal Modeling	87.10	82.50	83.60	84.90	3.20	180
SAM-Med2D Segmentation	89.30	85.10	85.80	86.90	2.80	190
Full Hybrid Mamba-Transformer	91.20	89.20	91.80	91.10	2.10	210

Table 6: Ablation study for classification performance, showing the contribution of each module to the total diagnosing performance of the Hybrid Mamba-Transformer model.

Model Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)	Specificity (%)	Sensitivity (%)	MCC	Log Loss
Transformer Block	84.50	83.80	81.40	82.90	88.20	86.60	82.10	0.70	0.35
Mamba Temporal Modeling	87.10	85.30	83.60	84.90	90.10	88.80	84.30	0.74	0.32
SAM-Med2D Segmentation	89.30	87.10	85.80	86.90	92.30	90.00	86.60	0.78	0.30
Full Hybrid Mamba-Transformer	92.10	92.50	91.80	91.15	95.50	94.00	93.10	0.84	0.24

The ablation analysis clearly demonstrates that every module plays an important role in the overall performance of the Hybrid Mamba-Transformer model. Starting from the baseline (already discussed), when adding the Transformer block, it adds to the spatial feature extraction, while Mamba temporal modeling enhances the ability to learn the sequential dependency. The full Hybrid Mamba-Transformer model with all modules integrated is able to provide the highest performance in all aspects, which shows that integrating Mamba and Transformer can perform effective spatio-temporal learning. This hybrid approach provides an improvement in both learning stability and accuracy, and hence could be a robust solution for complex tasks like cardiovascular disease detection.

4.5 Visual and Explainability Analysis

The Grad-CAM and SHAP visualizations illustrate the clinically relevant regions of the human heart that the Hybrid Mamba-Transformer model uses to predict CVD. As can be seen in Fig. 6, both approaches have consistently focused on the left ventricular walls, septal areas, and myocardial boundaries-regions typically

evaluated by cardiologists as the site of functional and structural abnormalities. The heatmaps show strong attention to motion-sensitive and morphology critical zones over multiple frames, showing that the model is making decisions that have an anatomical basis as opposed to spurious artifacts. This correspondence between model and clinically relevant attention aligns the clinical reliability and interpretability of this proposed framework in the real world.

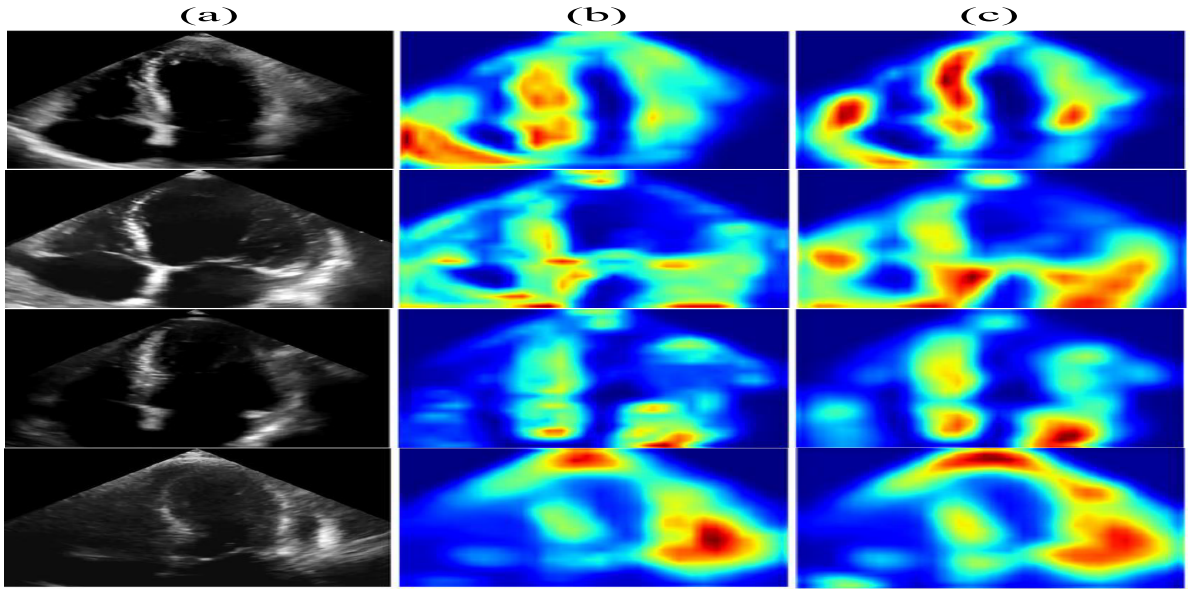

Figure 6: Explainability results with the use of Grad-CAM and SHAP. (**a**) Processed illustrations of echocardiograms. (**b**) Grad-CAM emphasizes discriminative regions in the cardiac images. (**c**) SHAP saliency maps with the feature contribution. Both methods highlight anatomically relevant structures to be used by the model for CVD prediction.

4.6 Comparative Analysis with Existing Studies

This section aims to place our proposed model, the Hybrid Mamba Transformer, in comparison to the recent state-of-the-art models in CVD detection and diagnosis. Table 7 shows us the performance comparison of the Hybrid Mamba-Transformer with existing models on segmentation tasks on key metrics, including Dice, IoU, Sensitivity and Specificity and Table 8 shows us the performance comparison of the Hybrid Mamba-Transformer against existing Models on classification tasks, point out Accuracy, Precision, Recall, Fl-Score and AUC. We compare the diagnostic performance of our model against five models from the recent literature, summarising key metrics, as well as pointing out our contributions across the modalities of hybridisation, interpretability and generalisation.

From the comparative outcome, it proves the efficiency of our Hybrid Mamba-Transformer model, which outperforms the previously used segmentation and classification model with a Dice score of 91.20% and AUC of 95.50%, showing a measured improvement of 4%–6% compared to the previous segmentation and classification methods. The temporal and spatial fact integration via Mamba and Transformer blocks boosts both the segmentation and the classification, whereas the addition of Grad-CAM and SHAP explainability helps ensure that the model is transparent. Additionally, the generalization of our model across multiple data sets and imaging modalities and the use of Focal Loss to mitigate the effect of class imbalance improve recall and sensitivity, making this deep learning model a reliable tool for clinical diagnostics. On the whole, the Hybrid Mamba-Transformer is here to set a new standard in cardiovascular image analysis by putting together segmentation, spatio-temporal learning, and interpretability in a frozen structure.

Table 7: Performance comparison of the Hybrid Mamba-Transformer with existing models on segmentation tasks on key metrics, including dice, IoU, sensitivity, and specificity.

Study	Model	Dice (%)	IoU (%)	Sensitivity (%)	Specificity (%)	HD95 (mm)	Inference Time (ms)
Hybrid Mamba Transformer (our model)	SAM-Med2D + Transformer with MAMBA	91.20	89.20	90.80	91.10	2.10	210
Lin et al. [21]	DSA	85.50	78.90	84.30	88.40	4.50	150
Mandava [15]	MDenseNet201 IDRSNet	87.00	82.30	86.10	87.50	4.00	140
Yang et al. [16]	Msv Mamba	88.20	83.10	87.00	88.50	3.80	160
Nazari et al. [22]	WGAN	84.60	79.60	82.40	86.70	5.00	155
Deng and Wu [23]	NCM-Net	83.70	78.50	80.20	85.30	5.20	165

Table 8: Performance comparison of the Hybrid Mamba-Transformer against existing models on classification tasks, point out accuracy, precision, recall, F1-score, and AUC.

Study	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC (%)	Specificity (%)	Sensitivity (%)	MCC	Log Loss
Hybrid Mamba Transformer (our model)	SAM-Med2D + Transformer with MAMBA	92.10	92.50	91.80	91.10	95.50	94.00	93.10	0.84	0.24
Lin et al. [21]	DSA	88.50	87.80	86.40	86.80	90.10	89.50	84.00	0.75	0.31
Mandava [15]	MDenseNet201 IDRSNet	89.10	88.20	86.70	87.40	91.20	89.80	84.50	0.76	0.28
Yang et al. [16]	Msv Mamba	86.30	84.90	83.10	83.90	88.50	87.30	80.60	0.70	0.35
Nazari et al. [22]	WGAN	85.20	83.00	80.70	81.80	87.10	85.80	79.40	0.68	0.37
Deng and Wu [23]	NCM-Net	84.70	82.10	79.80	80.90	86.00	84.50	–	–	–

4.7 Computational Efficiency and Robustness

This part measures Hypercube Hybrid ptiles' efficacy and sturdiness with key measures, such as time per sample for the integer, model parameters, and FLOP (Floating Point Operations Per Second). These characteristics are of significant importance to learn more about the complexity and the computational cost of the model, but also about its deployability for clinical use. In addition, we perform sensitivity tests to understand the robustness of the model under different kinds of real-world distortions (noise, image rotation, and changes in resolution). Computational efficiency comparison between Hybrid Mamba-Transformer and other state-of-the-art models in terms of inference time, model parameters, and FLOPs can be seen in Table 9. The tests deal with difficulties usually faced in medical imaging, such as noise, the rotation of the images, and varying the quality of an image. The comparison with state-of-the-art models demonstrates the trade-off between model complexity and clinical deployability, due especially to the complex components of the model in the spatial and temporal aspects, and interpretability.

Table 9: Computational efficiency comparison between Hybrid Mamba-Transformer and other state-of-the-art models in terms of inference time, model parameters, and FLOPs.

Model	Inference Time (ms)	Parameters (Millions)	FLOP (Billions)
Hybrid Mamba-Transformer (Our Model)	210	85	150
DSA	150	35	50
MDenseNet201 IDRSNet	140	60	80
Msv Mamba	160	45	70
WGAN	155	70	120

The Hybrid Mamba-Transformer model has a slightly increased inference time (210 ms) at a sacrifice of simpler models such as CardioTabNet (150 ms) and MDenseNet201-IDRSNet (140 ms) when compared to the model. This contributes to the enhanced performance, such as spatio-temporal learning and interpretability features (essential for clinical decision-making). The model's 85 million parameters and 150 billion FLOPS are much higher than simpler models, such as CardioTabNet (with 35 million parameters and 50 billion FLOP), needed to enable the model's high-level functionality of segmentation and classification. These increases in computational cost make it possible for this model to undertake complex tasks with high levels of accuracy, making it ideal for clinical environments where the ability to be transparent and perform well is essential. Table 10 shows us the sensitivity test results proving the Hybrid Mamba-Transformer robustness to noise, rotation, and resolution change, revealing its capability to keep its high performance under different imaging conditions.

Table 10: Sensitivity test results proving the Hybrid Mamba-Transformer robustness to noise, rotation, and resolution change, revealing its capability to keep its high performance under different imaging conditions.

Distortion Type	Dice (%)	IoU (%)	AUC (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
Original (no distortion)	91.20	89.20	95.50	91.80	92.10	91.15
Noise (Gaussian, σ = 0.5)	89.30	85.50	92.30	88.20	90.10	89.60
Rotation (±30 degrees)	90.20	87.40	93.70	89.60	91.20	90.30
Resolution Decrease (50%)	87.10	83.00	91.20	85.50	89.00	86.40
Noise (Gaussian, σ = 1.0)	85.90	81.80	89.80	83.00	87.50	84.70
Rotation (±45 degrees)	86.30	82.60	90.00	84.20	88.30	85.10

For testing the quality of the Hybrid Mamba-Transformer, we performed sensitivity testing by varying various real-world scenarios depending on noise, image rotation, and resolution. These tests simulate typical

issues that medical imaging systems encounter, including noise snapping and variation in the image quality due to rotation or image quality degradation.

The Hybrid Mamba-Transformer model shows great potential under various distortions, including Gaussian noise, rotation, and resolution reduction. Despite a small reduction in performance (e.g., Dice score decreased to 89.30% with noise), the model could achieve fairly high sensitivity (88.20%) and specificity (90.10%), indicating that the model performed well with noisy and rotated images. Even with elevated distortions such as increased noise or rotation, the model's AUC was still high, which suggests the generalization ability and robustness of the proposed model and can be well-suited for clinical applications where imaging conditions are varied. Two aspects of robustness, which are complementary in the proposed framework, are illustrated in Fig. 7. The lack of amalgamation between CVD and non-CVD prediction scores clearly demonstrates the stability of the probability outputs and high probability confidence in all the test data sets shown in the left panel. The distribution of frames per sequence for EchoNet-Dynamic, CAMUS, and UK Biobank CMR is shown in the right panel, which highlights the great temporal variability of these 3 datasets. Despite this heterogeneity, there is consistent performance from the model, even showing good robustness to sequence length, frame rate, and computational load variations.

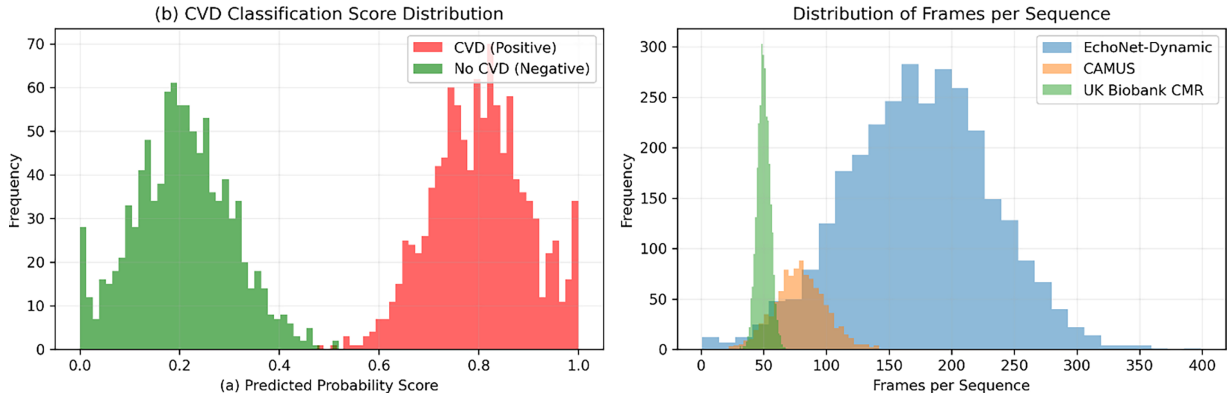

Figure 7: Gradable schemes for computational robustness evaluation of the proposed framework. (**a**) Distribution of predicted CVD and non-CVD probability scores. Strong class separability. (**b**) Variation in the number of frames per sequence in the three data sets springs from distance (e.g., temporal variability, stable model performance for variation in computational burden).

5 Conclusion and Future Work

In this paper, we have proposed the Hybrid Mamba-Transformer framework for cardiovascular disease detection, which effectively integrates SAM-Med2D for accurate segmentation, Mamba temporal modeling for efficient temporal feature extraction, and Transformer spatial feature extraction for robust representation learning. Infra architecture has also integrated Grad-CAM and SHAP (Model interpretability) to provide transparent and actionable insights and tell clinicians about their model. Our model showed that it is better at both segmentation and classification compared to state-of-the-art models in terms of several important metrics such as Dice, AUC, and MCC, across various datasets, such as EchoNet-Dynamic, CAMUS, and UK Biobank CMR. The spatio-temporal learning abilities of the Mamba model enable the framework to learn both temporal and spatial dependency in the medical images, which is essential to operations like dynamic cardiac MRI analysis. The Transformer backbone helps to extract spatial features better, and this helps to improve the robustness of the model in identifying complex patterns in the given imaging data. Despite the added complexity, the performance boost of the model offers a good trade-off in terms of the computational cost, thus making it very appropriate for clinical environments where accuracy and

interpretability are very important. The explainability features further enhance its clinical applicability and ensure that medical professionals can have trust and validation in the model's predictions, which is a key requirement for adoption in the real-world clinical setting.

While the Hybrid Mamba-Transformer shows good performance, it also has certain defects, such as comparatively high computational costs and longer inference time relative to simpler models, which affect its deployment to real environments with limited available time. Future work will be spent on providing an even more optimized framework by looking into more advanced optimization possibilities to reduce these computation costs and inference time while still providing high accuracy in the diagnosis. Additionally, combining federated learning may improve data privacy and make it possible for the model to be learned across various medical centres around the world without compromising data security. Another area of potential success is the investigation of multi-modal methods that use both imaging information and information from clinical examination and genetic profile. Finally, the capability of the model to handle a broader spectrum of cardiovascular conditions and longitudinal data will allow for further improvement in its clinical usefulness in monitoring progressive disease and planning tailored treatment.

Acknowledgement: We would like to thank Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R748), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia for funding this research.

Funding Statement: This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00218176) and the Soonchunhyang University Research Fund. Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R748), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Ghada Atteia; methodology, Ghada Atteia, Muhammad Umer and Abdulaziz Altamimi; software, Muhammad Umer, Khaled Alnowaiser and Nihal Abuzinadah; validation, Yunyoung Nam; formal analysis, Yunyoung Nam and Yongwon Cho; investigation, Ghada Atteia; data curation, Muhammad Umer and Abdulaziz Altamimi; writing—original draft preparation, Ghada Atteia, Muhammad Umer, Nihal Abuzinadah and Abdulaziz Altamimi; writing—review and editing, Khaled Alnowaiser, Yunyoung Nam and Yongwon Cho; visualization, Nihal Abuzinadah; supervision, Yunyoung Nam and Yongwon Cho; project administration, Yunyoung Nam and Yongwon Cho; funding acquisition, Yunyoung Nam and Yongwon Cho. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The dataset can be accessed from the following link: EchoNet-Dynamic: https://echonet.github.io/dynamic/index.html#dataset; CAMUS-Human Heart Data: https://www.kaggle.com/datasets/shoybhasan/camus-human-heart-data; UK Biobank Cardiac MRI: https://community.ukbiobank.ac.uk/hc/en-gb/articles/27830032450461-Cardiac-Magnetic-Resonance-Imaging-Derived-Phenotypes-CMR-IDPs; the dataset can also be requested from the corresponding authors.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tarek Z, Alhussan AA, Khafaga DS, El-Kenawy ESM, Elshewey AM. A snake optimization algorithm-based feature selection framework for rapid detection of cardiovascular disease in its early stages. Biomed Signal Process Control. 2025;102:107417. doi:10.1016/j.bspc.2024.107417.
2. Kiran S, Reddy GR, Dorthi K. A gradient boosted decision tree with binary spotted hyena optimizer for cardiovascular disease detection and classification. Healthc Anal. 2023;3:100173.

3. Marengo A, Pagano A, Santamato V. An efficient cardiovascular disease prediction model through AI-driven IoT technology. Comput Biol Med. 2024;183:109330. doi:10.1016/j.compbiomed.2024.109330.
4. Hoorali F, Khosravi H, Moradi B. An automatic method for microscopic diagnosis of diseases based on URCNN. Biomed Signal Process Control. 2023;80:104240. doi:10.1016/j.bspc.2022.104240.
5. Ying Y, Fang X, Zhao Y, Zhao X, Zhou Y, Du G, et al. SAM-MyoNet: a fine-grained perception myocardial ultrasound segmentation network based on segment anything model with prior knowledge driven. Biomed Signal Process Control. 2025;110:108117.
6. Wang Z, Stavrakis S, Yao B. Hierarchical deep learning with Generative Adversarial Network for automatic cardiac diagnosis from ECG signals. Comput Biol Med. 2023;155:106641. doi:10.1016/j.compbiomed.2023.106641.
7. Singh A, Nagabhooshanam N, Kumar R, Verma R, Mohanasundaram S, Manjith R, et al. Deep learning based coronary artery disease detection and segmentation using ultrasound imaging with adaptive gated SCNN models. Biomed Signal Process Control. 2025;105:107637. doi:10.1016/j.bspc.2025.107637.
8. Rehman A, Naijie G, Ojo S, Nathaniel TI, Samee NA, Umer M, et al. FISM: harnessing deep learning and reinforcement learning for precision detection of microaneurysms and retinal exudates for early diabetic retinopathy diagnosis. BioData Min. 2025;18(1):75.
9. Manocha A, Sood SK, Bhatia M. Federated learning-inspired smart ECG classification: an explainable artificial intelligence approach. Multimed Tools Appl. 2025;84(19):21673–96. doi:10.1007/s11042-024-20084-3.
10. Wang X, Hu J, Lin H, Liu W, Moon H, Piran MJ. Federated learning-empowered disease diagnosis mechanism in the internet of medical things: from the privacy-preservation perspective. IEEE Trans Ind Inform. 2022;19(7):7905–13. doi:10.1109/tii.2022.3210597.
11. Raghavan K, Sivaselvan B, Kamakoti V. Attention guided grad-CAM: an improved explainable artificial intelligence model for infrared breast cancer detection. Multimed Tools Appl. 2024;83(19):57551–78.
12. Wu Y, Zhao T, Hu S, Wu Q, Chen Y, Huang X, et al. Integrating multi-scale information and diverse prompts in large model SAM-Med2D for accurate left ventricular ejection fraction estimation. Med Biol Eng Comput. 2025;63(7):2161–71. doi:10.1007/s11517-025-03310-4.
13. Gurusubramani S, Latha B. Enhancing cardiac diagnostics through semantic-driven image synthesis: a hybrid GAN approach. Neural Comput Appl. 2024;36(14):8181–97. doi:10.1007/s00521-024-09452-0.
14. Naseer A, Khan MM, Arif F, Iqbal W, Ahmad A, Ahmad I. An improved hybrid model for cardiovascular disease detection using machine learning in IoT. Expert Syst. 2025;42(1):e13520. doi:10.22541/au.169358589.99602470/v1.
15. Mandava M. MDensNet201-IDRSRNet: efficient cardiovascular disease prediction system using hybrid deep learning. Biomed Signal Process Control. 2024;93:106147.
16. Yang X, Wang Q, Zhang K, Wei K, Lyu J, Chen L. Msv-mamba: a multiscale vision mamba network for echocardiography segmentation. IEEE Trans Comput Soc Syst. 2025:1–13. doi:10.1109/TCSS.2025.3562441.
17. Sumon MSI, Islam MSB, Rahman MS, Hossain MSA, Khandakar A, Hasan A, et al. CardioTabNet: a novel hybrid transformer model for heart disease prediction using tabular medical data. Health Inf Sci Syst. 2025;13(1):44.
18. Zhao W, Ma H, Jin N, Zheng Y, Guo X. Detection of coronary heart disease based on heart sound and hybrid vision transformer. Appl Acoust. 2025;230:110420. doi:10.1016/j.apacoust.2024.110420.
19. Qi X, Wang S, Fang C, Jia J, Lin L, Yuan T. Machine learning and SHAP value interpretation for predicting comorbidity of cardiovascular disease and cancer with dietary antioxidants. Redox Biol. 2025;79:103470. doi:10.1016/j.redox.2024.103470.
20. Sathi TA, Jany R, Ela RZ, Azad A, Alyami SA, Hossain MA, et al. An interpretable electrocardiogram-based model for predicting arrhythmia and ischemia in cardiovascular disease. Results Eng. 2024;24:103381. doi:10.1016/j.rineng.2025.104070.
21. Lin J, Xie W, Kang L, Wu H. Dynamic-guided spatiotemporal attention for echocardiography video segmentation. IEEE Trans Med Imaging. 2024;43(11):3843–55. doi:10.1109/tmi.2024.3403687.
22. Nazari M, Emami H, Rabiei R, Rabiee HR, Salari A, Sadr H. Enhancing cardiac function assessment: developing and validating a domain adaptive framework for automating the segmentation of echocardiogram videos. Comput Med Imaging Graph. 2025;124:102627.

23. Deng X, Wu H. Echocardiography video segmentation via neighborhood correlation mining. IEEE Trans Med Imaging. 2025;44(12):5172–82. doi:10.1109/tmi.2025.3588157.
24. Dong H, Gu H, Chen Y, Yang J, Chen Y, Mazurowski MA. Segment anything model 2: an application to 2D and 3D medical images. IEEE Trans Biomed Eng. 2026:1–17. doi:10.1109/TBME.2026.3653267.
25. Arif S, Son SH, Kim HY, Kim SC, Lee JY. A diagnosis tool for early detection and classification of heart disease in individuals using transformer mechanisms. Comput Methods Programs Biomed. 2026;277:109248. doi:10.1016/j.cmpb.2026.109248.
26. Mahmood AH, Hasan TM. A custom dilated-separable CNN for automated cardiovascular disease detection using electrocardiogram images. Archit Image Stud. 2026;7(1):1484–98.
27. Ouyang D, He B, Ghorbani A, Lungren MP, Ashley EA, Liang DH, et al. Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. In: Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019 Dec 8–14; Vancouver, BC, Canada.
28. Davi S, Kumar M, Hanif ZM, Kumar A, Kumari M, Ridham F, et al. Deep learning for early detection of cardiovascular diseases from medical imaging. Health Sci Rep. 2025;8(10):e71334. doi:10.1002/hsr2.71334.
29. Salatzki J, Condurache DG, D'Angelo S, Salih AM, Szabo L, Mahmood A, et al. Rheumatoid arthritis and cardiovascular disease associations in the UK Biobank. BMC Med. 2025;23(1):605. doi:10.1093/eurheartj/ehaf784.4210.
30. Singh G, Darji AD, Sarvaiya JN, Patnaik S. Preprocessing and frame level classification framework for cardiac phase detection in 2D echocardiography. Biomed Signal Process Control. 2025;107:107803.
31. Yu T, Chen K. Enhancing cardiac disease detection via a fusion of machine learning and medical imaging. Sci Rep. 2025;15(1):26269. doi:10.1038/s41598-025-12030-6.
32. Jabbar MK, Jianjun H, Jabbar A, Rehman ZU. Mamba-based VoxelMorph framework for cardiovascular disease imaging and risk assessment. IEEE Access. 2025;13:78120–37. doi:10.1109/access.2025.3564962.
33. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin/Heidelberg, Germany: Springer; 2015. p. 234–41.
34. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023 Oct 2–3; Paris, France. p. 4015–26.
35. Patra R, Dutta S, Roy IK, Basak P, Ghosh A. Heart disease detection using vision-based transformer ensemble models. Procedia Comput Sci. 2025;258:3554–69. doi:10.1016/j.procs.2025.04.611.
36. Alsayat A, Mahmoud AA, Alanazi S, Mostafa AM, Alshammari N, Alrowaily MA, et al. Enhancing cardiac diagnostics: a deep learning ensemble approach for precise ECG image classification. J Big Data. 2025;12(1):7.