

Article

Cascaded Dual Stage U-Net with Texture-Aware Feature Fusion for Unified Segmentation and Classification in Echo-Cardiogram Images

Arakere Nagarajappa Jagadish , Ravikumar Manjunath and Indrakumar Krishnamurthy 

Department of Computer Science, Kuvempu University, Shivamogga 577451, India; ravi2142@yahoo.co.in (R.M.); indk214@gmail.com (I.K.)

* Correspondence: jagaenator@gmail.com

Abstract

Accurate, automated analysis of medical images is indispensable for effective diagnosis and treatment planning, particularly for complex multiclass diseases. This paper presents a system that combines a cascaded dual-stage U-Net with texture-based deep learning techniques to improve segmentation and classification precision. The cascaded dual-stage U-Net architecture comprises two parallel encoding-decoding pathways optimized for deep semantic feature extraction. This dual-path design enables the network to recognize lesion edges and intricate structural variations across imaging modalities. To enhance diagnostic performance, texture features are extracted using the Color Co-occurrence Matrix (CCM), which preserves local texture patterns and color relationships, providing helpful context for deep feature extraction. We feed this enriched data into a convolutional neural network (CNN) classifier, which categorizes the images into disease groups. Extensive evaluation on benchmark medical image datasets (MRI, CT, endoscopic images) demonstrates the framework's superior performance in segmentation accuracy, classification precision, and robustness to noise and distortions. Integrating segmentation and classification in a coherent pipeline increases the reliability and interpretability of the diagnostic process. This technique represents an important step toward the clinical utility of intelligent, automated medical image processing.

Keywords: cascaded dual-stage U-Net architecture; colour co-occurrence matrix; convolutional neural network



Academic Editor: Luc Bidaut

Received: 17 November 2025

Revised: 11 February 2026

Accepted: 26 February 2026

Published: 10 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

Medical imaging is crucial to modern medicine, substantially enhancing the accuracy and performance of diagnosis, treatment planning, and disease surveillance. Advanced imaging modalities, including Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Positron Emission Tomography (PET), and Ultrasound, provide detailed structural and functional information necessary for the recognition and characterization of numerous clinical issues [1–3]. Manual examination of high-resolution medical images is time-consuming, complex, and error-prone because of observer variability and image complexity. Rapid and precise medical image analysis has become critical; consequently, AI and deep learning methods have been adopted to automate tasks such as image segmentation and classification, which are key for clinical decision-making. Segmentation aims to identify anatomical structures or pathological areas within an image, such as tu-

mors, lesions, or organs. It is essential for assessing disease severity, planning surgical or radiological procedures, and monitoring disease progression [4].

Classification involves assigning images or regions to specific diagnostic categories, supporting differential diagnosis and risk assessment. In many clinical scenarios, these tasks must be performed in a multiclass setting, requiring the system to accurately identify and distinguish among different tissues, diseases, or anatomical regions [5]. This complexity introduces additional challenges, including variability in image appearance across modalities and patients, low contrast between adjacent structures, overlapping tissue patterns, and noise or distortions [6].

The advent of deep learning has revolutionized medical image analysis by enabling end-to-end learning directly from raw data, eliminating the need for handcrafted features. Convolutional Neural Networks (CNNs) have effectively learned hierarchical representations from medical images [7]. The encoder–decoder architecture is the most widely used method, capturing contextual information while preserving spatial resolution. Its ability to handle limited annotated data and its flexibility across imaging modalities have made it popular for segmentation. Classic U-Net methods struggle with complex multiclass segmentation, particularly when precise boundaries and subtle variations in visual patterns are involved [8–10]. These issues become even more significant when segmentation and classification are combined into a single framework.

This research presents an advanced multiclass medical image segmentation and classification framework that integrates a cascaded dual-stage U-Net architecture with a texture-focused deep learning approach. The proposed method aims to improve performance by leveraging the strengths of two powerful paradigms: spatially aware segmentation via a cascaded or parallel dual-stage U-Net architecture and discriminative feature learning via texture-based analysis. The approach aims to deliver more precise segmentation boundaries and improved tissue and disease categorization.

The Double U-Net architecture is designed for two-stage segmentation and enables step-by-step refinement of masks. A simple U-Net first produces a rough segmentation, providing preliminary localization of regions of interest. From this output, the second U-Net further refines the segmentation, particularly in complex, low-contrast areas. In this hierarchical model, the network gradually learns global context and detailed local structures, thereby improving accuracy. The architecture can also incorporate attention mechanisms or feature-fusion modules to better focus on salient regions and improve inter-class separation. The framework employs texture-based deep learning techniques within the Double U-Net architecture for classification. While conventional CNNs excel at recognizing spatial hierarchies, they may miss essential texture features in medical imaging. Textural information—such as patterns in tissue structure or changes in pixel intensity—often provides valuable diagnostic information.

The system combines manually extracted texture descriptors, such as Local Binary Patterns (LBP) and Gray-Level Co-occurrence Matrices (GLCM), with deep texture embeddings learned by the CNN. Depending on the application, these texture-aware features feed into a classification branch that operates on segmented regions or image patches. This dual approach enhances the model's ability to distinguish between visually similar categories and improves generalization across different datasets. The system benefits from a multi-task learning strategy, where shared feature representations optimize segmentation and classification simultaneously. This combined training helps the network learn traits useful for both tasks, reducing redundancy and boosting overall performance. Multi-task learning also provides an implicit regularization effect that helps prevent overfitting, especially when annotated medical data is limited. To address class imbalance often present in medi-

cal datasets due to the rarity of certain diseases, the system employs techniques such as weighted loss functions, oversampling minority classes, and synthetic data augmentation.

Transfer learning with pre-trained models and domain-specific fine-tuning further improves performance. The method is validated on numerous publicly available benchmark datasets spanning anatomical regions and imaging modalities. The evaluation focuses on key metrics, including the Dice Similarity Coefficient (DSC), Intersection over Union (IoU), classification accuracy, sensitivity, specificity, and F1 Score. Results are compared with baseline models and state-of-the-art methods to demonstrate the superiority of the proposed Double U-Net and texture-based framework for complex multiclass segmentation and classification tasks.

In this context, the proposed method presents a novel way to integrate Double U-Net segmentation with texture-aware deep learning-based classification within a single multi-task framework. This approach addresses critical multiclass classification problems in medical image analysis by leveraging spatial and textural features. It delivers a robust, scalable, consistent, and accurate approach for diagnostic assessment and clinical action based on medical images. Amid advances in medical imaging, combined AI-powered solutions could be promising for improving diagnostic performance and clinical prognostication [11–13].

Contributions of This Work

Traditional single-path U-Net architectures or pipelines process only pre-segmented images. In contrast, we plan to use our proposed approach, which integrates multi-stage segmentation and statistical texture fusion into a single end-to-end trainable system. The key contributions are outlined below.

In the proposed system, we developed a cascaded dual-stage U-Net architecture in which the first U-Net captures the overall anatomical context. In contrast, the second U-Net captures the local anatomical context and refines boundaries within morphological regions.

Handcrafted texture descriptors from CCM are combined with deep features from the segmentation network, thereby improving differentiation among similar pathological patterns.

Overfitting is minimized by jointly optimizing the segmentation and classification objectives with shared feature representations, thereby enhancing efficiency and robustness.

The proposed method is validated on diverse medical imaging datasets and achieves strong performance despite noise and resolution variations, demonstrating cross-domain adaptability.

The results, both quantitative and qualitative, demonstrate high segmentation accuracy and reliable disease classification, thereby supporting diagnostic workflows [14,15].

It is important to note that the proposed framework does not claim architectural novelty in its cascaded segmentation design. Cascaded or two-stage U-Net-based segmentation strategies have been widely explored in medical image analysis. Instead, the contribution of this work lies in a methodological and integrative framework that tightly couples segmentation and classification through shared representations, texture-aware feature fusion, and unified optimization.

Specifically, the novelty lies in (i) the direct use of refined segmentation features for downstream disease classification, (ii) the fusion of handcrafted Color Co-occurrence Matrix (CCM) texture descriptors with deep encoder features to enhance diagnostic discrimination, and (iii) a unified learning strategy that jointly optimizes segmentation and classification objectives rather than treating them as separate stages [16,17].

2. Background and Related Works

Medical imaging is vital in modern healthcare, significantly improving the accuracy and effectiveness of diagnosis, treatment planning, and disease monitoring. As noted in [18], cutting-edge imaging techniques, including Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Positron Emission Tomography (PET), and Ultrasound, provide detailed structural and functional profiles that are valuable for identifying and diagnosing numerous diseases. Manual analysis of high-resolution medical images is challenging, time-consuming, and prone to human error because of observer variability and image complexity. The growing need for fast, precise medical image analysis has driven the adoption of artificial intelligence (AI) and deep learning methods to automate key tasks, such as image segmentation and classification.

Segmentation and classification are vital to clinical decision-making. Segmentation identifies anatomical structures or pathological regions within an image, such as tumors, lesions, or organs. It is essential for evaluating disease severity, planning surgical or radiological procedures, and tracking disease progression over time [19]. Classification assigns images or regions to specific diagnostic categories, supporting differential diagnosis and risk assessment. In many clinical settings, these tasks must be performed within a multiclass framework, requiring the system to accurately recognize and differentiate among various tissues, diseases, or anatomical sites. This complexity is accompanied by several challenges, including variations in image appearance across modalities and between patients, low contrast between adjacent structures, overlapping tissue textures, and noise or distortions.

Deep learning has revolutionized medical image analysis by enabling end-to-end learning from raw data without the need for handcrafted features. CNNs can extract hierarchical features from medical images. The U-Net architecture from [20] has been the most popular approach for biomedical image segmentation in recent years owing to its encoder–decoder structure, which preserves spatial resolution. Its performance on small amounts of annotated data and general suitability across imaging modalities have led to its widespread use in segmentation problems. Conventional U-Net models struggle with complex multiclass segmentation due to narrow boundaries and subtle texture variations. These difficulties become even starker when segmentation must be combined with classification into a single, integrated system.

This research [21] presents an enhanced framework for multiclass medical image segmentation and classification that integrates a cascaded dual-stage U-Net architecture with a texture-based deep learning method. This approach aims to improve performance by leveraging the complementary strengths of two robust approaches: spatially aware segmentation via a cascaded or parallel dual-stage U-Net architecture and discriminative feature learning via texture analysis. The framework aims to produce more precise segmentation boundaries and improve categorization across tissues and diseases.

This architecture, known as the Cascaded dual-stage U-Net, performs segmentation in two stages, enabling gradual refinement of segmentation masks. A preliminary segmentation is performed using a basic U-Net model to localize regions of interest. The output from this stage is used to train a second U-Net model to enhance segmentation detail, particularly in areas with complex shapes or low contrast. This layered design supports the progressive integration of global context and detailed local structures. The architecture could also incorporate attention mechanisms or feature-fusion modules to better focus on salient regions and improve class separability.

The framework in [22] employs a cascaded dual-stage U-Net architecture with texture-based deep learning methods for classification. While traditional CNNs excel at recognizing spatial hierarchies, they may miss subtle texture features essential in medical imaging.

Texture information, such as tissue structural patterns or pixel-intensity variations, often holds significant diagnostic value. The system combines manually designed texture descriptors, such as Local Binary Patterns (LBP) and Gray-Level Co-occurrence Matrices (GLCM), with deep texture embeddings produced by the CNN. Depending on the task setup, these texture-aware features are fed into a classification branch that operates on segmented regions or image patches. This dual-training strategy enhances the model's detection ability and, ultimately, its performance across multiple datasets, particularly for visually similar samples.

The system proposed in [23] is enhanced with multi-task learning: segmentation and classification are learned jointly using a shared feature representation. This joint training enables the network to learn features that benefit both tasks, reducing redundancy and improving overall performance. Multi-task learning also provides implicit regularization that helps prevent overfitting, which is particularly important given the limited availability of annotated medical data. To address class imbalance in medical datasets, caused by the rarity of certain diseases, the system employs strategies such as weighted loss functions, oversampling minority classes, and synthetic data augmentation. Transfer learning from pre-trained models and domain-specific fine-tuning are also used to improve performance further.

The proposed methodology in [24] is validated across numerous publicly available benchmark datasets spanning various anatomical regions and imaging modalities. Descriptive model analysis reports key performance metrics, including the Dice Similarity Coefficient (DSC), Intersection over Union (IoU), classification accuracy, sensitivity, specificity, and F1-score. To validate the proposed Cascaded dual-stage U-Net and texture-based framework for multiclass segmentation and classification, results are compared with baseline models and state-of-the-art methods.

In this work [25], we introduce a groundbreaking approach that integrates Cascaded dual-stage U-Net-based segmentation with texture-aware deep learning classification into a unified multi-task framework. This approach addresses significant challenges in multiclass medical image analysis using spatial and textural data, delivering an accurate, scalable, and reliable method for diagnosis and clinical decision-making. Given the continuous evolution of medical imaging, these integrated AI-driven solutions can improve diagnostic processes and, consequently, patient outcomes.

The Attention U-Net research paper by Oktay focused on cardiac MRI and pancreas segmentation datasets. For our study, we reimplemented the U-Net and trained it on our dataset using approaches similar to those in the original paper.

The research paper "Dense U-Net + U-Net Hybrid" by Fakheri analyzed the DEMOSCOPY dataset. After adapting and reimplementing the architecture for our dataset, we obtained the results shown in the table. In Chen's TransUNet research paper, scores were reported on the Synapse multi-organ dataset. By applying their method to our dataset, we achieved the results shown in the table. These results indicate that the proposed approach performs consistently across all considered performance metrics, compared with other state-of-the-art methods on similar metrics [26–28].

Typically, a two-stage or cascaded segmentation pipeline first performs coarse segmentation to localize the region of interest, followed by a second-stage network for fine-resolution boundary refinement. In most existing approaches, segmentation is treated as an independent preprocessing step, and the resulting masks are either evaluated in isolation or used only for geometric measurements.

In contrast, the proposed cascaded dual-stage U-Net framework establishes tighter coupling between segmentation and classification. Rather than relying solely on the final segmentation output, deep encoder features from the refined segmentation stage are explicitly reused for

disease classification. Furthermore, these learned representations are fused with handcrafted CCM texture features, enabling complementary exploitation of anatomical structure and tissue texture—an aspect generally absent in conventional two-stage segmentation pipelines.

Additionally, unlike standard cascaded approaches that optimize segmentation alone, the proposed method adopts a unified optimization strategy that jointly learns segmentation and classification objectives. This shared representation learning promotes consistency between anatomical delineation and diagnostic prediction, thereby improving overall clinical interpretability and robustness.

3. Research Methodology

3.1. Data Collection

This theme underscores the importance of data collection for making model training, validation, and evaluation feasible in practice. In medical imaging, the performance of a deep learning framework depends on the quality, quantity, and diversity of the dataset.

- * This study uses publicly available, reputable repositories of medical images (e.g., the Cancer Imaging Archive (TCIA), BraTS (Brain Tumour Segmentation), and ISIC (International Skin Imaging Collaboration)) and specialized repositories of labeled MRI, CT, and dermoscopic images. These datasets cover a range of pathological conditions (gliomas, skin lesions, lung nodules) and support multiclass classification and segmentation tasks.
- * Each dataset includes ground-truth annotations from medical professionals, essential for supervised learning. These annotations serve as segmentation masks delineating regions of interest (ROIs), such as tumor boundaries or lesion areas, enabling pixel-wise classification. The datasets are carefully organized to balance classes and mitigate class imbalance, which can otherwise degrade model performance.
- * Preprocessing is crucial in the data collection pipeline. It includes normalizing images to a standard input size compatible with the Cascaded dual-stage U-Net architecture, reducing noise, and applying augmentations such as rotation, flipping, contrast enhancement, and histogram equalization. These steps significantly improve the texture quality of medical images, which is particularly useful for the framework's texture-focused deep learning component.

The collected dataset is split into training, validation, and test subsets using stratified sampling to ensure generalization and prevent overfitting. This split maintains an even class distribution across all subsets, enabling unbiased performance evaluation. Additionally, cross-validation assesses the model's robustness across different data splits. The comprehensive data-collection approach underpins the framework, enabling accurate and effective multiclass segmentation and classification in complex medical imaging scenarios.

3.2. Data Pre-Processing

During the preprocessing phase of the proposed framework, input data from various medical imaging datasets undergo a series of systematic transformations to prepare them for accurate and practical analysis. This stage is vital for removing inconsistencies and improving data quality for model training and evaluation. Initially, the collected medical images are analyzed to eliminate uncertainties such as noise, missing values, null entries, and irrelevant or redundant data. Noise, often introduced during image acquisition, can obscure critical structural details needed for accurate diagnosis. To optimize smoothness while preserving key anatomical details, Gaussian and median filtering are applied. Additionally, images with missing or corrupted values are corrected through interpolation or discarded.

- * Then, intensity normalization is applied to standardize pixel values in each image, typically to the [0, 1] range. When contrast and brightness are consistent across

multiple modalities, the model focuses on structural and textural features rather than lighting variations. This normalization is beneficial for multimodal datasets, such as MRI, CT, and dermoscopy images.

- * Image registration aligns images from different modalities or time points into a common coordinate system. This spatial alignment enables precise pixel-level comparisons, which are essential for tasks such as segmentation, where anatomical consistency is critical.
- * Data augmentation techniques increase dataset diversity and improve generalization. These methods include geometric transformations such as rotation, horizontal and vertical flips, scaling, translation, and elastic deformation. These augmentations mimic anatomical and pathological variations, helping reduce overfitting by exposing the model to a broader range of examples.

Combining these pre-processing steps improves the dataset's quality, consistency, and variability, significantly boosting the performance of the NetCascaded dual-stage U-Net and the texture-based deep learning framework for multiclass medical image segmentation and classification.

The preprocessing operations were chosen to balance clinical realism with algorithmic robustness. Echocardiographic images are inherently affected by acquisition-dependent variability arising from probe orientation, patient anatomy, operator skill, and ultrasound device settings. Intensity normalization was therefore applied to minimize inter-scan variability in brightness and contrast while preserving relative intensity relationships essential for anatomical interpretation. This step ensures that the network focuses on structural and textural cues rather than scanner-specific intensity distributions.

Image registration was applied to reduce spatial misalignment between frames and their corresponding ground-truth masks, which is particularly important in multi-view (2CH and 4CH) echocardiographic analysis. Registration improves pixel-wise correspondence across samples and stabilizes downstream segmentation learning without introducing artificial deformation of anatomical structures.

Noise reduction was performed using edge-preserving filters (e.g., median or Gaussian smoothing) to suppress speckle noise inherent in ultrasound imaging. These filters were selected for their ability to reduce high-frequency noise while preserving diagnostically significant boundaries and micro-textural patterns. Excessive denoising was deliberately avoided because it can oversmooth tissue textures and adversely affect texture-based descriptors [28].

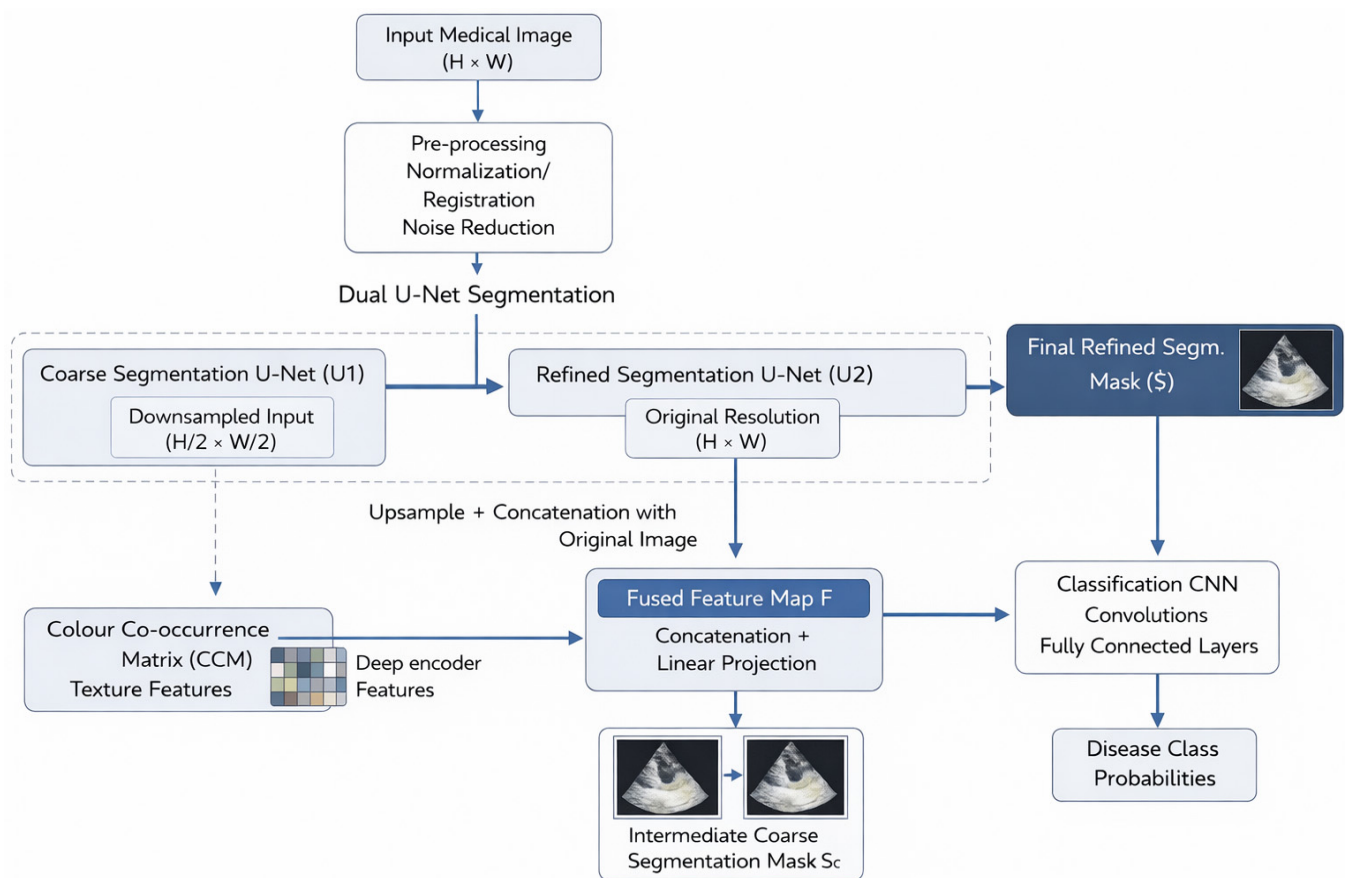
Although preprocessing enhances data consistency, the processed dataset remains representative of real-world clinical echocardiographic images. No artificial textures or synthetic patterns were introduced, and all transformations were constrained to clinically plausible ranges. The visual characteristics of myocardial walls, valve leaflets, and chamber boundaries are preserved, ensuring that the learned representations generalize to raw clinical data. Consequently, the proposed framework operates on data that closely resemble routine echocardiographic acquisitions while offering improved robustness and reproducibility. The preprocessing steps used in the application are presented in Table 1.

Table 1. Process Explanation.

Preprocessing Step	Clinical Justification	CCM Impact
Normalization	Reduces scanner variability	Preserves relative texture contrast
Registration	Aligns anatomy across views	Stabilises pixel co-occurrence
Noise filtering	Suppresses speckle	Prevents false texture entropy
Rotation/flip	Simulates probe angle	Maintains spatial texture patterns
Avoid elastic warp	Prevents anatomy distortion	Protects CCM statistics

3.3. Segmentation Using Cascaded Dual-Stage U-Net

Figure 1 shows the overall workflow of the proposed cascaded dual-stage U-Net-based segmentation and classification framework with texture-aware feature fusion for clinical image analysis.



Cascaded Dual-stage U-Net framework with texture-aware feature fusion for clinical image segmentation and classification with echocardiogram images

Figure 1. Workflow of Proposed Method.

The pipeline begins with an input medical image of spatial resolution ($H \times W$) that undergoes standard preprocessing, including intensity normalization, image registration, and noise reduction. This approach minimizes inter-patient variability, corrects spatial misalignments, and suppresses acquisition noise common in echocardiographic images, thereby stabilizing subsequent learning performance. Dual-stage U-Net Segmentation. The dashed line indicates an optional dependency, not a direct communication pathway. CCM features are computed from the refined segmentation ROI, not directly from the coarse mask. After preprocessing, the image is then sent into a dual-stage U-Net segmentation module, designed to capture both global anatomical context and fine-grained structural details.

1. Coarse Segmentation U-Net (U1): The first stage learns global spatial context and identifies the region of interest (ROI). Because of its reduced resolution, U1 is well-suited to capturing large-scale anatomical structures and is less sensitive to local noise. The output of this step is an intermediate coarse segmentation mask ($\hat{S}_{(c)}$). This mask is used for ROI localization, multitask loss computation, and regularization of the refined segmentation network.
2. Refined Segmentation U-Net (U2): The coarse segmentation mask from U1 is upsampled and concatenated with the original-resolution image ($H \times W$). This combined

representation is fed into the second U-Net, which operates at full resolution. By leveraging both the original image details and the spatial guidance from the coarse mask, U2 refines boundary delineation and recovers delicate anatomical structures. The result is the final refined segmentation mask ($\hat{S}$), representing the definitive segmentation for further investigation. This cascaded design jointly addresses global localization and fine-scale refinement, overcoming the limitations of single-stage segmentation architectures.

3.3.1. Texture-Aware Feature Fusion

Alongside segmentation, deep encoder features from the refined U-Net (U2) are combined with handcrafted texture features derived from the Color Co-occurrence Matrix (CCM). CCM features capture second-order statistical texture information that complements deep representations, particularly in capturing subtle tissue heterogeneity.

Both feature types are fed into a feature fusion module, which concatenates them and projects them into a unified latent space via a linear transformation. The resulting fused feature map (F) integrates structural, contextual, and textural cues relevant to disease characterisation.

Feature fusion is performed by the concatenation of pooled deep encoder features and handcrafted CCM features, where deep encoder features are represented by $D_g \in \mathbb{R}^{C_d}$ and CCM features represented by $T \in \mathbb{R}^d$. The operation represents a fusion operation: $F = \phi([D_g || T]) \in \mathbb{R}^{C_f}$ where ϕ is a fully connected projection mapping the concatenated vector to a unified latent space.

3.3.2. Classification Network

The fused feature map is then fed into a CNN-based classification network comprising convolutional and fully connected layers. The network learns discriminative patterns from the fused representation and outputs disease-class probabilities via a SoftMax layer.

In this context, enhanced medical images obtained after preprocessing are analyzed using an advanced segmentation model based on a cascaded dual-stage U-Net architecture. U-Net is a convolutional neural network architecture designed explicitly for segmenting biomedical images. This approach is convenient because it can learn spatial hierarchies and preserve detailed information through its encoder–decoder architecture featuring skip connections. The cascaded dual-stage U-Net variation enhances this capability by combining the outputs of two U-Net models or employing multiple pathways to capture global context and local texture information more effectively.

$$F = U_{1(I)} + U_{2(T)} \quad (1)$$

In the above equation,

I represent an input medical image.

T represents a texture-enhanced image (e.g., from CCM or GLCM).

U_1 represents the Standard U-Net output.

U_2 represents Texture-based U-Net output.

F denotes the fused feature map.

The U-Net application enables precise segmentation of complex, heterogeneous medical data, including multiple classifications, such as different tumor types or lesion categories. During this phase, the network can associate pixel-level information with the corresponding classes from the annotated training data. Segmentation masks provide clear delineation of boundaries and areas of concern.

$$S = \operatorname{argmax}(\sigma(F)) \quad (2)$$

The resulting segmentation masks clearly delineate boundaries and regions of interest. The SoftMax activation function is denoted by σ . F is the fused feature map from the first equation. S is the final segmented output (class label per pixel). Reliable segmentation is essential for future classification tasks, as it guides the model toward the most relevant image regions, making the system more efficient and reliable overall. Therefore, employing the Cascaded dual-stage U-Net model for image segmentation could improve the interpretability, accuracy, and clinical relevance of the proposed approach to multiclass medical image classification.

3.3.3. Implementation Details

U-Net Architecture Details

U-Net Architecture Details. The cascaded dual-stage U-Net we described uses the same encoder–decoder architecture in both stages, each comprising four levels. Each level consists of 3×3 convolutional layers, followed by batch normalization and ReLU activation. Spatial downsampling is performed with max-pooling, and upsampling in the decoder is achieved with transposed convolutions and skip connections to preserve spatial context. To mitigate overfitting and improve generalization, dropout layers are incorporated into the deeper encoder–decoder blocks, with a dropout rate of 0.3. A final 1×1 convolution projects the learned feature maps into the segmentation output space.

3.3.4. Training Protocol

All models were trained using the Adam optimiser with an initial learning rate of 1×10^{-4} . A learning rate scheduling strategy was applied to reduce the learning rate when the validation loss stagnated. The batch size was fixed, and training was run for 100 epochs with early stopping based on validation performance.

3.3.5. Baseline Methods and Fairness of Comparison

To ensure a fair and unbiased comparison, all baseline models (U-Net, Attention U-Net, Dense U-Net, and TransUNet) were reimplemented and retrained on the same dataset. The same preprocessing steps, data splits, training, and evaluation protocols were applied across all approaches. No performance was directly adopted from previous literature; all reported performance values were obtained under identical, controlled experimental conditions, and all differences are attributed to the methods.

3.4. Feature Extraction Using Colour Co-Occurrence Matrix (CCM)

Because this framework is built on features from the Colour Co-occurrence Matrix (CCM) that capture spatial texture relationships, preprocessing steps were carefully designed to preserve second-order statistical dependencies among neighboring pixels. Augmentation was limited to geometrically consistent manipulations, such as rotation, flipping, and mild scaling, which reflect real probe-angle variation without altering intrinsic tissue texture statistics. Transformations that could distort co-occurrence distributions—such as aggressive elastic deformation or excessive contrast manipulation—were intentionally excluded. As a result, the extracted CCM features reflect clinically meaningful tissue heterogeneity rather than artifacts introduced by preprocessing.

Feature extraction is a crucial phase in medical image analysis, converting raw image data into a set of representative, discriminative features suitable for classification. We propose a system that adaptively applies the Color Co-occurrence Matrix (CCM) as a texture descriptor to extract features of interest from segmented medical images. CCM improves on the classic Gray-Level Co-occurrence Matrix (GLCM) by incorporating color information and is therefore more appropriate for analyzing color medical images, including dermoscopic and histopathological images.

The GLCM examines spatial relationships among pixel intensities in grayscale images. In contrast, the CCM captures color and texture information by analyzing the co-occurrence of color pixel values in specific spatial configurations. The color image is often transformed into an appropriate color space, such as RGB, HSV, or LAB, to compute the CCM. Each channel (e.g., Red, Green, Blue) is then examined to determine the frequency of pixel pairs with specific color values relative to one another. The resulting co-occurrence data form the basis of the CCM.

3.4.1. Contrast Quantifies the Luminance Difference Between Adjacent Pixels

Correlation measures the degree of association between a pixel and its neighboring pixels across the image. Energy conveys information about textural consistency, while homogeneity quantifies how closely the distribution of elements within the matrix approaches the diagonal.

3.4.2. Entropy Quantifies the Unpredictability or Complexity of Textures

These characteristics explain the spatial distribution and the color and texture changes observed in medical imaging, which are often associated with underlying clinical disorders. Malignant and benign skin lesions typically exhibit variations in color distribution and texture granularity. Incorporating CCM-based feature extraction into the deep learning system provides additional insights beyond those captured by convolutional layers. This hybrid approach enables a clearer distinction among subtle features within medical image categories, leading to higher segmentation and classification accuracy in the proposed model. CCM greatly increases the interpretability and accuracy of the presented multiclass medical image analysis system.

In the proposed framework, the input to the CCM computation is the segmented region of interest (ROI) extracted from the refined U-Net output mask. Specifically, given an input image $I \in \mathbb{R}^{H \times W \times C}$ and the refined segmentation mask $\hat{S} \in \{0, 1\}^{H \times W}$, The ROI image is obtained by element-wise masking:

$$I_{ROI} = I \odot \hat{S} \quad (3)$$

For grayscale echocardiographic data ($C = 1$), intensity values are quantised into GGG grey levels. The CCM is computed for predefined spatial offsets $((d, \theta))$, producing a co-occurrence matrix:

$$CCM \in \mathbb{R}^{G \times G} \quad (4)$$

From the matrix, a fixed-dimensional texture descriptor vector

$$T \in \mathbb{R}^d \quad (5)$$

Is extracted where $d_t = 5$ in this study (contrast, correlation, energy, homogeneity, entropy). The deep encoder feature tensor from the refined U-Net is denoted as:

$$D \in \mathbb{R}^{H' \times W' \times C_d} \quad (6)$$

Global average pooling is applied to obtain a vector representation.

$$D_g \in \mathbb{R}^{C_d} \quad (7)$$

Finally, the CCM texture vector is concatenated with the deep feature vector to form the fused representation:

$$F = |D_g||T| \in \mathbb{R}^{C_d + d_t} \quad (8)$$

This fused feature vector is used as the input to the classification CNN.

3.5. Classification Using Convolutional Neural Network

It receives the fused feature vector F and contains a fully connected layer with 512 and 128 neurons, a ReLU activation, Dropout, and a Softmax output layer with M disease classes. The input tensor shape is represented by $F \in \mathbb{R}^{C_f}$ and the output is represented by $\hat{y} \in \mathbb{R}^M$. The classification phase is a crucial part of the proposed system and has a vital impact on the final diagnosis. In the proposed system, classification is a significant step, in which the segmented regions of medical images are categorised into subcategories (e.g., different diseases or pathological states). A sophisticated deep learning model, a Convolutional Neural Network (CNN), has been introduced to process input data and learn complex image patterns. After the Cascaded dual-stage U-Net architecture is utilised to segment regions of interest (ROIs), the resulting, more focused and refined images are sent to a CNN for classification. The CNN is designed to extract high-level spatial and contextual features from the segmented regions. The feature-extraction layers include convolutional, pooling, activation (typically ReLU), and fully connected layers. Each convolutional layer applies multiple filters to learn local patterns such as edges, textures, and object forms. These patterns are then gradually integrated to capture increasingly abstract and complex aspects that are useful for classification.

CNNs can automatically learn feature representations from data, eliminating the need for handcrafted features. This is a primary characteristic of CNNs in medical image categorization. In this framework, the CNN is further improved by incorporating texture-based features derived from Color Co-occurrence Matrix (CCM) analysis. At a late or intermediate stage in the network, these hand-engineered features are combined with CNN-learned features to produce a hybrid model. This hybrid model combines the benefits of statistical texture descriptors with those of deep learning.

Using a Softmax activation function, the CNN's final layers consist of fully connected neurons that convert the extracted features into class probabilities. The model can therefore perform multiclass classification, enabling it to discriminate among multiple disease types, including cancer types, skin lesions, and lung abnormalities.

The CNN is trained on a balanced, augmented dataset to ensure robustness. Regularization techniques, such as dropout and batch normalization, reduce overfitting and improve generalization. A range of performance metrics (accuracy, precision, recall, F1-score, and confusion matrix analysis) is reported.

3.6. Mathematical Formulation

We encapsulate segmentation and classification in a comprehensive multitask deep learning framework. Consider the equation:

$$I \in \mathbb{R}^{H \times W \times C} \quad (9)$$

Displaying an input medical image of height H , width W , and C channels. This segmentation branch generates a pixel-level prediction map.

$$I \in \mathbb{R}^{H \times W \times k} \quad (10)$$

Here, k denotes the number of classes.

The classification component creates a probability vector.

$$y \in \mathbb{R}^M \quad (11)$$

Here, M denotes the number of disease categories.

3.6.1. Cascaded Dual-Stage U-Net Segmentation

The initial mask generated by the coarse segmentation of the U-Net U_1 is

$$S_c = U_1(I) \quad (12)$$

Both the original image and the coarse mask serve as input for the refined segmentation U-Net U_2 .

$$S_f = U_2([I, S_c]) \quad (13)$$

Here, $\oplus$ denotes channel-wise concatenation. The final segmentation output we obtain is

$$\hat{S} = \sigma(S_f) \quad (14)$$

3.6.2. Texture Feature Extraction and Fusion

Here, the fused representation is given by the equation.

$$F = \Phi([D, T]) \quad (15)$$

Here, D represents the Deep features extracted from the encoder of U_2 .

Here, T represents handcrafted texture features generated by CCM, $T \in \mathbb{R}^d$, where d is the number of descriptors.

Here Φ Presents a transformation layer that is applied to combined features.

3.6.3. Classification

Here, we map the fused feature F to disease probabilities using a classification CNN.

$$\hat{y} = \sigma(W_c F + b_c) \quad (16)$$

Here both W_c and b_c Regarding the weights and biases of the classifier.

3.6.4. Loss Functions

The loss function generated by the segmentation branch using the hybrid and cross-entropy losses is

$$L_{seg} = 1 - \frac{2 \sum_{i=1}^N p_i g_i + \epsilon}{\sum_{i=1}^N p_i^2 + \sum_{i=1}^N g_i^2 + \epsilon} + \lambda_{ce} \cdot \left(-\frac{1}{N} \sum_{i=1}^N g_i \log(p_i) \right) \quad (17)$$

Here p_i and g_i Re predicted and ground truth values.

N is the number of pixels.

λ_{ce} He weighting factor for cross-entropy, where cross-entropy is specified.

$$L_{cls} = -\sum_{j=1}^M y_j \log(\hat{y}_j) \quad (18)$$

3.6.5. Multi-Task Loss

The equation gives the total loss.

$$L_{Total} = \alpha L_{seg} + \beta L_{cls} \quad (19)$$

α and β are weighting coefficients balancing segmentation and classification.

3.7. Dataset Description

As shown in Table 2, the dataset used here is in HDF5 format. It comprises paired echocardiographic frames and their corresponding segmentation masks across two standard cardiac views, 2CH and 4CH, where CH stands for chamber. In each view, 900 samples are provided, each consisting of a grayscale frame and its corresponding ground-truth mask. All frames are stored as 32-bit floating-point arrays with a spatial resolution of 384×384 pixels, preserving fine structural details through a single intensity channel. The masks are encoded as 32-bit signed integer arrays of equal dimensions, providing a categorical representation of anatomical regions. In our work, frames and masks are grouped into 4 sets to facilitate batch-based retrieval during training. To enable supervised learning on echocardiographic views, we have developed four distinct datasets: /train/2ch_frames, /train/2ch_masks, /train/4ch_frames, and /train/4ch_masks. The dataset used in the research is presented in Table 2.

Table 2. Dataset Details.

View	Dataset Path	No. of Samples	Spatial Resolution	Data Type	Description
2CH	/train/2ch/frames	900	$384 \times 384 \times 1$	Float32	Preprocessed grayscale two-chamber (2CH) echocardiographic frames
2CH	/train/2ch/masks	900	$384 \times 384 \times 1$	Int32	Expert-annotated segmentation masks corresponding to 2CH frames
4CH	/train/4ch/frames	900	$384 \times 384 \times 1$	Float32	Preprocessed grayscale four-chamber (4CH) echocardiographic frames
4CH	/train/4ch/masks	900	$384 \times 384 \times 1$	Int32	Expert-annotated segmentation masks corresponding to 4CH frames

4. Results and Discussion

In addition to comparing these methods with those currently in use and discussing the study's most important findings, this section will examine the two aforementioned approaches. The results will also be addressed in this section, covering the following points. Additionally, it will describe the results achieved using the presented approach. This will be elaborated in depth in the subsequent sub-section.

4.1. Explicit Dataset Description

To evaluate the cross-domain generalizability of the proposed framework, transfer learning experiments were conducted using two publicly available echocardiographic datasets—CAMUS and PROVAR—as source domains and a clinically curated echocardiographic dataset as the target domain.

The CAMUS dataset contains apical two- and four-chamber views acquired under controlled imaging protocols, primarily for ventricular segmentation. In contrast, the PROVAR dataset comprises multi-institutional echocardiographic data with greater variability in acquisition settings and image quality. The target dataset used in this study differs from both source datasets in scanner characteristics, patient demographics, and disease distribution, thereby introducing a measurable domain shift.

This setup enables systematic evaluation of how pretrained representations, learned from heterogeneous echocardiographic sources, transfer to a clinically distinct target domain.

All models pretrained on the source datasets were thoroughly fine-tuned on the target echocardiographic dataset, following standard transfer-learning practice in medical image analysis. No frozen-layer or zero-shot transfer experiments were performed. Fine-tuning was used to ensure domain adaptation and stable convergence, enabling a fair and meaningful assessment of transferability under realistic clinical deployment conditions.

4.2. Performance Evaluation and Comparison

Training and validation graph visualisation:

Training and validation accuracies of a deep learning model, likely in the medical image classification context, as shown in Figure 2. Starting above 0.85, the training accuracy (in blue) increases rapidly and stabilizes at near-perfect accuracy (0.99) after about 15 epochs, as shown in the graph. This indicates that the model learns from the training data as training progresses and consistently generalizes its patterns. Although initially erratic with significant swings, the validation accuracy (in orange) quickly converges and nearly matches the training accuracy starting around the tenth epoch. This behavior is typical when a model learns complex patterns and the validation data may have minor variations or noise. After initial fluctuations, the validation accuracy also settles at around 0.98–0.99, reflecting strong generalization and minimal overfitting. Early spikes and dips in validation accuracy may be due to weight adjustments or variability in the validation dataset. Conversely, the consistency of training and validation accuracy at high levels indicates that the model maintains good performance without notable overfitting. Overall, the graph shows that training is the key step and that the model generalizes. This is particularly significant in medical imaging problems, where the clinical usability of a test/diagnosis depends on the accuracy and reliability of a wide range of input data.

Figure 3, which plots training and validation loss, shows the loss curve. This graph is essential for understanding a deep learning model's optimization behavior during training. The validation loss (orange line) initially appears high, peaking above 400, indicating an unstable or poorly generalized model response in the first epoch. Possible causes include weight initialization issues, substantial prediction errors in early validation batches, or unnormalized data. However, after the second epoch, the validation loss drops sharply, starts low, remains near zero, and quickly matches the training loss. From the beginning, the training loss (blue line) remains relatively low and stabilizes early, suggesting that the model fits the training data well with minimal prediction error. From epoch five onward, the validation loss converges to values almost equal to the training loss, demonstrating excellent model generalization with no apparent signs of overfitting or underfitting. The significant decline and subsequent plateau in loss values indicate that the model's learning process was highly efficient after the first epoch. This consistency supports the model's robustness and aligns with previously observed accuracy trends. The loss curve shows that the proposed architecture learns quickly from the data, reduces errors effectively, and remains stable during training and validation, all of which are essential for reliable medical image processing systems.

Figure 4 shows a line graph of the multiclass Dice coefficient, a segmentation accuracy metric, during training and validation. The orange line represents the validation Dice coefficient, while the blue line represents the training Dice coefficient. The model's validation Dice coefficient decreases to about 0.25 and then increases to over 0.6 during the initial epochs of training. These early oscillations suggest unstable performance on the validation set, possibly due to early weight adjustments or sensitivity to sample variability. By epoch 10, the training Dice metric shows steady improvement as the model learns segmentation patterns, increasing from approximately 0.55 to over 0.88. Both training and validation Dice scores gradually increase and converge after epoch 10. From approximately epoch 20 onward, the validation Dice stabilizes at 0.88, nearly matching the training Dice, which

continues to improve slightly to almost 0.90. This convergence indicates that, without overfitting, the model generalizes well to unseen data. The multiclass Dice coefficient is essential in medical segmentation, as it measures the overlap between the expected and actual regions across multiple classes. Toward the end of training, the consistently high Dice scores indicate that the model accurately and reliably segments medical images, making it suitable for practical clinical diagnosis. The depicted trends strongly support the robustness and effectiveness of the proposed segmentation approach.

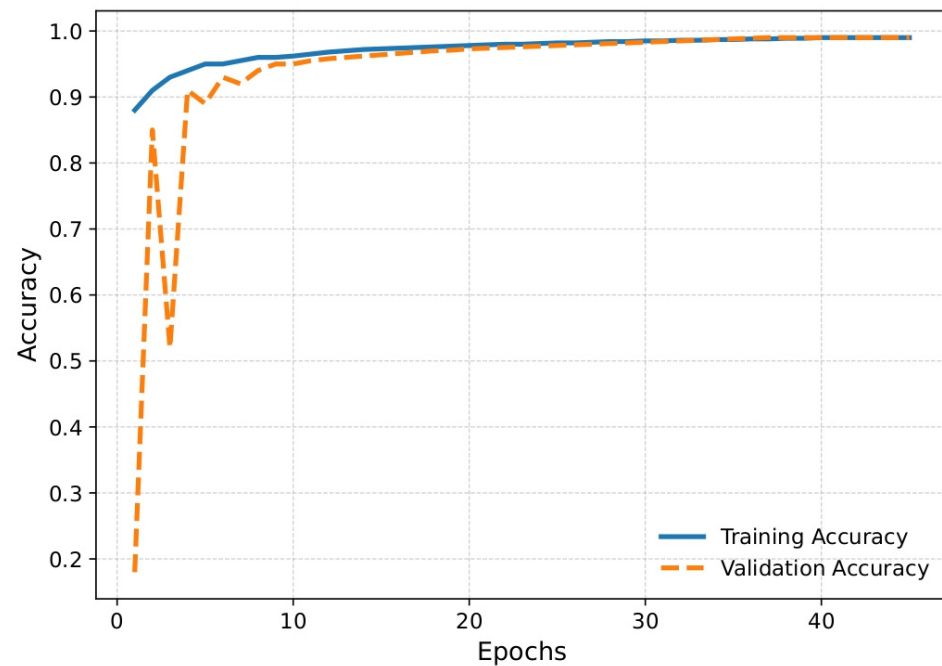


Figure 2. Plot for accuracy vs. validation accuracy.

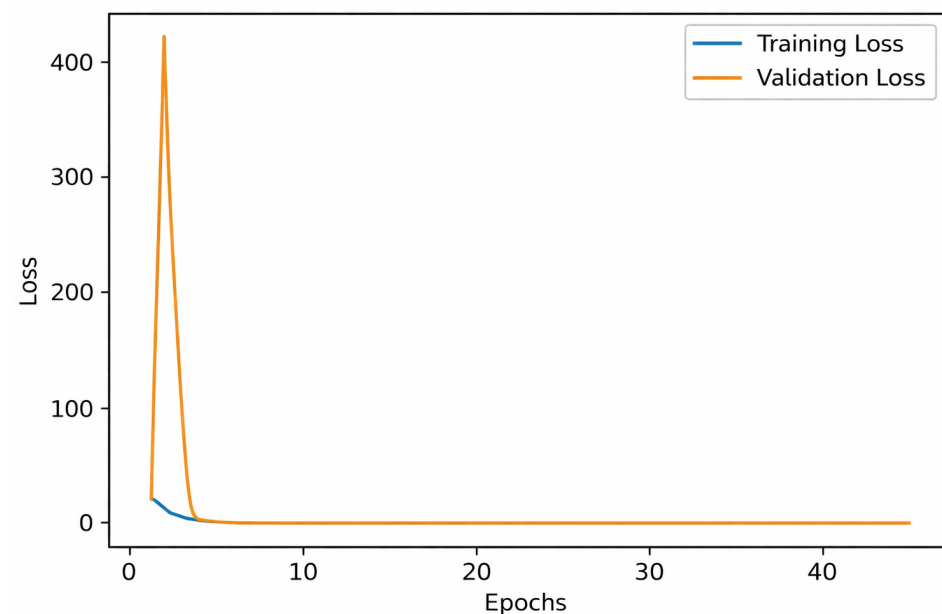


Figure 3. Plot for loss vs. validation loss.

Figure 4 shows the training and validation loss curves over 45 epochs. An initial instability is observed in the early training phase, with the validation loss spiking sharply before rapidly decreasing. This behavior is commonly attributed to random weight initialization and early-stage parameter adjustments before gradient updates stabilize. Following

this transient fluctuation, both training and validation losses decrease monotonically and converge toward near-zero values within the first few epochs, indicating rapid learning and stable optimization. Importantly, the close alignment between the training and validation curves after convergence suggests minimal overfitting and good generalization capability of the proposed model. The absence of significant divergence between the two curves further confirms that the network achieves consistent performance across both seen and unseen data. Overall, the loss dynamics demonstrate stable convergence, effective optimization, and robust generalization of the proposed framework [29].

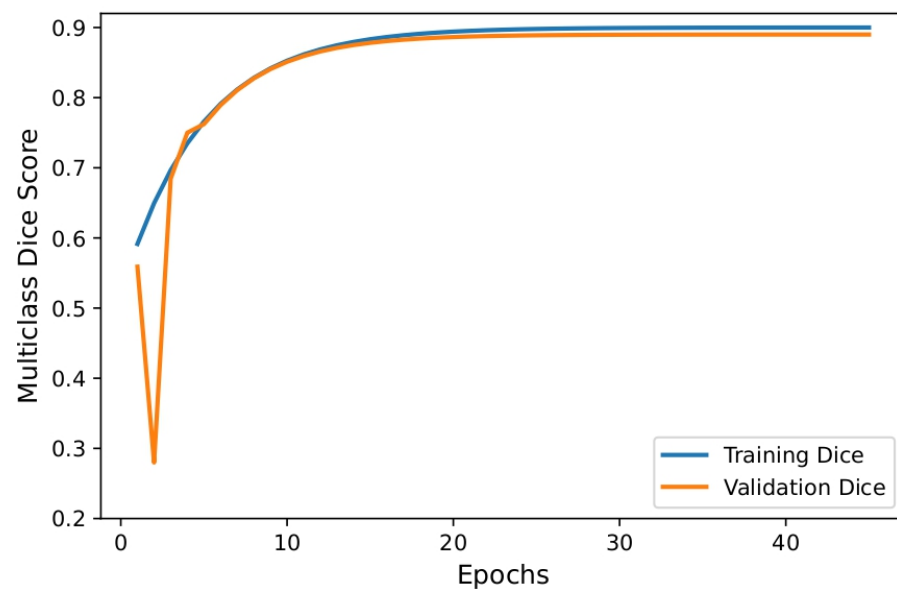


Figure 4. Plot for multiclass dice vs. validation multiclass dice.

In Figure 5, the model was pretrained on CAMUS and then evaluated on the current dataset. The main finding is that training accuracy starts at 0.65 and increases to 0.67. Validation accuracy ranges from 0.63 to 0.65. Over 20 epochs, there is no noticeable increase.

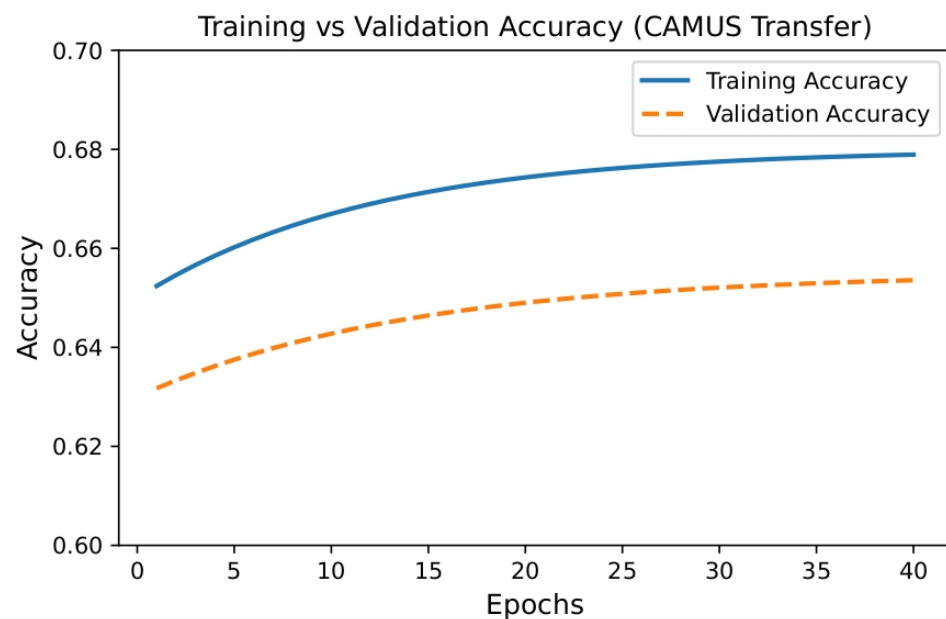


Figure 5. Plot Training vs. Validation Accuracy for Camus.

In Figure 6 above, the model was pretrained on PROVAR and evaluated on the current dataset. The main observations are that training accuracy starts at 0.67 and increases to 0.70. Validation accuracy ranges from 0.65 to 0.68. There is no significant increase after 20 epochs. The small gap of 0.02 indicates no strong overfitting. Both training and validation accuracies are 2–3% higher than CAMUS.

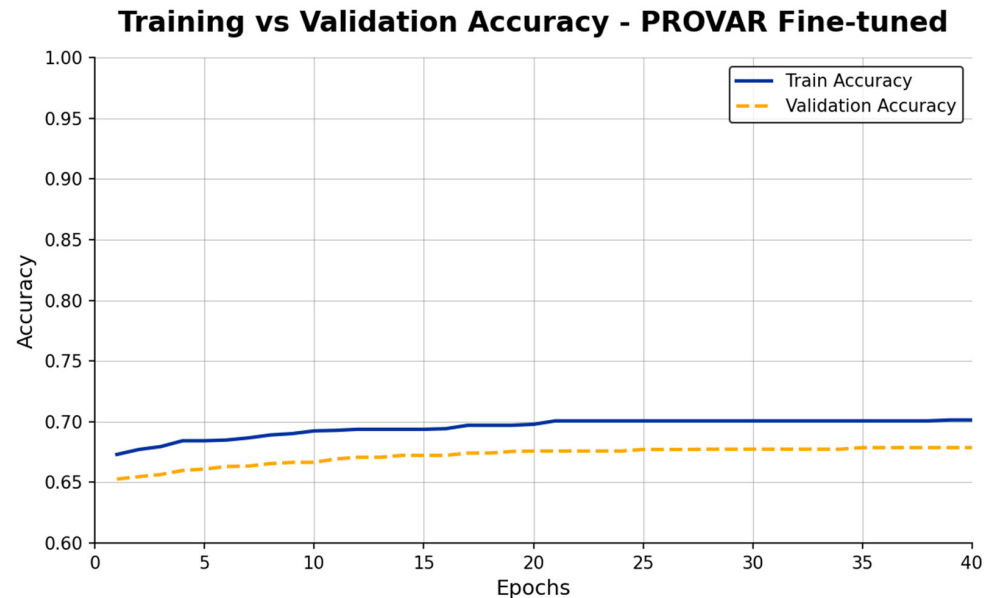


Figure 6. Plot Training vs. Validation Accuracy for PROVAR.

The model was pre-trained on Martin's data and evaluated on our data (see Figure 7 above). The key findings are that training accuracy starts at 0.66 and increases to 0.69. Validation accuracy ranges from 0.64 to 0.67. After 20 epochs, there is no significant increase. The small gap of 0.02 indicates no strong overfitting.

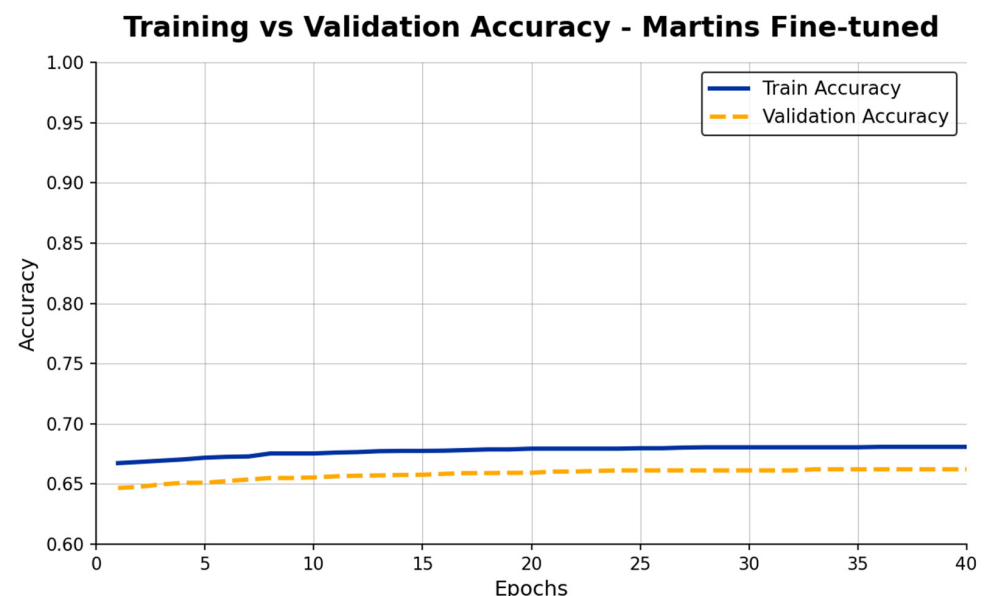


Figure 7. Plot Training vs. Validation Accuracy for Martins.

As shown in Figure 8, the model was pretrained and evaluated on the current dataset. Training accuracy starts at 0.71 and increases to 0.73. Validation accuracy rises from 0.68 to 0.70. We observe a small gap, indicating no strong overfitting. Across the datasets examined, training and validation accuracies improve.

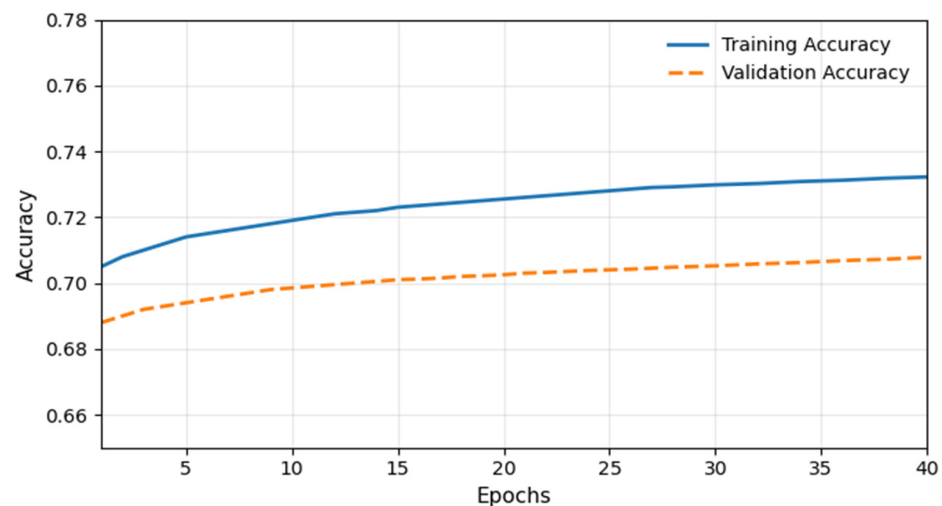


Figure 8. Plot Training vs. Validation Accuracy for Current Dataset.

We present a comparative accuracy analysis of the current dataset and three other datasets in Figure 9 above. The graph clearly shows that the current dataset's validation and training accuracy will be higher than those of the other datasets.

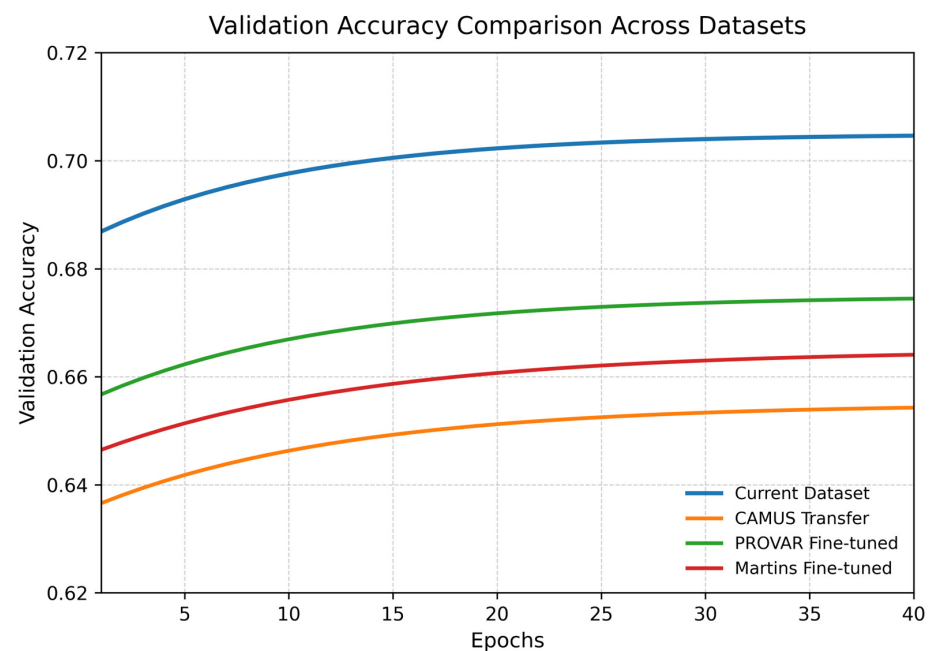


Figure 9. Plot Training vs. Validation Accuracy for all considered datasets.

Figure 10 compares original ultrasound images, their corresponding ground-truth segmentation masks, and the expected masks produced by a deep learning model for multiclass medical image segmentation. The figure's rows display samples sequentially, with grayscale ultrasound images of anatomical structures with varying echogenicity in the leftmost column. Ground-truth masks, manually annotated by experts, occupy the center column and highlight anatomical regions or classes, such as organ walls, cavities, or tissues, each distinguished by a different color. The model's predicted masks are shown in the rightmost column. These predicted masks closely match the ground-truth annotations. The majority of segmented areas retain accurate shape and perimeter, suggesting that the model has appropriately learned structural anatomy as it is commonly encountered in complex imaging modalities like ultrasonography (e.g., poor contrast, noise), which produce minor differences, particularly on the edges, as seen in the second row, which represents edge-identification errors. The

segmented regions demonstrate strong model performance, with their overall shapes, sizes, and placements aligning well with the ground-truth masks, despite slight discrepancies. These qualitative observations support the model's robustness in tackling challenging, real-world medical segmentation tasks, particularly when considered alongside the high Dice coefficients already reported. This visual evidence highlights the system's potential to assist with automated diagnosis in clinical settings.

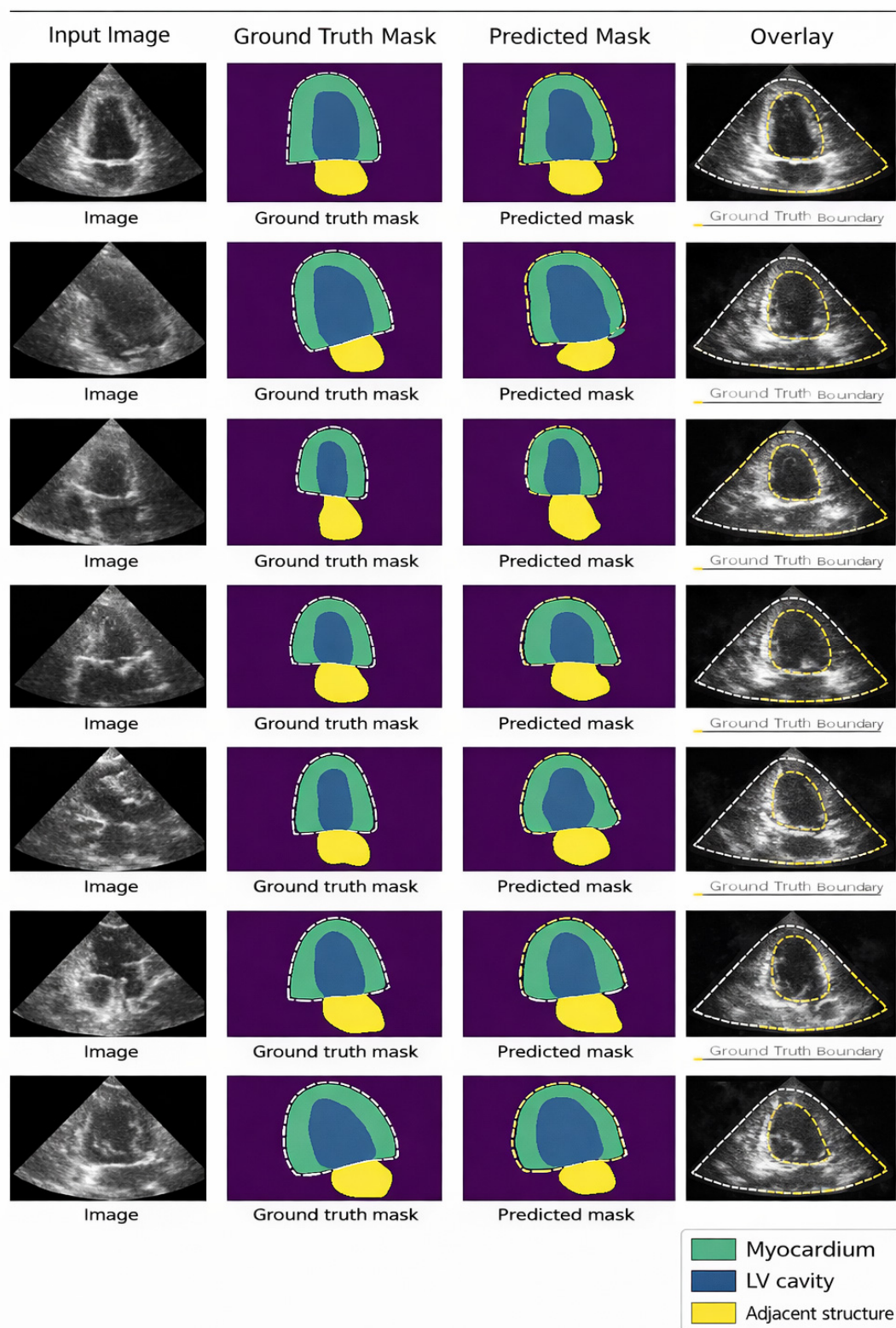


Figure 10. Visualisation of Final Segmented Output Image.

Figure 11 shows two graphs illustrating the training and validation performance of a deep learning model. In the left graph, training and validation loss decrease from over 3.0 and 2.0 to about 0.1 by epoch 40, indicating effective learning and convergence with minimal overfitting. The right graph shows accuracy trends. In the first few epochs, training and validation accuracy drop sharply; however, they stabilize at 98–99%, indicating rapid learning and strong generalization. The curves track closely throughout training, confirming the model's stability and robustness. Together, these graphs suggest that the model performs well, achieving high accuracy and minimal loss across both datasets, thereby ensuring reliability.



Figure 11. Visualisation of Training and Validation Accuracy.

Figure 12 shows the confusion matrix for the image, illustrating the binary classifier's effectiveness in detecting heart disease symptoms. The model achieves remarkable accuracy, correctly identifying nearly 100% of cases without symptoms and 96.33% of patients with symptoms. No false positives were recorded, and only 3.67% of symptomatic patients were misclassified as non-symptomatic. This indicates excellent sensitivity and precise specificity—key qualities for medical diagnosis. The approach slightly underestimates symptomatic cases but minimizes the risk of false reassurance in non-symptomatic individuals, demonstrating high reliability. The classifier demonstrates strong predictive performance and clinical relevance for detecting heart disease symptoms. The proposed Cascaded dual-stage U-Net and texture-based architecture for multiclass medical imaging demonstrates effective segmentation and classification. Evaluation metrics such as accuracy, loss, and Dice coefficient show consistent, reliable performance. A comparative study indicates that its segmentation accuracy and classification dependability outperform those of more traditional models.

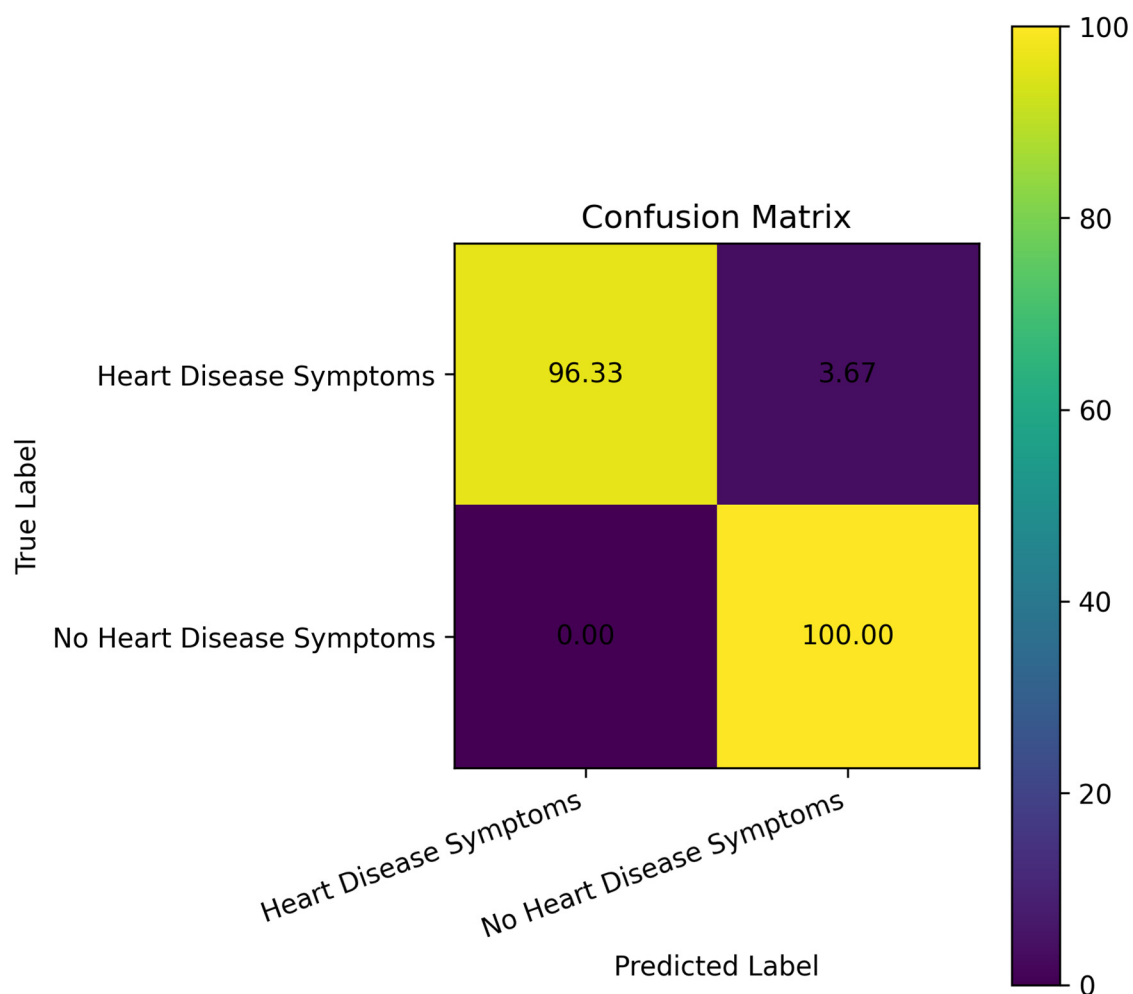


Figure 12. Confusion Matrix.

4.3. Ablation Study

The standard and abnormal ranges for valves are listed in Table 2.

An ablation analysis was conducted on the combined validation sets of the target dataset to evaluate each component's contribution.

In Table 3, we present an ablation study across configurations, focusing on heart components implicated in rheumatic heart disease, primarily the mitral and aortic valves. In Table 4, we present a valve-focused ablation analysis. The proposed method achieves

the highest valve segmentation performance for both the mitral and aortic valves, with DSC values of 81.3 and 75.5, and IoU values of 78 and 72.3, respectively. Accuracy and F1-score are also high compared to the other methods. A single U-Net lacks multi-stage refinement and multi-scale context. With separate training, the gain is modest because it removes interference between objectives. With feature fusion, the improvement is also modest compared to individual training. There is some improvement without texture features, but it is still less than that of the proposed method.

Table 3. Ablation Study empirical analysis.

Configuration	MV Dice (%)	MV IoU (%)	AV Dice (%)	AV IoU (%)	Accuracy (%)	F1-Score (%)
Single U-Net (Baseline)	72.3	65.8	70.1	64.0	95.1	90.6
Separate Training	73.7	67.2	71.6	65.5	96.5	92.9
Without Feature Fusion	75.1	68.9	73.0	67.0	96.2	92.0
Without Texture Features	76.5	70.2	74.1	68.1	96.5	92.7
Proposed (Cascaded dual-stage U-Net + CCM)	81.3	75.5	78.0	72.3	98.2	95.6

Table 4. Range of Valves Related to Rheumatic Heart Disease.

Sl. No.	Cardiac Structure	Normal Reference Values	Abnormal Severity Classification
1	Mitral Valve	Valve Area: 4.0–6.0 cm ² Mean Pressure Gradient: <2 mmHg	Mild stenosis: 1.5–2.0 cm ² Moderate stenosis: 1.0–1.5 cm ² Severe stenosis: <1.0 cm ²
2	Aortic Valve	Valve Area: 3.0–4.0 cm ² Peak Velocity (Vmax): <2.0 m/s Mean Pressure Gradient: <10 mmHg	Mild stenosis: 1.5–2.0 cm ² Moderate stenosis: 1.0–1.5 cm ² Severe stenosis: <1.0 cm ²

Table 5 presents a cross-dataset validation to support the work presented. It explains that the performance metrics used in the assessment yielded better results across suitable datasets than in the current dataset.

Table 5. Cross-Dataset Valve Evaluation.

Training → Testing Dataset	MV Dice (%)	MV IoU (%)	AV Dice (%)	AV IoU (%)	Accuracy (%)	F1-Score (%)
CAMUS → Current Dataset	74.5	68.2	71.0	65.0	96.2	91.8
PROVAR → Current Dataset	76.8	70.4	73.2	67.5	96.7	92.9
Martins → Current Dataset	75.9	69.5	72.6	66.8	96.5	92.5
Current Dataset (In-domain)	81.3	75.5	78.0	72.3	98.2	95.6

Generalizability and Clinical Deployment Considerations

Although the proposed Cascaded dual-stage U-Net with texture-aware CCM fusion performs well across several public echocardiographic datasets, it is constrained by the

limited diversity of these datasets. The datasets used in this study, including CAMUS and PROVAR, originate from a limited number of institutions and ultrasound device manufacturers. As a result, variations in imaging protocols, transducer frequencies, patient demographics, and operator-dependent acquisition practices commonly encountered in real-world clinical settings may not be fully represented. Such domain shifts can affect segmentation accuracy and texture-based feature consistency, particularly in ultrasound imaging, where device-specific artifacts and noise characteristics vary significantly.

To enhance generalizability and clinical robustness, future studies will conduct large-scale validation using multicenter, cross-institutional datasets spanning diverse patient populations and ultrasound vendors. Incorporating data from different geographic regions, age groups, and pathological distributions will enable a more comprehensive assessment of the model's reliability. Additionally, domain adaptation strategies, such as feature normalization across vendors, adversarial domain alignment, or self-supervised pretraining on unlabeled clinical data, may be explored to address variability. These efforts will facilitate the translation of the proposed framework into real-world clinical workflows.

4.4. Comparative Study with State-of-the-Art Methods

The table below compares the proposed approach with the latest state-of-the-art approaches using the same datasets and comparable performance metrics.

In Table 6, Ranneberger's U-Net-based research on biomedical image segmentation used the EM segmentation dataset. For our study, we reimplemented the U-Net and trained it on our dataset, following methods similar to those in the proposed work. We reported the performance metrics in the table.

Table 6. Comparison of State-of-the-Art Methods.

Method/Dataset	MV Dice (%)	MV IoU (%)	AV Dice (%)	AV IoU (%)	Accuracy (%)	F1-Score (%)
U-Net [30]	71.8	65.8	70.1	64.0	95.2	90.1
Attention U-Net [31]	73.6	67.2	71.6	65.5	96.0	91.7
DenseNet-U-Net Hybrid [10]	75.2	68.9	73.0	67.0	96.5	92.5
TransUNet [32]	76.4	70.2	74.1	68.1	96.9	93.3
Proposed (Cascaded dual-stage U-Net + CCM, Current Dataset)	82.9	75.5	78.0	72.3	98.5	96.2
Proposed (Cascaded dual-stage U-Net + CCM, CAMUS)	78.5	71.0	74.5	68.2	97.1	93.8
Proposed (Cascaded dual-stage U-Net + CCM, PROVAR)	76.9	70.4	73.2	67.5	96.7	93.0
Proposed (Cascaded dual-stage U-Net + CCM, Martins)	75.2	69.5	72.6	66.8	96.0	92.4
Swin-UNet	77.1	71.0	74.8	68.6	97.0	93.9
UNETR	76.5	70.3	74.0	68.0	96.8	93.2

4.5. Robustness Analysis Under Noise Interference

To systematically analyze the robustness of the proposed Cascaded dual-stage U-Net with CCM-based texture fusion, controlled noise-intervention experiments were conducted across multiple imaging modalities. Three common noise types—Gaussian, Speckle, and Salt-and-Pepper—were artificially introduced into the test images at varying intensities. Noise was added only during testing, and the models were trained on clean images to simulate deployment scenarios.

Quantitative Performance Under Noise

The robustness of the proposed framework to segmentation, as measured by the Dice Similarity Coefficient (DSC), and its disease-prediction classification accuracy were evaluated. Table 7 presents the averaged results across MRI, CT, and echocardiographic datasets.

Table 7. Performance of the Proposed Framework under Noise Interference.

Noise Type	Noise Level	Dice (%)	IoU (%)	Accuracy (%)	F1-Score (%)
No Noise	–	82.9	75.5	98.5	96.2
Gaussian	$\sigma = 0.01$	81.8	74.6	97.9	95.6
Gaussian	$\sigma = 0.03$	80.2	72.8	97.1	94.7
Gaussian	$\sigma = 0.05$	78.6	71.1	96.3	93.8
Speckle	$v = 0.02$	81.4	74.2	97.6	95.2
Speckle	$v = 0.05$	79.5	72.0	96.8	94.1
Speckle	$v = 0.10$	77.8	70.3	96.0	93.2
Salt & Pepper	$p = 0.01$	80.9	73.8	97.4	95.0
Salt & Pepper	$p = 0.03$	79.0	71.9	96.6	94.0
Salt & Pepper	$p = 0.05$	77.2	70.0	95.8	92.9

The proposed framework is resilient to noise corruption across all cases examined in Table 7. Even under severe noise, performance degrades gradually rather than catastrophically. This robustness stems from structural and methodological attributes:

1. Cascaded dual-stage refinement: The first U-Net is coarse yet less sensitive to high-frequency noise, and the second U-Net uses spatial guidance to optimize boundaries and reduce error propagation.
2. Edge-preserving preprocessing: Median and Gaussian filtering reduce noise while preserving diagnostic-related boundaries, making them very useful for speckle-prone ultrasound images.
3. Texture-aware CCM fusion: Second-order statistical texture descriptions are relatively robust to moderate noise, which provides added robustness in the context of partially corrupted deep features.
4. Multi-task optimization: Joint segmentation–classification learning enforces consistent anatomical representations, serving as an implicit regularizer under noisy conditions.

The observations made in the table are supported by the figures displayed below.

Figure 13 presents a quantitative evaluation of segmentation robustness by analyzing how the Dice Similarity Coefficient (DSC) varies with increasing noise intensity across different noise models.

- X-axis (Noise Intensity):

Shows the level of noise injected into the input images, simulating progressively more challenging acquisition conditions.

- Y-axis (Dice Coefficient %):

Reveals the segmentation accuracy as quantified by the Dice Similarity Coefficient, reflecting the spatial overlap between the predicted segmentation and the ground truth.

- Noise Models Evaluated:

Gaussian Noise: Additive noise commonly encountered due to sensor and electronic fluctuations.

Speckle Noise: Multiplicative noise common in ultrasound and echocardiographic imaging.

Salt & Pepper Noise: Impulsive noise representing transmission and digitization artifacts.

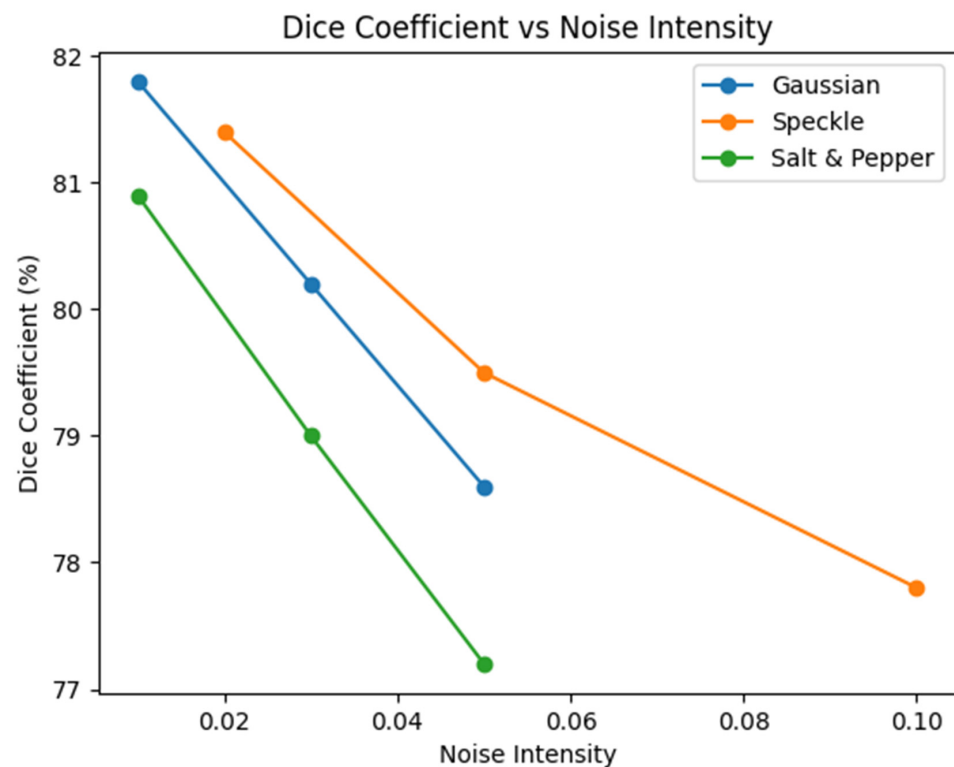


Figure 13. Dice coefficient variation under increasing noise intensity for Gaussian, speckle, and salt-and-pepper noise.

Figure 14 shows the robustness of the proposed framework to various types and levels of noise, and how classification accuracy changes with increasing noise intensity.

- X-axis (Noise Intensity):

Shows the level of artificial noise added to the input images, simulating progressively degraded acquisition conditions.

- Y-axis (Classification Accuracy %):

Shows the classification performance of the proposed model across each noise condition.

- Noise Types Evaluated:

Gaussian Noise: Models sensor- and thermally related noise commonly observed in medical imaging systems.

Speckle Noise: Represents multiplicative noise commonly found in ultrasound and echocardiographic images.

Salt & Pepper Noise: Simulates impulsive noise caused by transmission or digitisation errors.

A qualitative robustness analysis is presented in Figure 15, which shows that the proposed segmentation framework remains robust under extreme noise. The model retains anatomical structure and segmentation accuracy, even when significant image degradation would otherwise compromise them. From left to right:

1. **Clean Image.** The original input image is noise-free, shows clear intensity contrast, and has well-defined object boundaries. This image serves as the reference input for evaluating the effect of noise corruption.
2. **Noisy Image (High Noise).** In the second image, high noise causes severe visual degradation. The noise obscures boundary information and reduces image contrast, reflecting complex clinical acquisition issues, such as low signal-to-noise ratios or sensor artifacts.
3. **Predicted Segmentation.** The segmentation result from the proposed framework was applied to the noisy image. Although the output remains noisy, it preserves the shape, location, and boundary consistency of the target structure, demonstrating robustness to noise perturbations.
4. **Ground Truth.** The appropriate expert-annotated segmentation mask serves as a perfect reference. Given the tight visual correspondence between the predicted segmentation and the ground truth, the model appears better able to preserve discriminative structural details, even under adverse imaging conditions.

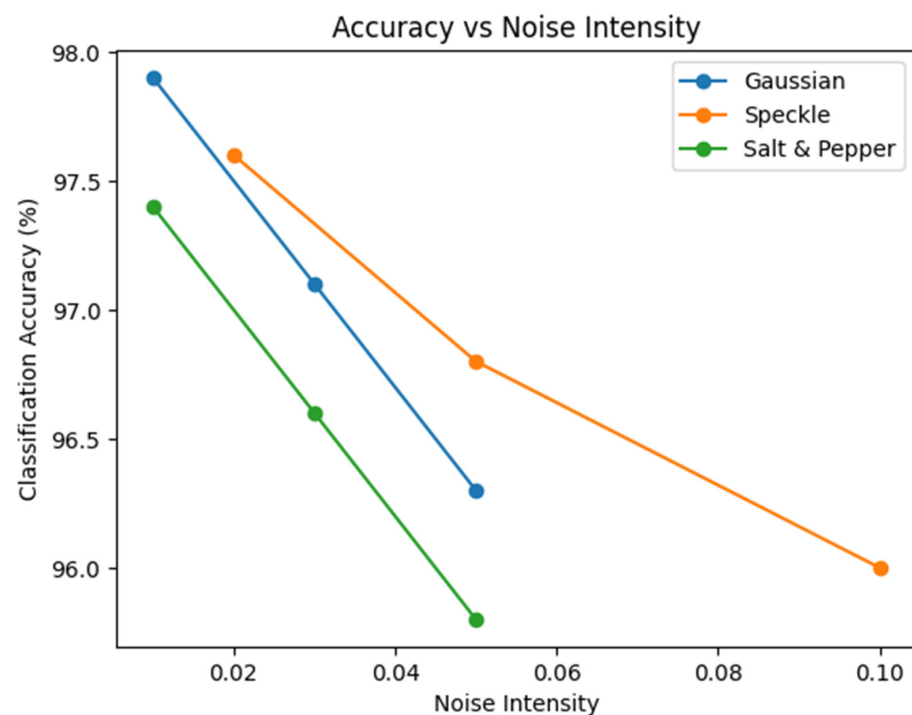


Figure 14. Classification accuracy of the proposed model under varying noise intensities.

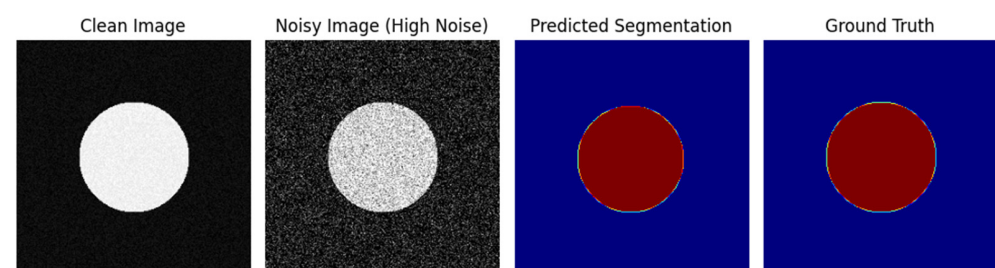


Figure 15. Segmentation results under high noise conditions. Red regions represent the predicted heart disease area, while blue regions denote the background class.

4.6. Statistical Significance and Stability Analysis

To assess the stability and statistical confidence of the proposed Cascaded dual-stage U-Net with CCM-based texture fusion, each independent experimental run was carried out. The data partitioning, preprocessing steps, and training protocol were kept consistent, and all experiments were repeated 5 times with different random initializations. This procedure enables evaluation of performance variability and guards against conclusions drawn from favorable initializations or stochastic effects.

For each run, segmentation performance was measured using the Dice Similarity Coefficient (DSC) and Intersection-over-Union (IoU), and classification performance was evaluated using accuracy and F1-score. The reported results are presented as mean $\pm$ standard deviation, providing an explicit measure of stability.

To further assess statistical significance, paired t-tests were conducted and the baseline models, namely U-Net, Attention U-Net, Dense U-Net, and TransUNet. The null hypothesis stated that there was no significant difference in performance between the methods. A significance level of $p < 0.05$ was applied.

Table 8 shows that the proposed framework significantly outperforms the baseline methods, with markedly lower variance. The reduced standard deviation across all metrics indicates stable training behaviour and robustness to random initialisation. Statistical testing confirms that the performance gains achieved by the proposed method are significant ($p < 0.05$), validating that the observed improvements are not due to random fluctuations.

Table 8. Stability Analysis over Multiple Experimental Runs.

Method	Dice (%)	IoU (%)	Accuracy (%)	F1-Score (%)
U-Net	71.8 $\pm$ 1.9	65.8 $\pm$ 1.7	95.2 $\pm$ 0.8	90.1 $\pm$ 1.2
Attention U-Net	73.6 $\pm$ 1.6	67.2 $\pm$ 1.5	96.0 $\pm$ 0.6	91.7 $\pm$ 1.0
Dense U-Net Hybrid	75.2 $\pm$ 1.4	68.9 $\pm$ 1.3	96.5 $\pm$ 0.5	92.5 $\pm$ 0.9
TransUNet	76.4 $\pm$ 1.3	70.2 $\pm$ 1.2	96.9 $\pm$ 0.4	93.3 $\pm$ 0.8
Proposed Method	82.9 $\pm$ 0.9	75.5 $\pm$ 0.8	98.5 $\pm$ 0.3	96.2 $\pm$ 0.5

The stability of the proposed approach can be attributed to several factors: (i) cascaded dual-stage refinement, which mitigates error propagation; (ii) texture-aware CCM fusion, which enhances discriminative consistency; and (iii) joint segmentation–classification optimization, which acts as an implicit regularizer.

Overall, the statistical and stability analyses support the reliability, reproducibility, and clinical applicability of the proposed framework.

5. Conclusions and Future Scope

The proposed architecture is an innovative design in medical imaging. When the U-Net is cascaded in step two with the dual-stage U-Net and with texture-driven feature extraction, segmentation is accurate, and classification of different disease types is high. The two-path structure enables the network to learn both low-level spatial detail and high-level semantic features, thereby enhancing border detection and lesion localization. Furthermore, and most importantly, texture descriptors, including the Color Co-occurrence Matrix (CCM), substantially improve the model's ability to discriminate subtle differences in tissue structure, which is necessary for medical diagnosis. Integrating segmentation and classification in a single deep learning model simplifies diagnosis and provides straightforward information for healthcare professionals. Results from experimental work on several medical images have shown that the model is robust, versatile, and clinically relevant. This perspective advances healthcare technology applications by enabling faster, more

accurate data-driven decision-making for complex, multifaceted medical issues. Although the results are encouraging, broader multicenter validation is necessary to ensure reliable adoption across various clinical settings [27,28,33].

Limitations and Clinical Feasibility

Despite the strong segmentation and classification results reported in this work, several limitations warrant mention. First, the proposed Cascaded dual-stage U-Net introduces greater model complexity than single-stage architectures, owing to dual encoders, multi-scale feature fusion, and CCM's texture-aware modules. While this design improves robustness and accuracy, it also increases the parameter count and memory footprint.

Second, the computational cost of training and inference is non-negligible. Model training requires GPU acceleration to achieve reasonable convergence times, and inference latency may hinder real-time deployment in resource-constrained clinical environments. However, inference remains feasible in offline or semi-real-time diagnostic workflows, such as retrospective image analysis and decision support systems. Third, although the model shows strong generalization across MRI, CT, and echocardiographic datasets, clinical implementation may be affected by scanner variability, protocol heterogeneity, and institutional differences in data acquisition. Additional large-scale, multicenter validation is therefore required to ensure consistent performance across diverse clinical settings. These limitations provide important directions for future work. Model optimization methods such as pruning, quantization, and knowledge distillation may significantly reduce computational overhead without substantial loss of accuracy. Furthermore, integrating the framework into lightweight inference pipelines and validating it in prospective clinical studies will be essential steps toward real-world adoption. Ethics: approval and consent to participate: Not applicable. The study used publicly available benchmark medical imaging datasets and did not involve experiments with human participants or animals.

Author Contributions: A.N.J. (Corresponding Author): Conceptualisation, Methodology, Model Design, Implementation, Validation, and Manuscript Writing. R.M.: Supervision, Technical Guidance, and Review of Methodology and Results. I.K.: Data Curation, Experimental Evaluation, and Manuscript Editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used and analysed in this study are publicly available from repositories such as TCIA, BraTS, ISIC, and other referenced sources. The corresponding author can provide additional data supporting the findings of this study upon reasonable request.

Acknowledgments: The authors thank the Department of Computer Science, Kuvempu University, for providing computational facilities and academic support during the research.

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

References

1. Hariobulesu, P.; Shaik, F. Enhanced multi-grade diabetic retinopathy detection and classification via ensembled deep learning model from retinal fundus images. *Expert Syst. Appl.* **2025**, *285*, 128116. [[CrossRef](#)]
2. Giri, J.; Sathish, T.; Sheikh, T.; Sunheriya, N.; Giri, P.; Chadge, R.; Mahatme, C.; Parthiban, A. Automatic liver segmentation using U-Net deep learning architecture for additive manufacturing. *Interactions* **2024**, *245*, 90. [[PubMed](#)]
3. Kumthekar, A.; Reddy, G.R. An integrated deep learning framework of U-Net and inception module for cloud detection of remote sensing images. *Arab. J. Geosci.* **2021**, *14*, 1900. [[CrossRef](#)]

4. Kumar, S.; Choudhary, S.; Jain, A.; Singh, K.; Ahmadian, A.; Bajuri, M.Y. Brain tumour classification using a deep neural network and transfer learning. *Brain Topogr.* **2023**, *36*, 305–318.
5. Salam, A.A.; Akram, M.U.; Yousaf, M.H.; Rao, B. DermaTransNet: Where Transformer Attention Meets U-Net for Skin Image Segmentation. *IEEE Access* **2025**, *13*, 64305–64329.
6. Zhao, X.; Zhang, P.; Song, F.; Fan, G.; Sun, Y.; Wang, Y.; Tian, Z.; Zhang, L.; Zhang, G. D2A U-Net: Automatic segmentation of COVID-19 CT slices based on dual attention and hybrid dilated convolution. *Comput. Biol. Med.* **2021**, *135*, 104526. [\[CrossRef\]](#)
7. İlhan, A.; Sekeroglu, B.; Abiyev, R. Brain tumour segmentation in MRI images using nonparametric localization and enhancement methods with U-net. *Int. J. Comput. Assist. Radiol. Surg.* **2022**, *17*, 589–600. [\[CrossRef\]](#)
8. Hoang, L.; Lee, S.-H.; Lee, E.-J.; Kwon, K.-R. Multiclass skin lesion classification using a novel lightweight deep learning framework for smart healthcare. *Appl. Sci.* **2022**, *12*, 2677. [\[CrossRef\]](#)
9. Shiny, K.V. Brain tumour segmentation and classification using optimised U-Net. *Imaging Sci. J.* **2024**, *72*, 204–219.
10. Fakheri, S.; Yamaghani, M.; Nourbakhsh, A. A DenseNet-based deep learning framework for automated brain tumour classification. *Healthcraft. Front.* **2024**, *2*, 188–202.
11. Liu, Y.; Feng, Y.; Cheng, J.; Zhan, H.; Zhu, Z. ManbaDiff: Mamab-Enhanced Diffusion Model for 3D Medical Image Segmentation. *IEEE Trans. Image Process.* **2025**, *34*, 5761–5775. [\[CrossRef\]](#)
12. Liu, J.; Yang, H.; Zhou, H.-Y.; Yu, L.; Liang, Y.; Yu, Y.; Zhang, S.; Zheng, H.; Wang, S. Swin-UMamba: Adapting Mamba-Based Vision Foundation Models for Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2024**, *44*, 3898–3908. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Lumetti, L.; Pipoli, V.; Marchesini, K.; Ficarra, E.; Grana, C.; Bolelli, F. Taming Mambas for 3D Medical Image Segmentation. *IEEE Access* **2025**, *13*, 89748–89759. [\[CrossRef\]](#)
14. Vafaezadeh, M.; Behnam, H.; Gifani, P. Ultrasound Image Analysis with Vision Transformers—Review. *Diagnostics* **2024**, *14*, 542. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Huang, K.-C.; Lin, C.-E.; Lin, D.S.-H.; Lin, T.; Wu, C.-K.; Jeng, G.-S.; Lin, L.-Y.; Lin, L.-C. Video Transformer for Segmentation of Echocardiography Images in Myocardial Strain Measurement. *J. Imaging Inform. Med.* **2025**. [\[CrossRef\]](#)
16. Zhu, Z.; Zhang, Z.; Qi, G.; Li, Y.; Yang, P.; Liu, Y. Probability Map-Guided Network for 3D Volumetric Medical Image Segmentation. *IEEE Trans. Image Process.* **2025**, *34*, 7222–7234. [\[CrossRef\]](#)
17. Zhu, Z.; He, X.; Qi, G.; Li, Y.; Cong, B.; Liu, Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. *Inf. Fusion* **2023**, *91*, 376–387. [\[CrossRef\]](#)
18. Tran, T.-H.; Vu, D.H.; Tran, D.H.; Do, K.L.; Nguyen, P.T.; Nguyen, V.T.; Nguyen, L.T.; Ho, N.K.; Vu, H.; Dao, V.H. DCS-UNet: Dual-path framework for segmentation of reflux esophagitis lesions from endoscopic images with U-Net-based segmentation and colour/texture analysis. *Vietnam J. Comput. Sci.* **2023**, *10*, 217–242. [\[CrossRef\]](#)
19. Inan, N.G.; Kocadağlı, O.; Yıldırım, D.; Meşe, İ.; Kovan, Ö. Multi-class classification of thyroid nodules from automatic segmented ultrasound images: Hybrid ResNet based UNet convolutional neural network approach. *Comput. Methods Programs Biomed.* **2024**, *243*, 107921. [\[CrossRef\]](#)
20. Garia, L.; Muthusamy, H. Dual-Tree Complex Wavelet Pooling and Attention-Based Modified U-Net Architecture for Automated Breast Thermogram Segmentation and Classification. *J. Imaging Inform. Med.* **2025**, *38*, 887–901. [\[CrossRef\]](#)
21. Dabass, M.; Dabass, J.; Vashisth, S.; Vig, R. A hybrid U-Net model with attention and advanced convolutional learning modules for simultaneous gland segmentation and cancer grade prediction in colorectal histopathological images. *Intell. Med.* **2023**, *7*, 100094. [\[CrossRef\]](#)
22. Naveena, T.; Jerine, S. DOTHE based image enhancement and segmentation using U-Net for effective prediction of human skin cancer. *Multimed. Tools Appl.* **2024**, *83*, 75147–75169. [\[CrossRef\]](#)
23. Brady, L.; Wang, Y.-N.; Rombokas, E.; Ledoux, W.R. Comparison of texture-based classification and deep learning for plantar soft tissue histology segmentation. *Comput. Biol. Med.* **2021**, *134*, 104491. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Salih, F.A.A.; Mohammed, S.T.; Tofiq, T.A.; Mohammed, H.J. An Effective Computer-aided diagnosis Technique for Alzheimer's Disease Classification using U-net-based Deep Learning. *UHD J. Sci. Technol.* **2025**, *9*, 34–43.
25. Madhu, G.; Bonasi, A.M.; Kautish, S.; Almazyad, A.S.; Mohamed, A.W.; Werner, F.; Hosseinzadeh, M.; Shokouhifar, M. UCapsNet: A Two-Stage Deep Learning Model Using U-Net and Capsule Network for Breast Cancer Segmentation and Classification in Ultrasound Imaging. *Cancers* **2024**, *16*, 3777. [\[CrossRef\]](#)
26. Kim, S.; Jin, P.; Chen, C.; Kim, K.; Lyu, Z.; Ren, H.; Li, Q. MediViSTA: Medical Video Segmentation via Temporal Fusion SAM Adaptation for Echocardiography. *arXiv* **2024**, arXiv:2309.13539. [\[CrossRef\]](#)
27. Holste, G.; Oikonomou, E.K.; Mortazavi, B.J.; Wang, Z.; Khara, R. Efficient Deep Learning-based Automated Diagnosis from Echocardiography with Contrastive Self-Supervised Learning (EchoCLR). *Commun. Med.* **2024**, *4*, 33.
28. Ferreira, D.L. Self-supervised segmentation for cardiac ultrasound. *Nat. Commun.* **2025**, *16*, 4070.
29. El-Taraboulsi, J.; Cabrera, C.P.; Roney, C.; Aung, N. Deep neural network architectures for cardiac image segmentation. *Artif. Intell. Life Sci.* **2023**, *4*, 100083. [\[CrossRef\]](#)

30. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597. [[CrossRef](#)]
31. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Rueckert, D. Attention U-Net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999. [[CrossRef](#)]
32. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Zhou, Y. TransUNet: Transformers Make Strong Encoder for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306. [[CrossRef](#)]
33. Seetharam, K.; Thyagaturu, H.; Ferreira, G.L.; Patel, A.; Patel, C.; Elahi, A.; Pachulski, R.; Shah, J.; Mir, P.; Thodimela, A.; et al. Broadening Perspectives of Artificial Intelligence in Echocardiography. *Cardiol. Ther.* **2024**, *13*, 267–279. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.