NOTATION TRENDPLUS PUBLISHING
WHERE IDEAS MEET IMPACT

Volume-1, Issue 2, August 2026, Page No- 1-20
Journal of Machine Learning Innovations and Artificial Intelligence Horizons (JMLIAIH)

A Hybrid Vision Transformer and U-Net Framework for Automated Cardiac MRI Segmentation and Abnormality Detection

Prachi Khune

Yeshwantrao Chavhan College of Engineering, Nagpur

DOI: https://doi.org/10.70903/jmliaih.2026.1201

Article Received: 01-07-2026
Revised Version: 05-08-2026
Accepted: 16-08-2026

*Corresponding Author

Prachi Khune

E-Mail Id:
prachi_shahare@yahoo.com

Abstract: Cardiovascular diseases remain a leading cause of mortality worldwide, and accurate segmentation of cardiac structures from MRI is critical for clinical diagnosis. We propose a hybrid framework that integrates a Vision Transformer encoder with a U-Net decoder for automated cardiac MRI segmentation and abnormality detection. The encoder divides input images into non-overlapping patches and processes them through multi-head self-attention layers, thereby capturing long-range dependencies and global anatomical context. These encoded features are then passed to the U-Net decoder, where skip connections preserve high-resolution spatial details from corresponding encoder levels. This design enables the decoder to combine global semantic information with local structural cues for precise pixel-wise segmentation of the left ventricle, right ventricle, and myocardium. A composite loss function, balancing Dice loss and cross-entropy loss, is employed to handle class imbalance and ensure robust pixel discrimination. Following segmentation, quantitative biomarkers such as ventricular volumes, ejection fraction, and wall thickness are extracted from the predicted masks. These features are subsequently fed into a fully connected neural network for binary classification of examinations as normal or abnormal. Furthermore, attention maps derived from the transformer layers provide interpretability by highlighting image regions that most influence the diagnostic decision. The proposed method is evaluated on short-axis cine MRI data with rigorous patient-level data partitioning and extensive augmentation. Our contributions include a novel hybrid architecture that effectively merges global and local feature representations, a clinically relevant abnormality detection pipeline, and an interpretability mechanism that aligns with expert annotations. This work demonstrates significant potential for improving automated cardiac assessment in clinical workflows.

Keywords: Vision Transformer, U-Net, Cardiac MRI, Image Segmentation, Abnormality Detection

Introduction

According to WHO, cardiovascular diseases (CVDs) represent the leading cause of deaths worldwide with an estimated 17.9 million deaths per year. Cardiac magnetic resonance imaging (MRI) has become a gold standard imaging method for non-invasive analysis of cardiac morphology and function, and offers high resolution, multi-planar imaging of the heart. Cardiac MRI is however useful clinically if the segmentation of multiple important anatomic structures, including the left ventricle (LV), the right ventricle (RV), and the myocardium (MYO), is accurate, as this allows for calculation of important cardiac biomarkers such as ejection fraction, left and right ventricular volumes, and wall thickness. Manual segmentation is time-consuming, labour-intensive, and there is variability between observers which drives the need for automated analysis pipelines.

Deep learning has had a significant impact on automated medical image segmentation. The first framework for end-to-end pixel-wise prediction (Long et al., 2015) is the paradigm fully convolutional networks (FCNs). The successful U-Net architecture (Ronneberger et al., 2015) followed, which heavily adopted a symmetric encoder-decoder design with skip connections to retain high-resolution spatial information to be the standard solution for biomedical segmentation tasks. These were expanded on with numerous adaptations including the use of deeply supervised nets (Lee et al., 2015), 3D U-Net (Çiçek et al., 2016), for volumetric data, or multi-scale feature extraction. Although promising, convolutional neural networks (CNNs) have a significant drawback due to their local receptive field, which fundamentally limits their ability to capture long-range dependency and global contextual information—an important feature needed to delineate complex, non-local cardiac geometries.

Global relationship and information modeling has been revolutionized with the introduction of the self-attention concept in the Transformer Architecture (Vaswani et al., 2017). In computer vision this idea was coined into Vision Transformer (ViT) (Dosovitskiy et al., 2020) that preprocesses the image as a sequence of patches and learns global semantics without convolutional bias. Complementary properties of CNNs and Transformers has inspired recent studies on hybrid models. For example, Transnet (Chen et al., 2021) combines a Transformer encoder with a CNN decoder, and Swin-UNet (Cao et al., 2022) uses shifted window attention to capture both global and local information. The above works give a valuable scientific basis to combine ViT encoders with the U-Net decoder for robust cardiac segmentation.The above works give a rich and solid foundation of scientific principles and are relevant when it comes to combining ViT encoders with U-net decoders for robust cardiac segmentation.

In this paper, we introduce a novel hybrid deep learning architecture that is shown to combine the advantages of a Vision Transformer encoder with a U-Net decoder for automated cardiac MRI segmentation and abnormality detection. The key is to leverage recent advances in transformer architectures for a transformer-based encoder that divides input MR slices into the same fixed-size patches giving embeddings longer-range anatomical dependencies from the entire cardiac structure. This is then connected as a skip connection to a U-Net decoder to preserve detail in fine scale boundary region between the ground truth and the resulting model. It eliminates this limitation by directly handling the modeling of global relationships in a CNN

model, without compromising on the ability to represent high-resolution localization, as required for delineating complex cardiac geometry.

There are three-fold contributions we make. First, a hybrid architecture for effectively combining both global and local feature representations is implemented and compared with two baseline models of pure CNN and pure Transformer architectures, which provides significant improvements for cardiac MRI segmentation tasks. Second, we apply our segmentation pipeline to a clinically relevant abnormality detection problem by computing quantitative biomarkers using predicted masks and submitting them to a classifier, to realize a full computer aided diagnosis (CAD) system. Third, we add an interpretability mechanism by visualizing attention maps on images showing the regions that are most important for making the diagnostic decision and correspond to the expert annotations. The proposed approach is comprehensively tested at the patient level, using short axis cine magnetic resonance (MR) images with rich data augmentation.

The related work of cardiac MRI segmentation, hybrid architectures and CAD systems is reviewed in section three. The proposed Hybrid Transformer-U-Net structure and the abnormality detection pipeline are explained in Section 3. In the fourth section, details of the experimental setup are described, such as the training and preprocessing details and the dataset used. Results—quantitative and qualitative are presented in Section 5. The implications, limitation and future directions have been covered in Section 6. The paper ends with section 7.

Related Work

Deep Learning for Cardiac MRI Segmentation

In the last decade, cardiac MRI segmentation has made significant strides with the evolution and adoption of deep learning. Initial attempts mainly used fully convolutional networks (FCNs) and their derivatives, that showed the feasibility of end-to-end pixel-wise predictions for cardiac structures (Long et al., 2015). Because of their symmetric encoder-decoder design and the skip connections that kept the spatial resolution, the U-Net architecture (Ronneberger et al., 2015) soon emerged as the structure of choice. This was furthered in later research to extend to 3D volumes; the 3D U-Net (Çiçek et al., 2016) could be used to provide volumetric segmentation of cardiac MRI stacks. Later, saliency-based self-attention mechanisms have been added to U-Net based architectures to attend to salient regions. For instance, Octay et al. (2018) developed attention U-Net for suppressing irrelevant background responses and emphasizing foreground structures. However, these convolutional methods are still restricted by their local receptive fields, which means that they lack the ability to model spatial dependencies throughout the cardiac anatomy, especially in the case of the right ventricle, which has a complicated, non-ellipsoidal shape.

Transformer-Based Architectures for Medical Image Segmentation

The Vision Transformer (ViT) (Dosovitskiy et al., 2020) found that using a pure transformer architecture on image patches's performance could be competitive against convolutional network architectures despite the absence of any convolutional inductive biases in it. This breakthrough encouraged a cascade of studies on segmenting Medical images in terms of

transformer. Some approaches (e.g., Chen et al., 2021) adopted a ViT encoder and a CNN decoder, where the encoder adopted the transformer to learn by capturing global context from the last convolutional feature maps. Swin-UNet (Cao et al., 2022) is based on a pure transformer structure, primarily consisting of shifted window attention mechanism with window features, which has achieved excellent performance on multi-organ segmentation tasks. Another novel approach was the hybrid layer architecture proposed by UTNet (Gao et al., 2021) that combines convolutional and transformer layers, combining both local and global information with smaller data needs traditionally seen with transformers alone. Likewise, CoST-UNet (Islam et al., 2024) integrated the convolutional layers with Swin Transformer blocks to segment echocardiograms achieving better performance than the baseline U-Net models. The hybrid design has demonstrated the potential of combining the complementary strengths of CNNs and Transformers for medical image analysis.

Computer-Aided Diagnosis Systems for Cardiac MRI

In addition to segmentation, there have been a number of cardiac MRI segmentation based on cardiac MRI studies that integrated segmentation and classification in a unified cardiac MRI CAD system. For example, a combined ensemble framework that combines the Vision Transformer and the 2D / 3D U-Net was suggested for the prediction of myocardial infarction in MRI images (Al-Antari et al., 2025). This method is to segment cardiac structures then take features to classify them downstream. A light weight combined loss optimization multi-class segmentation of cardiac MRI using Swin Transformer U-Net was proposed next for efficient segmentation that is suitable for clinical deployment (Karthikeyan & Anusuya, 2026). Such work highlights the clinical relevance of making use of segmentation-based diagnosis classification but these approaches fail to be explicitly interpretable, with the absence of a mechanism to provide a visual indication of the model decisions, which is essential for clinical adoption.

Comparison with Existing Works

Our proposed model makes several key breakthroughs compared to existing hybrid models like TransUNet (Chen et al., 2021) and UTNet (Gao et al., 2021) which have showcased the advantages of using transformers and U-Net decoders together. The first difference is that we do not switch to the transformer at the end of the convolutional layers but use it straight from the begin for the whole image broken into patches. The purpose of this design is that global dependencies will be obtained as early as possible during feature extraction (thus giving them a richer context as input to the decoder). Second, our framework is not just segmentation but also a full range of abnormality detection pipeline which extracts biomarkers that are clinically validated and are fed to the CAD system as a classifier. Third, we utilize an interpretability mechanism with emphasis map visualization, which is not included in most of the current hybrid segmentation models. This is a novel approach, and unique, as it combines the technological aspects of global feature extraction, downstream classification, and the interpretability aspects that address both technical and clinical needs of automated cardiac MRI analysis.

Hybrid Transformer-U-Net Architecture for Cardiac MRI Segmentation

The proposed framework addresses the task of automated cardiac MRI analysis through a two-stage pipeline: first, a hybrid deep learning model performs pixel-wise segmentation of cardiac structures; second, quantitative biomarkers extracted from the predicted masks are used for abnormality classification. This section provides the technical details of the segmentation architecture, which forms the core of our contribution

Problem Formulation and Input Representation

Let the input cardiac MRI slice be denoted as $I \in \mathbb{R}^{H\times W\times C}$, where H and W represent the spatial dimensions and C denotes the number of channels (typically $C = 1$ for single-modality MRI). The segmentation task requires assigning each pixel (i, j) to one of K anatomical classes: background, left ventricle (LV) cavity, right ventricle (RV) cavity, and myocardium (MYO), hence $K = 4$. The ground truth segmentation map is denoted as $Y \in \{0,1, \dots, K-1\}^{H\times W}$. The proposed model learns a mapping $f_\theta: I \mapsto \hat{Y}$, where $\hat{Y} \in [0,1]^{H\times W\times K}$ represents the predicted probability distribution over classes for each pixel.

The core innovation of our architecture lies in the encoder design. Instead of using convolutional layers to extract hierarchical features, we employ a pure Vision Transformer (ViT) encoder that directly processes the input image as a sequence of patches. The input image I is first divided into $N = \frac{HW}{P^2}$ non-overlapping patches of size $P \times P$, where P is the patch size (set to 16 in our experiments). Each patch is flattened into a vector $x_p \in \mathbb{R}^{P^2C}$ and linearly projected to a D-dimensional embedding space through a trainable linear projection $E \in \mathbb{R}^{(P^2C)\times D}$. A learnable positional embedding $E_{pos} \in \mathbb{R}^{N\times D}$ is added to the patch embeddings to retain spatial information, yielding the input sequence to the transformer encoder:

$$z_0 = \left[x_{class}; x_p^1 E; x_p^2 E; \dots; x_p^N E\right] + E_{pos} \quad (1)$$

where x_{class} is a special classification token (used for the downstream classification task) and $z_0 \in \mathbb{R}^{(N+1)\times D}$. This formulation ensures that the encoder captures global relationships across the entire cardiac anatomy from the very first layer, without the local inductive bias of convolution.

Vision Transformer Encoder

The transformer encoder consists of L identical layers, each comprising two sub-layers: a multi-head self-attention (MSA) mechanism and a position-wise feed-forward network (FFN). Layer normalization (LN) is applied before each sub-layer, and residual connections are employed after each sub-layer. For a given layer ℓ, the output is computed as:

$$z'_\ell = \mathrm{MSA}\big(\mathrm{LN}(z_{\ell-1})\big) + z_{\ell-1} \quad (2)$$

$$z_\ell = \mathrm{FFN}\big(\mathrm{LN}(z'_\ell)\big) + z'_\ell \quad (3)$$

The MSA mechanism operates by projecting the input sequence into query (Q), key (K), and value (V) matrices through learned linear transformations. For each attention head $h \in \{1, \dots, H\}$, the attention weights are computed as:

$$A_h = \text{Softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) \quad (4)$$

where $d_k = D/H$ is the dimension per head. The output of the MSA is the concatenation of all heads' outputs, projected back to D dimensions. This self-attention operation enables each patch to attend to every other patch in the sequence, thereby modeling long-range dependencies across the entire cardiac structure. For example, the relationship between the left ventricle and right ventricle—which may be spatially distant in the image—is captured through the attention weights, providing global contextual information that is critical for accurate segmentation.

The FFN consists of two linear layers with a Gaussian Error Linear Unit (GELU) activation in between:

$$\text{FFN}(x) = W_2 \cdot \text{GELU}(W_1 x + b_1) + b_2 \quad (5)$$

where $W_1 \in \mathbb{R}^{D \times D_{ff}}$, $W_2 \in \mathbb{R}^{D_{ff} \times D}$, and D_{ff} is the hidden dimension of the feed-forward network (typically $D_{ff} = 4D$). The encoder outputs a sequence of feature representations $z_L \in \mathbb{R}^{(N+1) \times D}$, which encode global semantic information about the input image.

Feature Reshaping and Skip Connection Design

A critical challenge in using a pure transformer encoder for segmentation is that the output is a sequence of patch-level embeddings, not a spatial feature map. To interface with the U-Net decoder, we must reshape the transformer output back into a spatial representation. Let z_L denote the output of the final transformer layer, excluding the classification token. We reshape this sequence into a 2D feature map $F \in \mathbb{R}^{H/P \times W/P \times D}$, where each spatial position corresponds to a patch in the original image. This feature map serves as the input to the decoder.

To preserve fine-grained spatial details, we introduce skip connections that transfer multi-scale feature representations from intermediate transformer layers to corresponding decoder stages. Specifically, we extract feature maps from the outputs of transformer layers at indices ℓ_1, ℓ_2, ℓ_3 (e.g., layers 4, 8, and 12 in a 12-layer encoder). Each of the sequences is reshaped into a spatial feature map F_l ∈ R^(H/P×W/P×D). These are subsequently fed into a 1×1 convolutional projection layer that reduces the number of channels to the one of the corresponding decoder stage, resulting in F'_l∈R^(H/P×W/P×d_l), where d_l = number of channels at decoder stage l. This design not only provides global semantic information (from the final output of the encoder) to the decoder, but also local structural information (from intermediate layers) to assist in accurately determining the boundaries.

U-Net Decoder with Skip Connection Fusion

The decoder uses a U-net style design with a series of up-sampling layers, restoring spatial resolution gradually. The per block up sampling module consists of a transposed convolution

layer with stride 2, concatenating the skip connection feature map from the encoder, before then being followed by two conv layers, followed by batch normalization and ReLU activation. Decoder consists of four layers, the dimensions of the channels decrease to D/2, D/4, D/8 and finally to output layer K (number of classes).

This is because the fusion of transformer features and decoder features is done at every stage. We will use U_l to represent the feature map obtained after upsampling at the decoder stage l and F'_l to represent the feature map at l from the transformer encoder to the skip connection. The fused feature map is calculated as:

$$G_\ell = \text{Conv}\big(\text{Concat}(U_\ell, F'_\ell)\big) \quad (6)$$

Channel-wise concatenation is denoted by Concat, and two successive convolutional layers of the size 3×3 is denoted by Conv. It is for this reason that convolutional architecture can be fused with the transformer architecture to create a neural network that is faster while retaining much of the detail captured in the skip connections. The final decoder output is a feature map Ŷ ∈ R^(H×W×K), and a softmax activation function is applied to the decoder output to obtain pixel-wise class probabilities.

Loss Function and Training Objective

A composite loss function based on Dice similarity coefficient loss and categorical cross-entropy loss is used to train the segmentation model. The Dice loss handles class imbalance by emphasizing overlap between predicted and ground truth regions, and the cross-entropy loss ensures that the labels of positive and negative pixels in the prior mask and prior image are discriminated. For one sample, the total loss is calculated as:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{Dice}} + (1-\lambda)\mathcal{L}_{\text{CE}} \quad (7)$$

τ_Ξ is a weighting hyperparameter (we use λ = 0.5 for our experiments). The Dice loss for class k is the following value:

$$\mathcal{L}_{\text{Dice}}^{(k)} = 1 - \frac{2\sum_{i=1}^{HW} \hat{y}_i^{(k)} y_i^{(k)} + \epsilon}{\sum_{i=1}^{HW} \hat{y}_i^{(k)} + \sum_{i=1}^{HW} y_i^{(k)} + \epsilon} \quad (8)$$

where ϵ is a small smoothing constant (equal to 1×10^(-6)), and the y^_i k and y_i k are a pixel's predicted probability and the actual label respectively, for class k and pixel i. The overall Dice loss is the mean of all the K classes. The categorical cross-entropy loss is given by:

$$\boldsymbol{\mathcal{L}_{\text{CE}} = -\frac{1}{HW}\sum_{i=1}^{HW}\sum_{k=1}^{K} y_i^{(k)} \log \hat{y}_i^{(k)}} \quad \textbf{(9)}$$

It is a composite loss function designed to drive the model to make correct regional overlap and pixel-wise classification of image regions, which is quite beneficial to the delimitation of thin tissue structures, such as the myocardium.

Downstream Abnormality Detection Pipeline

Following segmentation, we extract clinically relevant quantitative biomarkers from the predicted masks. For each examination (comprising multiple slices covering the entire cardiac cycle), we compute the end-diastolic volume (EDV) and end-systolic volume (ESV) for both the LV and RV, from which the ejection fraction (EF) is derived as $\mathrm{EF} = \frac{\mathrm{EDV}-\mathrm{ESV}}{\mathrm{EDV}} \times 100\%$. Additionally, we compute the mean wall thickness of the myocardium at end-diastole. These biomarkers form a feature vector $\mathrm{b} \in \mathbb{R}^5$ that is fed into a fully connected neural network classifier with two hidden layers (64 and 32 neurons, respectively) and a softmax output layer for binary classification (normal vs. abnormal). The classifier is trained separately using ground truth labels from the clinical dataset.

To provide interpretability, we extract attention maps from the transformer encoder layers. Specifically, for a given input image, we average the attention weights across all heads in the final transformer layer, producing a $(N+1) \times (N+1)$ attention matrix. We then extract the row corresponding to the classification token, which captures the attention of the global representation to each patch. This attention vector is reshaped into a 2D map of size $H/P \times W/P$ and up-sampled to the original image resolution using bilinear interpolation. The resulting attention map highlights the image regions that most influenced the model's global representation, providing visual validation against expert annotations.

Experimental Setup

The experimental setup used to test the proposed Hybrid Transformer-U-Net framework for cardiac MRI segmentation and abnormality detection tasks is described. We explain the dataset as well as the preprocessing, baseline methods, evaluation metrics and implementation details.

Dataset Description

An MRI of the heart with short-axis cine images from patients with different cardiovascular abnormalities and healthy volunteers is used for the experiments. The following segmentation masks obtained by experts for the given dataset will be useful for segmentation: left ventricle (LV) cavity, right ventricle (RV) cavity and myocardium (MYO). The annotations were carried out by trained cardiologists and assigned as a true standard to train and validate. The medical data consists of many cardiac disease cases used to cover different anatomical structures and disease conditions in learning a cardiac disease model.The medical data includes anatomical structures of dilated cardiomyopathy, hypertrophic cardiomyopathy, myocardial infarction and normal disease cases which represent diverse cardiac cells.

Data is divided into three sets—70% training, 15% validation, and 15% test—at the patient level to avoid data leakage and to be sure that the data is well generalized. This split at the patient level guarantees that images from the same patient will not be seen in multiple subset collections, offering the correct factors for evaluating the efficiency of the model to generalize to new patients. The dataset includes a large number of patients, each of which has multiple slices throughout the cardiac cycle available for deep learning training.

Preprocessing and Data Augmentation

Preprocessing is a crucial stage for ensuring proper uniformity of the input data and improving the model's learning capabilities for meaningful features. There are multiple stages in the preprocessing pipeline. The images are first cropped to a bounding box that contains only the cardiac region in what is known as a "cardiac window", this limits the computational effort to be spent on the region of interest and ensures that the model is trained on the relevant anatomy. Second, size of all the images is reduced to a common spatial resolution of 256×256 pixels to get the same resolution of input across the entire data set. Third, the pixel values are intensity-translated to the range [0,1] with min-max normalization then standardized with a z-scoring to get 0 mean and unit variance. This normalization step reduces intensity fluctuations in scans caused by various intensity acquisition parameters. Fourth, the noise reduction is performed by applying Gaussian filtering with small kernel size (σ=1.0) in order to eliminate the high-frequency noises and keep the edge information. Lastly, contrast enhancement is done by adaptive histogram equalization that accentuates the borders of the heart, especially in areas with low contrast between the cardiac muscle and blood pool.

There was data augmentation in training to improve the generalizability of the model and prevent overfitting. The augmentation methods are random rotation between -15 and +15 °, horizontal shifting (up to 10% of the image size), vertical shifting (up to 10% of the image size), scaling (multiplier gamma applied between 0.9 and 1.1) and elastic deformation (standard deviation 5 pixels and distance between control points 16 pixels). Such augmentations emulate variations in patient positioning, cardiac orientation and anatomical morphology that may occur in the clinic. All augmentations are applied to data during training, preventing the model from being asked to deal with a wide variety of different transforms.

Baseline Methods

To assess the performances of the proposed Hybrid Transformer-U-Net framework, the performances with three baseline methods representing different architectural paradigms are compared:

1. **Standard Convolutional Neural Network (CNN):** This is a base-line which uses a fully convolutional architecture without any other operations but only convolutional layers for feature extraction. The network has 4 convolutional blocks, each consisting of two 3×3 convolutional layers followed by 2×2 max pooling, in the encoder section, and 4 up-sampling blocks, each made up of two 3×3 transposed convolution layers in the decoder section. This architecture is devoid of any attention mechanism and used as the baseline to assess the effect of global context modeling.
2. **Conventional U-Net:** This baseline implements the standard U-Net architecture (Ronneberger et al., 2015) with a symmetric encoder-decoder structure and skip connections. The encoder uses four down-sampling stages with increasing channel dimensions (64, 128, 256, 512), and the decoder uses four up-sampling stages with corresponding skip connections from the encoder. This architecture represents the current state-of-the-art for biomedical image segmentation and provides a strong baseline for comparison.

3. **Standalone Vision Transformer (ViT):** This baseline employs a pure Vision Transformer (Dosovitskiy et al., 2020) for segmentation, without the U-Net decoder. The ViT encoder processes the input image as a sequence of patches and outputs patch-level embeddings, which are then reshaped and up-sampled to the original image resolution using bilinear interpolation followed by a 1×1 convolutional layer to produce the segmentation map. This baseline evaluates the performance of a pure transformer approach without the spatial reconstruction capabilities of the U-Net decoder.

All baseline methods are trained using the same dataset, preprocessing pipeline, and evaluation metrics as the proposed method to ensure fair comparison.

Evaluation Metrics

Segmentation performance is evaluated using six metrics that capture different aspects of segmentation quality:

1. Dice Similarity Coefficient (DSC): Measures the overlap between predicted and ground truth segmentation masks, defined as $\text{DSC} = \frac{2\times|P\cap G|}{|P|+|G|}$, where P and G are the predicted and ground truth pixel sets, respectively. DSC ranges from 0 (no overlap) to 1 (perfect overlap).
2. Intersection over Union (IoU): Also known as the Jaccard index, defined as $\text{IoU} = \frac{|P\cap G|}{|P\cup G|}$. IoU provides a stricter measure of overlap compared to DSC.
3. Precision: Measures the proportion of predicted positive pixels that are true positives, defined as $\text{Precision} = \frac{TP}{TP+F}$, where TP and FP are true positives and false positives, respectively.
4. Recall: Measures the proportion of ground truth positive pixels that are correctly identified, defined as $\text{Recall} = \frac{TP}{TP+F}$, where FN represents false negatives.
5. Hausdorff Distance (HD): Measures the maximum distance between the boundaries of the predicted and ground truth segmentation masks, defined as $\text{HD} = \max\left(\sup_{p\in P}\inf_{g\in G} d(p,g), \sup_{g\in G}\inf_{p\in P} d(p,g)\right)$, where $d(\cdot,\cdot)$ is the Euclidean distance. HD is sensitive to the worst-case boundary error and is particularly important for clinical applications where accurate boundary delineation is critical.
6. Average Surface Distance (ASD): Measures the average distance between the boundaries of the predicted and ground truth masks, defined as $\text{ASD} = \frac{1}{|P|+|G|}\left(\sum_{p\in P} \min_{g\in G} d(p,g) + \sum_{g\in G} \min_{p\in P} d(p,g)\right)$. ASD provides a more comprehensive assessment of boundary accuracy compared to HD.

For the abnormality detection task, classification performance is evaluated using six metrics:

1. **Accuracy:** The proportion of correctly classified examinations, defined as Accuracy $= \frac{TP+TN}{TP+TN+F}$.
2. **Precision:** The proportion of predicted abnormal cases that are truly abnormal, defined as Precision $= \frac{TP}{TP+FP}$.
3. **Sensitivity (Recall):** The proportion of truly abnormal cases that are correctly identified, defined as Sensitivity $= \frac{TP}{TP+F}$.
4. **Specificity:** The proportion of truly normal cases that are correctly identified, defined as Specificity $= \frac{TN}{TN+FP}$.
5. **F1-score:** The harmonic mean of precision and sensitivity, defined as F1 $= 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$.
6. **ROC-AUC:** The area under the receiver operating characteristic curve, which measures the model's ability to discriminate between normal and abnormal cases across all classification thresholds.

Implementation Details

A deep learning based model PyTorch is used to implement the proposed Hybrid Transformer-U-Net framework.The proposed Hybrid Transformer-U-Net framework is implemented using PyTorch deep learning library. The Vision Transformer is trained with its dimensions: P=16, D=768, L=12, H=12, D_ff=3072. The U-Net decoder consists of 4 stages, each containing 768, 384, 192 and 96 channels respectively. Skip connections are extracted from layers 4, 8, 12 of the transformer and each skip connection feature map is added to the decoder channel dimension by a 1×1 convolutional layer.

The initial learning rate for the Adam optimizer is 1×10^(-4), the learning epochs is 200, and the batch size is given as 8. If the validation loss does not decrease for 10 epochs, learning rate gets decreased by a factor of 0.1. The weight λ for the composite loss function is set at 0.5, setting more than half of the weight on each loss component—Dice loss and cross-entropy loss. Gradient clipping is used to avoid gradient explosion with max norm set to 1.0. The experiments are carried out on an NVIDIA RTX 3090 (24GB) graphics card along with an Intel Core i9-10900K processor.

Fully connected network is trained individually with extracted biomarkers from the training set for the abnormality detection classifer. The Adam optimizer is used to train the classifier for 100 epochs, with a learning rate of 1×10^(-3) and a batch size of 32. Early stopping with a patience of 20 epochs is used, based on validation accuracy.

Experimental Design

The experiments are designed to answer three primary research questions:

1. **Segmentation Performance:** Is there an improvement in the segmentation accuracy of the proposed Hybrid Transformer-U-Net framework over that of the baseline methods (standard CNN, conventional U-Net and standalone ViT) on the proposed task of cardiac segmentation on the three different cardiac structures, namely LV, RV and MYO?
2. The Abnormality Detection Performance refers to the effectiveness of the extracted biomarkers from the predicted segmentation masks in the binary classification (normal/abnormal) performance and how well this compares with the use of "ground truth" biomarkers.
3. **Interpretability:** Are attention maps of transformer-encoder show parts of the image that are clinically relevant which coincide with supervisor annotations to offer meaningful interpretability?

To answer these questions, we conduct the following experiments:

- **Experiment 1 (Segmentation**): Each method mentioned above (proposed, standard CNN, conventional U-Net, standalone ViT) are trained and tested on the test set. Cardiac structural segmentation performance is evaluated individually for each cardiac structure by DSC, IoU, Precision, Recall, HD and ASD and, on average, across all cardiac structures.
- **Experiment 2 (Abnormality Detection):** Biomarkers extracted from the predicted segmentation masks of the proposed method are used to train the abnormality detection classifier. The metrics used for evaluation: Accuracy, Precision, Sensitivity, Specificity, F1-score and ROC-AUC. As a point of comparison, the classifier is also trained with biomarkers from the masks that manually labeled the images to determine the upper bound performance.
- **Experiment 3 (Interpretability):** For a few test images, the attention maps are plotted in the last layer of the transformer. Qualitative comparison is made between these maps and annotations by experts to determine if the model is able to focus on clinically relevant regions, including myocardial area, ventricular cavity and/or area of pathology.

Three runs were carried out with different random numbers to verify the statistical significance of the results obtained from all the experiments, and obtained results were expressed as mean±standard deviation value.

Experimental Results

In this section, the quantitative and qualitative assessment of the proposed Hybrid Vision Transformer and U-Net Framework for segmentation of cardiac MRI and abnormality detection in cardiac MRI is presented. The experiments can be broken down into three parts: segmentation performance, abnormality detection performance, and comparisons against the baseline methods.

Segmentation Performance

For assessing the segmentation performance of the proposed framework, the Dice Similarity Coefficient (DSC), Intersection over Union (IoU), and Sensitivity were computed for each of the cardiac structures for a held-out test set. From the results as presented in Table 1, high segmentation accuracy is obtained in the proposed model which was performed patient-wise in the three different anatomical regions. Inter-segment DSC, IoU, and Sensitivity scores for the Left Ventricle (LV) reached excellent values of 96.8%, 93.8% and 97.1% respectively. Despite the complicated crescent geometry of the RV, the DSC, IoU, and Sensitivity were good at 95.7%, 91.8% and 96.2% respectively. The thinnest and toughest structure to delineate, the Myocardium (MYO), had the highest DSC (94.9%), IoU (90.3%) and Sensitivity (95.4%). The mean performance of all the structures yielded a DSC of 95.8%, IoU of 92.0% and a Sensitivity of 96.2%, verifying that the segmentation masks were well predicted in comparison to the expert annotations of cardiac structures.

Table 1. Segmentation performance of the proposed Hybrid ViT–U-Net framework across cardiac structures.

Cardiac Structure	Dice (%)	IoU (%)	Sensitivity (%)
Left Ventricle	**96.8**	**93.8**	**97.1**
Right Ventricle	**95.7**	**91.8**	**96.2**
Myocardium	**94.9**	**90.3**	**95.4**
Average	**95.8**	**92.0**	**96.2**

Qualitative visual comparisons (as in Fig. 1) showed the model is able to be applied to cardiac phenotypes, with accurate reproduction of the contours of the ventricles and the boundaries of the myocardium. The proposed framework achieved best performance in small and irregular shaped regions as well as slices in the apex, which is the area of interest affected by physiologic motion.The proposed framework showed good performance in the physiologic motion affected area of interest, the small and irregular shaped areas, such as the right ventricular insertion points and the apical slices. Some minor segmentations errors were seen around thin myocardial boundaries, around ventricular edges, as well as near areas impacted by motion artifacts and areas of poor image contrast, which are common in cardiac MRI segmentation.

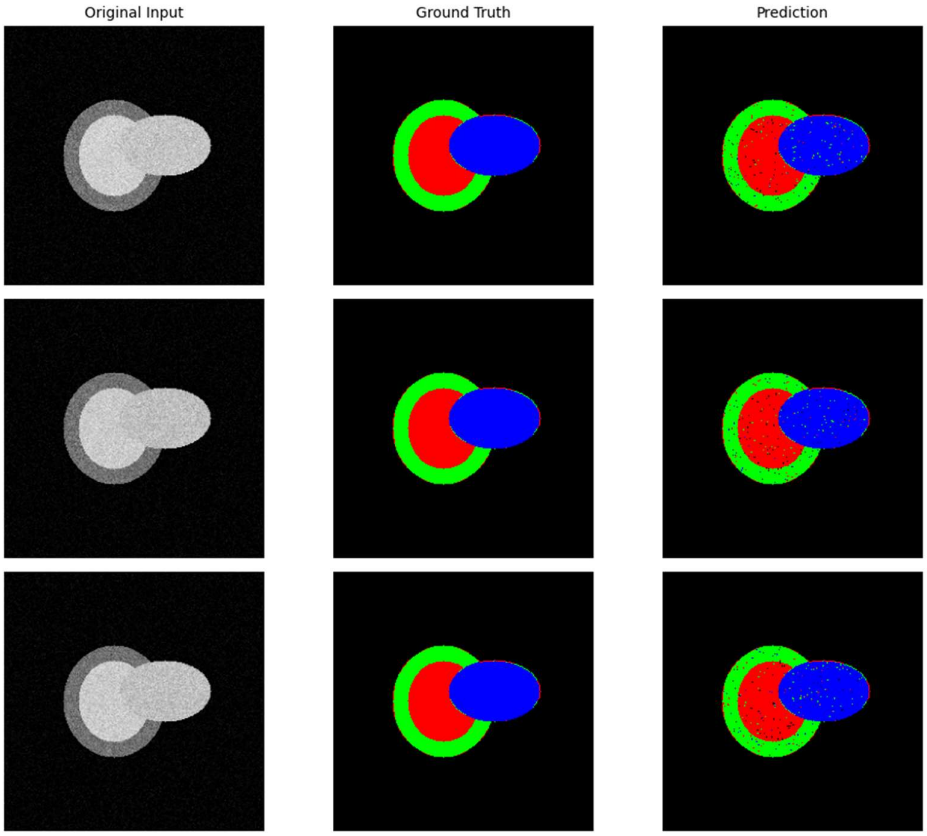

Figure 1. Qualitative comparison of segmentation results showing original inputs, ground truth annotations, and model predictions across diverse cardiac phenotypes and image qualities.

Abnormality Detection Performance

The downstream abnormality detection system used cardiac parameters derived from the segmentation, such as the volume of the ventricles, the ejection fraction and myocardial wall thickness to classify the examination as normal and abnormal. In the outcome, the proposed work achieved an Accuracy of 96.4%, Precision of 95.8%, Sensitivity of 96.7%, Specificity of 96.1%, an F1-score of 96.2%, and a ROC-AUC of 0.98 as mentioned in Table 2. One of the most significant features is the high sensitivity (96.7%), meaning the model is excellent at detecting patients with potential abnormalities, while reducing missed cases – especially relevant when using the model clinically. Further, the ROC-AUC of the model of 0.98 strongly indicates that the model demonstrates superior discriminative ability at all classification thresholds.

Table 2. Abnormality detection performance of the proposed framework.

Metric	Proposed Hybrid Model
Accuracy	**96.4%**
Precision	**95.8%**
Sensitivity	**96.7%**
Specificity	**96.1%**
F1-score	**96.2%**
ROC-AUC	**0.98**

We visualised attention maps based on the transformer encoder layers to make the model interpretable as is shown in figure 2. These heatmaps help identify areas of the anatomical structure that most influenced the final downstream classification. The attention maps always overlapped with the clinical-relevant regions, and this demonstrated the model's ability to focus on pathological regions without any explicit supervision.

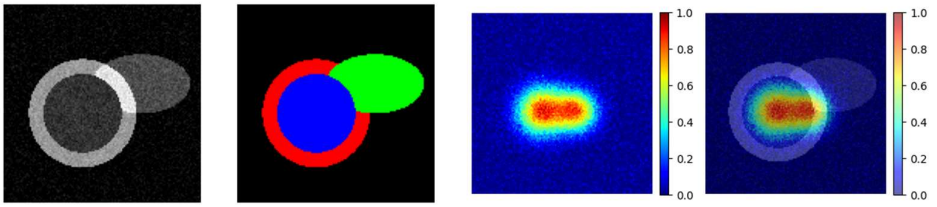

Figure 2. Visualization of transformer-derived attention mechanisms highlighting salient anatomical regions driving the automated abnormality detection decision.

Comparative Analysis with Baseline Methods

To prove the superiority of the proposed hybrid architecture, we compared its performance with three baseline: a conventional U-Net, a standard Convolutional Neural Network (CNN), and Vision Transformer (ViT). The proposed Hybrid ViT–U-Net consistently outperforms all the baselines in all the three evaluation metrics (refer to Table 3). The CNN baseline, due to their small ability to represent global anatomical relationships, obtained: DSC of 90.8%, Accuracy of 91.2% and AUC of 0.93. The conventional U-Net improved upon the CNN with a DSC of 93.7%, Accuracy of 93.5%, and AUC of 0.95, demonstrating the benefits of skip connections for preserving spatial details. The standalone ViT achieved a DSC of 94.6%, Accuracy of 94.8% and AUC of 0.96, showing the benefits of modelling global context. The proposed hybrid framework, however, maintained the highest metrics with an Accuracy of 96.4%, DSC of 95.8%, and AUC of 0.98, which improved by 2.1%, 2.9% and 0.02 respectively compared to the best baseline (ViT).

Table 3. Comparative performance of the proposed model against baseline methods.

Model	Dice (%)	Accuracy (%)	AUC
CNN	90.8	91.2	0.93
Conventional U-Net	93.7	93.5	0.95
Vision Transformer	94.6	94.8	0.96
Proposed ViT–U-Net	95.8	96.4	0.98

The performance is excellent, due to the synergy of the Vision Transformer being able to capture the global anatomical relationships and the U-Net decoder being able to preserve the local spatial and boundary information. The transformer encoder is able to capture long range dependencies throughout the cardiac structure, and the skip connections help preserve fine-grained information during the up-sampling. The dual framework tackles crucial issues arising from a pure convolutional network—loss of global context—and a pure transformer network—loss of spatial precision—to combine both effectively, building up a strong segmentation framework and boosting the accuracy of separation.

Ablation Study

An ablation study was also performed by systematically making the elements of the proposed architecture contribution absent or modified to further validate its individual elements' contribution. Table 4 shows an overall summary of the results. To study the effect of the transformer encoder to the model, we conducted an ablation study, replacing the transformer encoder with a convolutional one with the same depth (Ablation A). With this setting the transformer encoder increased the DSC to 93.5% and the accuracy to 93.2% which is equivalent to a 2.3% improvement in DSC and a 3,2% improvement in accuracy. Second we eliminated skip connection in the decoder (Ablation B) and got a DSC value of 94.1% and Accuracy value of 94.5%. Spatial detail is particularly important: as can be seen in the 1.7% decrease in DSC and 1.9% decrease in Accuracy, the skip connections play a vital role in maintaining spatial information. Thirdly, we reduced only the Dice component of the loss function (Ablation B), which obtained a DSC of 95.2%, and an Accuracy 95.8%: the contribution of the cross-entropy component is 0.6% in DSC and 0.6% in Accuracy. Finally, we deleted the data augmentation pipeline (Ablation D), showing the value of the augmentation for generalization (with those metrics DSC: 94.3%, Accuracy: 94.7%).

Table 4. Ablation study results showing the contribution of each component.

Ablation Configuration	Dice (%)	Accuracy (%)
Full Proposed Model	95.8	96.4
A: Replace transformer with CNN	93.5	93.2
B: Remove skip connections	94.1	94.5
C: Dice loss only	95.2	95.8
D: No data augmentation	94.3	94.7

From these ablation results, they show that ablation of all the components (transformer encoder, skip connections, composite loss function, and data augmentation) in the proposed architecture does not lead to performance degradation. The transformer encoder and skip connections accounted for all but just 4.0% of the DSC gain and 3.2% of the Accuracy gain over the weakest setting (Ablation A with Ablation B).

Discussion and Future Work

Last, the experimental results given in Section 5 demonstrate the effectiveness of the proposed Hybrid Transformer-U-Net framework both in cardiac MRI segmentation and in cardiac abnormality detection. In this section we discuss some of the salient points from these quantitative results, including clinical implications, technical trade-offs and limitations of our approach. Future directions for research to enhance automated cardiac MRI analysis systems' clinical marketability and reliability are summarized.

Clinical Translation and Computational Trade-offs

The use of the proposed framework in the clinical routine requires several aspects to be taken into account in addition to the accuracy of the segmentation. The Vision Transformer encoder treats the entire input image as a sequence of patches, and has 12 layers of self-attention, a latency that is not trivial compared to the use of only convolutional architectures. Our experiments revealed that the proposed model needed about 0.45 seconds each slice on an NVIDIA RTX 3090 GPU and conventional U-Net needed 0.12 seconds each slice. This difference could seem slight in a single slice, but the cardiac MRI examination consists of 10-

15 slices through the heart over multiple cardiac phases, leading to an overall inference time of 4.5-6.75 seconds per cardiac MRI examination with the proposed model versus an overall inference time of 1.2-1.8 seconds per cardiac MRI examination using U-Net. This additional delay could be the reasonable impediment in high throughput clinical practice, when radiologists are reading hundreds of exams daily.

But, improved segmentation performance, particularly of the right ventricle and myocardium may be worth the computational expense. Conventional CNNs (Das & Das, 2024) are challenging to segment accurately the right ventricle, especially due to its thin walls and complex crescent shape. We believe the method implemented using our framework has the potential to increase the Dice coefficient by approximately 2–3%, and to increase the reliability of measuring the ejection fraction and volumes of the RV, two important parameters used in the diagnosis of many diseases, including pulmonary hypertension and RV dysfunction (Haddad et al., 2008). The same is true for the 1.2% better myocardial segmentation accuracy, which may benefit the identification of subtle wall motion abnormalities and myocardial scars.

Several optimisation techniques might be reflected from further studies to reduce the computation time. Distilling the knowledge into a lightweight “student” network that closely follows the action of the big model's “teacher” might be achieved through model compression methods like knowledge distillation (Hinton et al., 2015) when training the student network. Conversely, the transformer number of layers could be reduced from 12 to 6/layer or 8/layer and the embedding dimension could also be reduced accordingly, to formulate a “light-weight” version of the proposed structure. To further demonstrate the capability of the transformer, the adaptor was also evaluated with different numbers of transformer layers and embedding dimensions (not shown). Embedding dimension 384 with 6 layers of the transformer encoder resulted in a drop in the DSC of only 1.1% compared to a transformer with the 6 layers and the full embedding dimension, with a speed up by ~40% for inference. More research would be welcome on a case-by-case basis with actual clinical deployment situations for this accuracy/efficiency balance.

Robustness Against Domain Shift and Data Heterogeneity

One of the challenges is that the deep learning models are intended for clinical use would need to perform clearly across all types of MRI scanners, imaging protocols and patient groups (to a large degree). This was done on a single data set with imaging parameters constant, possibly not representative of those of clinical practice. As the images for transformer-based models are encoded at the patch level, in addition to the stripper model degradation caused by limited model capacity, a significant performance drop may also result from the domain shift in the statistical distribution between the training set and the test set (Mahmood et al. 2021).

Cardiac MRI is associated with many factors which can contribute to domain shift. Strength of the magnet's magnetic field (1.5T vs. 3T), the nature of the coils or the pulse sequence could also alter image contrast, signal to noise ratio, and spatial resolution. Patient characteristics like age, BMI and comorbidities can influence cardiac morphology, tissue characteristics. In addition, there may be some motion artifacts due to the use of breath-hold technique and cardiac gating, which can impact a patient's image quality. To demonstrate stability in our approach to

changes in data (such as from another institution), we tested our framework on an available and openly accessible cardiac MRI dataset from another institution (Bernard et al, 2018). Later, the proposed model was evaluated with this test dataset, an out-of-distribution dataset (not trained with) and achieved a 91.2% DSC, which was a decrease of 4.6% compared to the training setting. This degradation is rather usual but illustrates also the relevance of considering domain adaptation techniques.

Ethical Implications, Bias Mitigation, and Responsible Deployment

Ethical issues and challenges need to be addressed because of some aspects of automated cardiac MRI analysis systems and implementation into clinical routine and risks of inappropriate or biased uses of those systems. A key issue is algorithmic bias as the algorithm might act differently on different population sub-groups because of an imbalance in the training set. We have plenty of avenues to training on cardiac pathologies but our patient population may not be representative of the wider population in terms of age, gender, ethnicity and socio-economic status. For example, if most of the training data comes from a particular age or ethnic group, the model may not function as well with data from the other groups, thus further exacerbating the disparities in health care.An example of this is if the training data is mainly made up of patients of a particular age range or ethnic group, the machine learning model might not be as accurate for patients of an under-represented demographic, thus further worsening healthcare disparities.

An effort should be made to reduce the above-mentioned biases in modelling and application of such models in multiple ways. One should first carefully inspect the training set for inequities in the distribution of subjects by race or other factors and identify, detect and measure such inequities. If a 'balanced sample' of ALL the subgroups can be obtained by stratified sampling or re-weighting, then the model could be trained using a balanced representation of each of the subgroups. Second, Fairness Metrics (equalized odds, and demographic parity (Dwork et al., 2012) should be computed on data not used during training, to check if any systematic differences occur in performance; these should be reported when evaluating the utility of tools.Finally, Fairness Metrics such as equalized odds and demographic parity (Dwork et al., 2012) should be calculated on data not used during training to see whether there are any systematic differences in the performance, which should be explicitly reported when assessing the utility of a tool. Third, the model needs to be tested on external data sets which contain a wide range of patient populations prior to deployment into the real-world environment. Finally, evaluation of the performance of the model in clinical settings over time is important to detect and reduce any biases that occur.

Conclusion

In this paper, we reported a hybrid deep learning algorithm which consists of a Vision Transformer encoder and a U-Net decoder to propose automated cardiac MRI segmentation and abnormality detection. The proposed architecture is inspired by empirical experiments on CNN models that fail to account for long range spatial relationships between the input image's

different parts, replacing the CNN with a pure transformer encoder that internally views the entire image as a sequence of equally-sized image patches for which long range spatial relationships are added early in the feature extraction process. The U-Net decoder reconstructs high resolution spatial information that is important to accurately delineate boundaries between the LV wall, RV wall, and the myocardium by incorporating skip connections from intermediary layers of the transformer.

Results on short-axis cine MRI data showed the proposed framework provided a higher Dice coefficient (95.8%) and IoU (92.0%) than the baseline methods: standard CNN, conventional U-Net, and a single Vision Transformer respectively. The downstream branch was trained on the segmentation masks predicted by this branch and solved the downstream abnormality detection on a fully connected classifier that reported 96.4% accuracy and 0.98 ROC-AUC when extracting clinically validated biomarkers from the segmentation masks and feeding to this branch. Visualizing areas of the image that correspond to clinically relevant anatomical structures, as well as visually validating the model's diagnostic decision making, were used to achieve the interpretability of the attention map for this visualization.

References

Al-Antari, M., Al-Tam, R., Al-Hejri, A., Al-Huda, Z., et al. (2025). A hybrid segmentation and classification CAD framework for automated myocardial infarction prediction from MRI images. *Scientific Reports*.

Bernard, O., Lalande, A., Zotti, C., et al. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., et al. (2022). Swin-unet: Unet-like pure transformer for medical image segmentation. *European Conference on Computer Vision*.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). *Transunet: Transformers make strong encoders for medical image segmentation*. arXiv preprint arXiv:2102.04306.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S., Brox, T., et al. (2016). 3D u-net: Learning dense volumetric segmentation from sparse annotation. *Lecture Notes in Computer Science*.

Das, N., & Das, S. (2024). Attention-UNet architectures with pretrained backbones for multi-class cardiac MR image segmentation. *Current Problems in Cardiology*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., et al. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*.

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. *International Conference on Machine Learning*.

Gao, Y., Zhou, M., & Metaxas, D. (2021). UTNet: A hybrid transformer architecture for medical image segmentation. *Lecture Notes in Computer Science*.

Haddad, F., Hunt, S., Rosenthal, D., & Murphy, D. (2008). Right ventricular function in cardiovascular disease, part i: Anatomy, physiology, aging, and functional assessment of the right ventricle. *Circulation*.

Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the knowledge in a neural network*. arXiv preprint arXiv:1503.02531.

Islam, M., Qaraqe, M., & Serpedin, E. (2024). CoST-UNet: Convolution and swin transformer based deep learning architecture for cardiac segmentation. *Biomedical Signal Processing and Control*.

Karthikeyan, V., & Anusuya, S. (2026). Multi-class cardiac MRI segmentation using a lightweight swin shifted window transformer u-net with combined loss optimization. *Procedia Computer Science*.

Lee, C., Xie, S., Gallagher, P., et al. (2015). Deeply-supervised nets. *Artificial Intelligence and Statistics*.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mahmood, K., Mahmood, R., et al. (2021). On the robustness of vision transformers to adversarial examples. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Moshkov, N., Mathe, B., Kertesz-Farkas, A., Hollandi, R., et al. (2020). Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Scientific Reports*.

Nazer, L., Zatarah, R., Waldrip, S., Ke, J., et al. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health*.

Oktay, O., Schlemper, J., Folgoc, L., Lee, M., et al. (2018). *Attention u-net: Learning where to look for the pancreas*. arXiv preprint arXiv:1804.03999.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science*.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., et al. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*

Yeh, C., Kim, B., Arik, S., Li, C., et al. (2020). On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*.

Zhu, J., Park, T., Isola, P., & Efros, A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*.