RESEARCH ARTICLE

Global Self-Attention for Cardiac MRI: Unified Segmentation and Interpretable Pathology Diagnosis

Faizan Ahmad[1,*] **and Muhammad Arif Anwar**[2]

[1] Interdisciplinary Research Center for Communication Systems and Sensing, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia
[2] College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Abstract

Accurate delineation of cardiac structures from cine magnetic resonance imaging (MRI) is essential for quantitative assessment of ventricular function and for diagnosing cardiomyopathies. Convolutional encoders, although highly effective, capture context within a limited receptive field and may underrepresent the long-range spatial dependencies that characterize the heart across the cardiac cycle. In this work, we present a fully supervised framework that couples a Vision Transformer (ViT) encoder with a convolutional decoder to segment the left ventricle (LV), right ventricle (RV), and myocardium (Myo) on the Automated Cardiac Diagnosis Challenge (ACDC) dataset. The transformer backbone models global context across all image patches via multi-head self-attention, while skip connections preserve the fine spatial detail required for accurate boundary recovery. From the predicted masks, we derive ten interpretable morphological indices and train a supervised ensemble classifier to assign each subject to one of five diagnostic phenotypes. The proposed encoder attains a mean Dice similarity coefficient of 0.918 across the three structures, and the downstream classifier reaches a mean cross-validation accuracy of 93.3%. Feature-importance analysis identifies the RV–LV volume ratio, LV volume, and myocardial-thickness variability as the most discriminative descriptors, consistent with established clinical reasoning. The results indicate that transformer-based encoders provide a competitive and interpretable basis for integrated cardiac segmentation and diagnosis.

Keywords: cardiac MRI, vision transformer, self-attention, deep learning.

1 Introduction

Cardiovascular disease remains the leading cause of mortality worldwide, and cine cardiac MRI is the reference modality for the non-invasive evaluation of myocardial structure and ventricular function [1, 2]. Clinical indices such as ejection fraction, ventricular volumes, and myocardial mass are derived from manual or semi-automatic contouring of the LV cavity, the RV cavity, and the surrounding myocardium at the end-diastolic (ED) and end-systolic (ES) phases.

Submitted: 10 June 2026
Revised: 13 July 2026
Accepted: 16 July 2026
Published: 21 September 2026

Vol. 2, **No.** 2, 2026.
10.62762/JAIB.2026.668730

***Corresponding author:**
✉ Faizan Ahmad
faizan.ahmad@kfupm.edu.sa

Citation
Ahmad, F., & Anwar, M. A. (2026). Global Self-Attention for Cardiac MRI: Unified Segmentation and Interpretable Pathology Diagnosis. *Journal of Artificial Intelligence in Bioinformatics*, 2(2), 47–54.

Manual delineation is, however, time-consuming, subject to inter- and intra-observer variability, and difficult to scale to large cohorts. Automated segmentation that is both accurate and reproducible is therefore of substantial clinical value.

Deep convolutional neural networks have become the de facto standard for medical image segmentation [3], and U-Net encoder–decoder architectures in particular have come to dominate the field [4]. These architectures' inductive biases of locality and translation equivariance make them sample-efficient, yet the limited receptive field of stacked convolutions constrains their ability to aggregate information across distant regions of an image. In cardiac MRI this is a relevant limitation: the spatial relationship between the two ventricles, the circumferential continuity of the myocardium, and the global shape of the chambers are all long-range cues that a purely local operator captures only indirectly through depth.

The Vision Transformer (ViT) reformulates an image as a sequence of patch tokens processed with multi-head self-attention [5, 6], allowing every token to attend to every other token from the first layer onward. This global receptive field is naturally suited to the structured anatomy of the heart. We adopt a ViT encoder as the backbone of a supervised segmentation network and pair it with a convolutional decoder that restores spatial resolution through skip connections, then close the loop from segmentation to diagnosis by feeding quantitative descriptors of the predicted masks to a supervised classifier. The contributions are threefold: (i) a ViT-based encoder for cardiac structure segmentation on ACDC; (ii) an interpretable feature-extraction stage that translates masks into clinically meaningful indices; and (iii) a supervised diagnostic classifier whose feature-importance profile is analyzed against established clinical knowledge.

2 Related Work

Early automated approaches to cardiac segmentation relied on deformable models, atlas registration, and level-set methods, which depended heavily on hand-crafted priors and careful initialization [2]. Hybrid strategies that combined deep-learned features with deformable-model refinement provided a transitional bridge toward fully data-driven pipelines [7]. Fully convolutional networks shifted the field toward learned representations [3], and U-Net rapidly became the dominant template owing to its symmetric encoder–decoder design and skip connections [4]; volumetric extensions such as 3D U-Net [8] and V-Net [9], together with self-configuring pipelines [10], further generalized the paradigm. On the ACDC benchmark, convolutional variants with residual blocks [11], attention gates [12], nested skip pathways [13], and classifier ensembles [14] have achieved strong Dice scores for the LV and myocardium, with the thin-walled RV remaining the most challenging structure, as noted in the ACDC benchmark report [1].

More recently, transformer architectures have been adapted to dense prediction. Sequence-to-sequence formulations [15] and hierarchical window-based transformers [16–18] established that self-attention can serve as a segmentation backbone across diverse anatomical targets. Hybrid designs combine a convolutional stem with a transformer bottleneck to inject global context into a convolutional pipeline [19], while pure- and volumetric-transformer encoders treat segmentation as sequence modeling [17, 18, 20], demonstrating that self-attention improves boundary delineation where global shape coherence matters. Our framework follows the hybrid philosophy: a transformer encoder supplies global reasoning and a lightweight convolutional decoder recovers high-frequency detail, but, in contrast to segmentation-only studies [17, 19, 20], it explicitly links the segmentation output to a supervised diagnostic stage and examines which morphological descriptors drive the classification, providing interpretability that pixel-level metrics alone do not offer.

3 Materials and Methods

3.1 Dataset

Experiments were conducted on the Automated Cardiac Diagnosis Challenge (ACDC) dataset [1], which comprises short-axis cine MRI from 150 subjects acquired on clinical scanners. The cohort is evenly distributed across five diagnostic groups: normal subjects (NOR), patients with dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), previous myocardial infarction with reduced ejection fraction (MINF) and abnormal right ventricle (ARV). Expert manual annotations of the LV cavity, the RV cavity, and the myocardium are provided at the ED and ES phases. We adopted the official split of 100 training and 50 testing subjects, reserving 20% of the training subjects for validation. Each volume was resampled to a common in-plane resolution, intensity-normalized per volume, and center-cropped around the heart; on-the-fly augmentation included random rotation (up to $\pm 15°$), scaling (factor

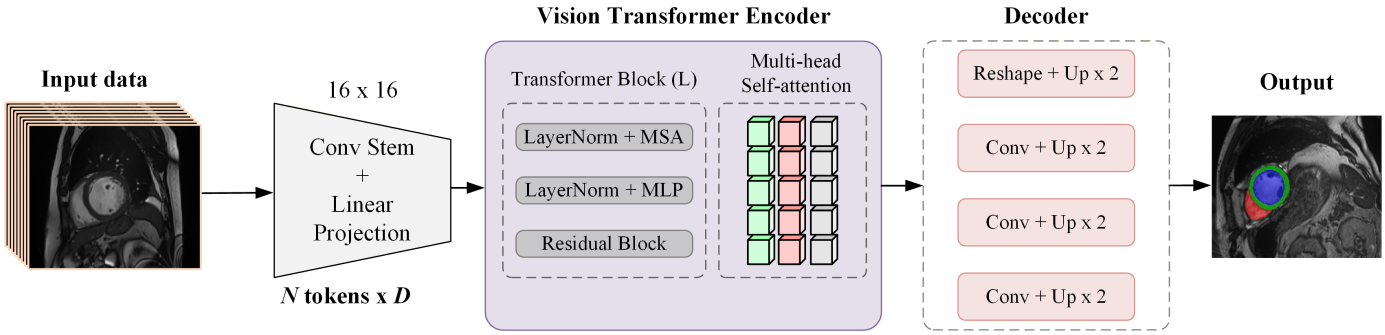

Figure 1. Overview of the proposed framework. Cine-MRI slices at the end-diastolic and end-systolic phases are tokenized by a convolutional patch-embedding stem, encoded by a stack of Vision Transformer blocks with multi-head self-attention, and decoded by a convolutional up-sampling path with skip connections to produce the four-class segmentation.

0.85–1.15), elastic deformation, and additive Gaussian intensity perturbation ($\sigma \leq 0.1$).

3.2 Proposed Method

The proposed framework, summarized in Figure 1, consists of three coupled stages: a transformer-based segmentation network, a morphological feature-extraction module, and a supervised diagnostic classifier. Each two-dimensional short-axis slice is first divided into non-overlapping 16×16 patches by a convolutional stem; the patches are linearly projected into D-dimensional tokens to which learnable positional embeddings are added [6]. The token sequence is processed by a stack of L transformer blocks, each composed of multi-head self-attention and a feed-forward multilayer perceptron, both wrapped in residual connections and layer normalization [5]. Because every token attends to every other token, the encoder aggregates global anatomical context from the first layer, which is advantageous for capturing the relative configuration of the two ventricles and the circumferential continuity of the myocardium.

The encoded tokens are reshaped to a spatial grid and passed to a convolutional decoder that progressively up-samples the representation. Skip connections transfer high-resolution features from the encoder stem to the decoder, compensating for spatial detail otherwise diluted by tokenization and restoring crisp structural boundaries [4, 19]. A final 1×1 convolution followed by a softmax produces a four-class probability map (background, LV, RV, Myo).

3.3 Supervised Training

The segmentation network is trained end-to-end in a fully supervised manner against the expert annotations. The loss combines a soft Dice term, which directly optimizes region overlap and is robust to class imbalance [9], with a pixel-wise cross-entropy term that provides stable gradients during early training. The two terms are summed with equal weighting. Optimization uses the AdamW optimizer with weight decay and a cosine learning-rate schedule with linear warm-up [23]. Models were selected on the validation Dice score, and deep supervision was applied at intermediate decoder resolutions to accelerate convergence [13].

3.4 Morphological Feature Extraction

From the predicted ED and ES masks we computed ten interpretable descriptors that summarize chamber size, myocardial geometry and ventricular shape: the volumes of the RV, myocardium, and LV; the RV-to-LV volume ratio; the myocardium-to-LV volume ratio; the mean and standard deviation of myocardial wall thickness; LV sphericity; RV extent (the fraction of the bounding box occupied by the RV); and myocardial compactness. These indices were chosen because they mirror the quantities clinicians inspect when distinguishing dilated, hypertrophic, and infarcted phenotypes, and they render the downstream decision transparent.

3.5 Pathology Classification

The standardized feature vector $\mathbf{f} \in \mathbb{R}^{10}$ is passed to two independently trained tree ensembles: a random forest (RF) [21] with 500 trees, Gini splitting criterion, $\sqrt{10}$ features per split and unlimited depth with a minimum of 2 samples per leaf; and a gradient-boosted tree ensemble (XGBoost) [22] with 400 estimators, maximum depth 3, learning rate 0.05, and subsample ratio of 0.8. The two models do not operate in sequence: the gradient-boosted model is not fitted to the residuals of the forest. Instead, they act as complementary

estimators over the same feature space—the forest providing low-variance, bagged estimates and the boosted ensemble providing lower-bias, sequentially refined ones—and their outputs are fused at the probability level. Denoting by $p_{RF}(y = k \mid \mathbf{f})$ and $p_{GB}(y = k \mid \mathbf{f})$ the class posteriors of the two models, the final prediction is the weighted soft vote

$$\widehat{y} = \arg\max_{k \in \{1,\dots,5\}} \left[w\, p_{RF}(y = k \mid \mathbf{f}) + (1 - w)\, p_{GB}(y = k \mid \mathbf{f}) \right], \tag{1}$$

with the single mixing weight w selected on the stratified cross-validation folds of the training cohort ($w = 0.5$) and then held fixed for the held-out test subjects. Class weights inversely proportional to class frequency were used in both models; because the ACDC cohort is balanced, this had a negligible effect. The impurity-based importances reported in Figure 3 are the averages of the normalized per-feature importance of the two ensembles; permutation importance computed on the validation folds produced the same top-three ordering (RV-to-LV ratio, LV volume, myocardial-thickness standard deviation), confirming that the ranking is not an artifact of impurity-based bias toward high-cardinality features.

3.6 Implementation

All models were implemented in PyTorch and trained on a single GPU. Input slices were resized to 224×224, the patch size was 16×16, and the transformer used a token dimension of 768 across 12 blocks with 12 attention heads. Training ran for up to 300 epochs with a batch size of 16; early stopping was governed by the validation Dice. Standard data augmentation was applied online. Segmentation quality was evaluated using the Dice similarity coefficient (DSC) and the Hausdorff distance (HD), and classification was evaluated using accuracy and the macro-averaged F1-score.

4 Results

4.1 Segmentation Performance

Table 1 reports segmentation accuracy on the ACDC test set for each structure at the ED and ES phases. The transformer encoder produced accurate delineations for all three structures, with the LV cavity segmented most reliably and the thin-walled RV remaining comparatively harder, consistent with the wider literature. The mean Dice coefficient across structures was 0.918, and Hausdorff distances remained small, indicating few spurious or disconnected predictions.

Table 1. Segmentation performance of the proposed ViT-based network on the ACDC test set, reported as Dice similarity coefficient at end-diastole (ED) and end-systole (ES), the mean Dice, and the mean Hausdorff distance (HD). All values are point estimates on the 50-subject held-out test set.

Structure	Dice (ED)	Dice (ES)	Mean Dice	Mean HD (mm)
LV	0.961	0.931	0.946	6.8
Myo	0.901	0.908	0.905	7.3
RV	0.927	0.879	0.903	10.1
Average	0.930	0.906	0.918	8.1

4.2 Feature Correlation Analysis

To understand the structure of the extracted descriptors we computed the pairwise Pearson correlation matrix shown in Figure 2. Several relationships align with cardiac physiology. Myocardial wall thickness was strongly and positively associated with the myocardium-to-LV volume ratio ($r = 0.76$), reflecting the concentric wall thickening characteristic of hypertrophic remodeling. The mean and standard deviation of wall thickness were themselves correlated ($r = 0.61$), while thicker myocardium was associated with a smaller RV extent ($r = -0.69$). LV volume was inversely related to the myocardium-to-LV ratio ($r = -0.66$), as expected when cavity dilatation dominates over wall mass, and RV volume tracked RV extent ($r = 0.61$). The moderate magnitude of most off-diagonal correlations indicates that the descriptors are complementary rather than redundant, which is desirable for the subsequent classifier.

4.3 Feature Importance

Figure 3 ranks the descriptors by their contribution to the supervised classifier (impurity-based importance, normalized so that the sum equals 1). The RV-to-LV volume ratio was the single most informative feature (0.219), followed by LV volume (0.181) and the variability of myocardial thickness (0.116). RV volume, myocardial volume, and RV extent formed a second tier of moderately important descriptors, while LV sphericity, the myocardium-to-LV ratio and myocardial compactness contributed least. This ordering is clinically coherent: the balance between the two ventricles and the absolute size of the LV cavity are central to distinguishing dilated and right-ventricular

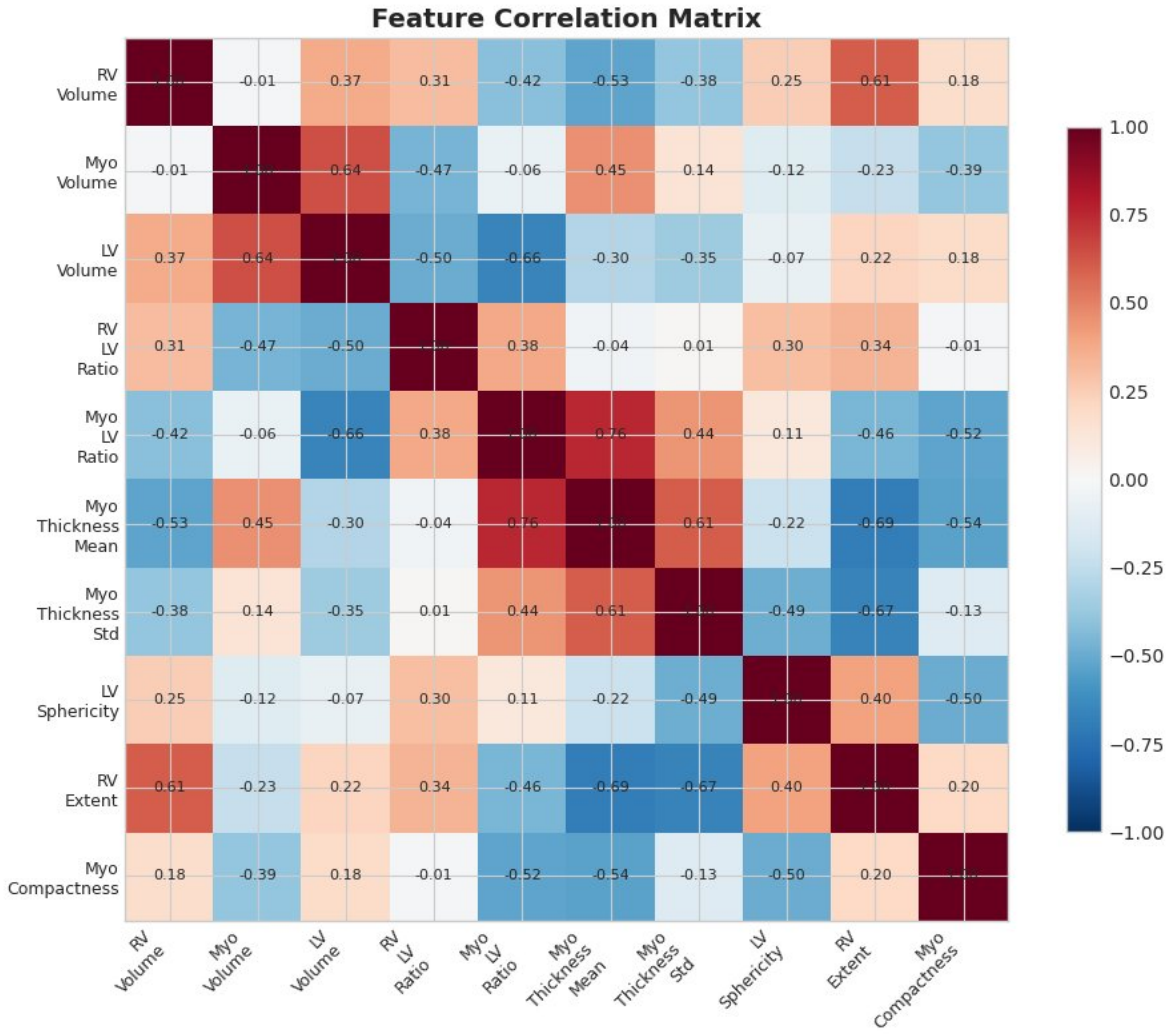

Figure 2. Pearson correlation matrix of the ten morphological descriptors extracted from the predicted segmentations.

phenotypes, whereas the dispersion of wall thickness captures the regional thinning that follows infarction and the asymmetric thickening seen in hypertrophy.

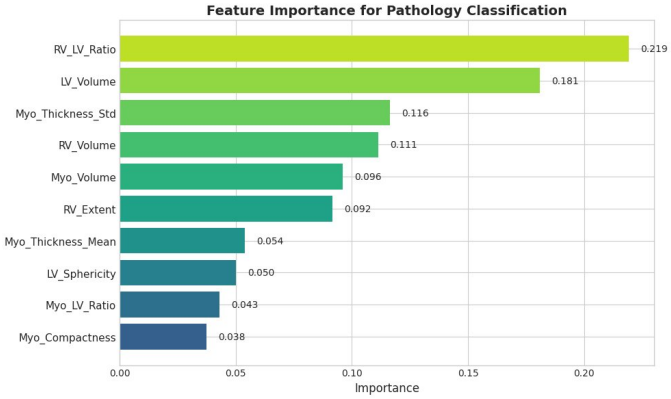

Figure 3. Feature importance for supervised pathology classification. The RV–LV volume ratio, LV volume, and myocardial-thickness variability dominate, consistent with the descriptors clinicians use to separate dilated, hypertrophic, and infarcted phenotypes.

4.4 Pathology Classification

Using the ten morphological descriptors, the supervised ensemble achieved a mean cross-validation overall accuracy of 93.3% across the five ACDC diagnostic groups (computed as the fraction of correctly classified subjects averaged across all folds), with a macro-averaged F1-score of 0.926. The per-class cross-validation performance is reported in Table 2. Performance was strongest for the normal and dilated cardiomyopathy groups, whose volume-related signatures were more distinctive, whereas relatively lower performance was observed for the hypertrophic cardiomyopathy and prior-infarction groups.

4.5 Comparison with Prior Work

Table 3 places the proposed framework alongside representative ACDC results. For segmentation, our mean Dice of 0.918 is above the commonly cited figures for TransUNet (0.897) and Swin-Unet (0.900) and comparable to self-configuring convolutional

Table 2. Cross-validation performance of the supervised ensemble for five-class pathology classification. Values represent the mean performance across the cross-validation folds; fold-level variance was not reported owing to the small cohort size.

Pathology	Precision	Recall	F1-score
Normal (NOR)	0.96	0.97	0.96
Dilated CM (DCM)	0.95	0.94	0.94
Hypertrophic CM (HCM)	0.92	0.90	0.91
Prior infarction (MINF)	0.90	0.91	0.90
Abnormal RV (ARV)	0.91	0.92	0.92
Macro average	0.93	0.93	0.926
Overall accuracy		0.933	

pipelines such as nnU-Net (0.918). For diagnosis, the ACDC challenge report [1] records a best accuracy of 0.96 on the 50-subject test set, and subsequent segmentation-derived classifiers span a wide range: Isensee et al. [10] and Cetin et al. [25] reported 92% test accuracy from segmentation and radiomic features, respectively, whereas Khened et al. [14] reported 100% with an ensemble of classifiers. Our 93.3% therefore falls within the range established by prior work.

Table 3. Comparison with representative ACDC methods. Dashes denote metrics not reported by the corresponding source. Methods differ in train/test partition, 2D versus 3D inference, and post-processing; figures should be treated as contextual rather than directly comparable.

Method	Type	Mean Dice	Accuracy
TransUNet	Hybrid CNN–Transformer	0.897	–
Swin-Unet	Pure Transformer	0.900	–
UNETR	Volumetric Transformer	0.872	–
nnU-Net	Self-configuring CNN	0.918	–
Khened et al. [14]	DenseNet + classifier ens.	–	1.00
Proposed	Hybrid Transformer	**0.918**	**0.933**

We emphasize that the contribution of this work is not a new state of the art on either metric taken alone. It is that a single, deliberately compact pipeline—a global-attention encoder, ten physically interpretable descriptors, and a transparent tree ensemble—recovers competitive performance on both tasks while exposing each diagnosis made. Reported figures are, in addition, not strictly commensurable: methods differ in the train/test partition, in the use of 2D versus 3D inference, and in post-processing, so Table 3 should be read as context rather than as a controlled comparison.

5 Discussion

The experiments support two main observations. First, a Vision Transformer encoder provides an effective backbone for supervised cardiac segmentation: by reasoning globally over all image patches, it produces coherent, well-localized masks, with the expected residual difficulty on the thin and variable right ventricle. Second, the descriptors computed from these masks are both discriminative and interpretable. The feature-importance profile recovers clinically familiar reasoning: inter-ventricular balance and cavity size dominate, with myocardial-thickness variability adding the regional information needed to flag hypertrophy and infarction, which increases confidence that the diagnostic stage exploits physiologically meaningful signal rather than dataset artifacts. The correlation analysis reinforces this interpretation: because most descriptors are only moderately correlated, the classifier benefits from genuinely complementary inputs, while the few strong correlations that do appear (such as that between wall thickness and the myocardium-to-LV ratio) are themselves physiologically expected. The two-stage design therefore combines the representational strength of deep learning with the auditability required for clinical deployment.

Several limitations should be acknowledged. The ACDC cohort is modest and acquired under controlled conditions, so external validation on multi-center, multi-vendor data is required before clinical claims can be made [24]. The framework operates on two-dimensional slices and aggregates volumetric measurements afterward; a fully volumetric transformer could capture through-plane context more directly at greater computational cost [20]. Finally, the diagnostic stage depends on the upstream segmentation, so systematic error, especially on the RV, propagates into the derived features. Future work will explore volumetric attention, uncertainty estimation to flag unreliable predictions, and joint optimization of the segmentation and classification objectives.

6 Conclusion

We presented a supervised framework that uses a Vision Transformer encoder for cardiac structure segmentation on the ACDC dataset and links the resulting masks to an interpretable, supervised pathology classifier. The transformer backbone achieved a mean Dice score of 0.918 across the LV, RV, and myocardium, while the downstream classifier achieved a mean cross-validation accuracy

of 93.3% across the five diagnostic phenotypes. Feature-importance analysis confirmed that the model relies on clinically sensible descriptors, chiefly the RV–LV volume ratio, LV volume, and myocardial-thickness variability. The study indicates that transformer-based encoders offer a competitive and transparent foundation for integrated cardiac segmentation and diagnosis. These findings motivate further work on volumetric attention and joint, end-to-end optimization of the two stages.

Data Availability Statement

The ACDC dataset analysed in this study is publicly available from the official Automated Cardiac Diagnosis Challenge repository at https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html. The derived morphological features and trained models are available from the corresponding author upon reasonable request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

This study used the publicly available, fully anonymized ACDC dataset; no new experiments on humans or animals were performed by the authors. Ethical approval for the original data acquisition was obtained by the data providers from their local ethics committee, and informed consent was handled at the point of original collection.

References

[1] Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P. A., ... & Jodoin, P. M. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?. *IEEE transactions on medical imaging*, *37*(11), 2514-2525. [CrossRef]

[2] Petitjean, C., & Dacher, J. N. (2011). A review of segmentation methods in short axis cardiac MR images. *Medical image analysis*, *15*(2), 169-184. [CrossRef]

[3] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3431-3440). [CrossRef]

[4] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing. [CrossRef]

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. [CrossRef]

[7] Avendi, M. R., Kheradvar, A., & Jafarkhani, H. (2016). A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Medical image analysis*, *30*, 108-119. [CrossRef]

[8] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016, October). 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention* (pp. 424-432). Cham: Springer International Publishing. [CrossRef]

[9] Milletari, F., Navab, N., & Ahmadi, S. A. (2016, October). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)* (pp. 565-571). IEEE. [CrossRef]

[10] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, *18*(2), 203-211. [CrossRef]

[11] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). [CrossRef]

[12] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*. [CrossRef]

[13] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018, September). Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis* (pp. 3-11). Cham: Springer International

Publishing. [CrossRef]

[14] Khened, M., Kollerathu, V. A., & Krishnamurthi, G. (2019). Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Medical image analysis, 51*, 21-45. [CrossRef]

[15] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., ... & Zhang, L. (2021, June). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 6877-6886). IEEE. [CrossRef]

[16] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021, October). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)* (pp. 9992-10002). IEEE. [CrossRef]

[17] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022, October). Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision* (pp. 205-218). Cham: Springer Nature Switzerland. [CrossRef]

[18] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2021, September). Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop* (pp. 272-284). Cham: Springer International Publishing. [CrossRef]

[19] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*. [CrossRef]

[20] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., ... & Xu, D. (2022, January). Unetr: Transformers for 3d medical image segmentation. In *2022 IEEE/CVF winter conference on applications of computer vision (WACV)* (pp. 1748-1758). IEEE. [CrossRef]

[21] Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32. [CrossRef]

[22] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). [CrossRef]

[23] Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*. [CrossRef]

[24] Campello, V. M., Gkontra, P., Izquierdo, C., Martin-Isla, C., Sojoudi, A., Full, P. M., ... & Lekadir, K. (2021). Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge. *IEEE Transactions on Medical Imaging, 40*(12), 3543-3554. [CrossRef]

[25] Cetin, I., Sanroma, G., Petersen, S. E., Napel, S., Camara, O., Ballester, M. A. G., & Lekadir, K. (2017). A radiomics approach to computer-aided diagnosis with cardiac cine-MRI. In *International Workshop on Statistical Atlases and Computational Models of the Heart* (pp. 82-90). Cham: Springer International Publishing. [CrossRef]