Which Anatomy Matters Under Limited Labels? A Data-Efficient Anatomy-Aware Benchmark for Cardiac Pathology Prediction

Himanshu Singh * [1]

Abstract

Numerous medical imaging problems must be solved under limited labels and constrained compute, yet it remains unclear whether performance gains are driven mainly by more expressive models or by better representation of clinically meaningful anatomy. We study this question through a low-data anatomy-aware benchmark for 5-class cardiac pathology prediction on the public ACDC MRI dataset. Using segmentation-derived patient descriptors from the right ventricle, myocardium, and left ventricle, we compare anatomy-specific and multi-structure representations across linear, kernel, and tree-based classifiers. We find that under limited label settings, representation dominates complexity. These results suggest that in resource-constrained healthcare settings, identifying and representing the most informative anatomy may matter more than the increasing complexity of the model alone.

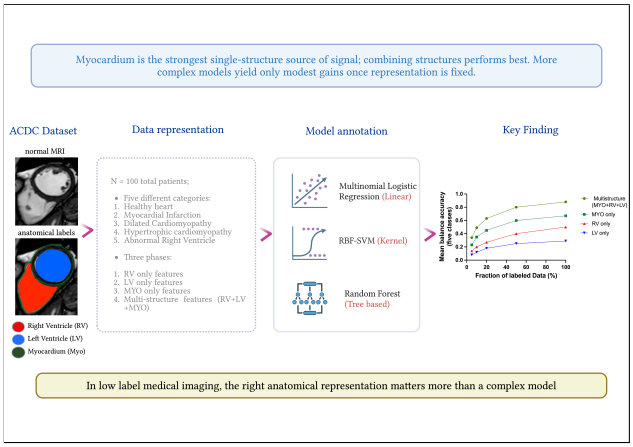

Figure 1. **Representation before complexity.** Our benchmark asks whether, under limited labels, selecting the right anatomical representation matters more than increasing model complexity. We illustrate this by decomposing a representative short-axis cardiac MR image into RV-only, MYO-only, LV-only, and ALL-structures views, which form the basis of the anatomy ablation study. Background cardiac MR image adapted from Adomat et al. (2026).

1. Introduction

In the low-label medical imaging regime (Jin et al., 2026; Zhou et al., 2021), researchers often attempt to improve performance using more complex models. However, it remains unclear whether the real bottleneck is the complexity of the model (Varoquaux & Cheplygina, 2022) or how the clinical structure is represented. In particular, in cardiac imaging (Buja & Butany, 2022; Counseller & Aboelkassem, 2023; Flachskampf et al., 2015), this question is especially relevant because the pathology is expressed through anatomically meaningful structure rather than arbitrary input variation.

On the other hand, in AI for medical imaging, practical bottlenecks often lie not only in model design but also in data preparation, annotation, and deployment infrastructure (Willemink et al., 2020). These constraints are especially acute in resource-constrained healthcare settings, where the limited radiology infrastructure can hinder the adoption of compute-intensive AI pipelines (Yousef & Schmollgruber, 2024).

In the present work, inspired by previous benchmark-driven progress in medical imaging (Bernard et al., 2018; Blagec et al., 2023), we address this research direction through a reproducible low-data benchmark built on Automated Cardiac Diagnosis Challenge (ACDC) MRI dataset (Bernard et al., 2018). The intuition behind our benchmark is illustrated in Figure 1, where a representative cardiac MR image is decomposed into structure-specific views to address which anatomy carries the dominant predictive signal.

This question is also well motivated by prior cardiac MRI diagnosis pipelines built on segmentation-derived representations. In particular, Isensee et al. (Isensee et al., 2017) and Khened et al. (Khened et al., 2019) combined segmentation outputs with clinically inspired handcrafted features

[1]Computing & Artificial Intelligence Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. Correspondence to: Himanshu Singh (https://himanshuvnm.github.io/) <singh_h@lanl.gov>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

for automatic disease assessment on ACDC, while Zheng et al. (Zheng et al., 2019) showed that explainable cardiac pathology classification can be achieved by combining shape-related features with motion characterization. These works demonstrate that anatomically grounded descriptors can support interpretable pathology prediction; our focus is to understand, under limited labels, which anatomical structures carry the dominant predictive signal and how much that matters relative to classifier complexity.

Our Contributions. We construct an anatomy-aware benchmark for low-data cardiac pathology prediction using ACDC segmentation masks. We then show that the myocardial morphology is the strongest single-structure source of predictive signal, while multi-structure anatomical representation yields the best overall performance. Finally, we show that simple handcrafted inter-phase delta features do not improve over static multi-structure descriptors, and we validate the benchmark using label-shuffle controls, confusion analysis, and patient-level visualization.

2. Anatomy-Aware Benchmark Setup & Models

2.1. Dataset and Task

To begin with, we consider the balanced 5-class pathology setting on ACDC, which are dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), myocardial infarction (MINF), normal subjects (NOR), and abnormal right ventricle (RV). In this dataset, each class contains 20 patients, giving a total of 100 subjects. For brevity, we use the abbreviations DCM, HCM, MINF, NOR, and RV throughout the remainder of the paper. For each patient, we use the annotated segmentation masks provided at the labeled cardiac phases and build patient-level features from the three principal anatomical structures: RV, MYO, LV.

Task. Let $\{(x_i, y_i)\}_{i=1}^{N}$ denote the patient-level dataset, where $y_i \in \{1, \dots, K\}$ is the pathology label, and x_i denotes the cardiac MRI study for patient i; N is the number of patients and K denotes the number of classes. From each study, we construct anatomy-aware feature representations

$$\phi_r(x_i) \in \mathbb{R}^{d_r}, \qquad r \in \mathcal{R},$$

where $\mathcal{R} := \{\text{RV-only}, \text{MYO-only}, \text{LV-only}, \text{ALL}\}$. For each representation r, we train classifiers $f_\theta^{(r)} \in \mathcal{F}$, where $\mathcal{F}$ includes linear, kernel, and tree-based models, to predict

$$\hat{y}_i = f_\theta^{(r)}(\phi_r(x_i)).$$

Our goal is not only to maximize predictive performance, but also to identify which anatomical representation r contributes the strongest signal under limited labels and whether the variation across representations is larger than the variation across classifier families.

2.2. Patient-Level Anatomical Features

In order to isolate the structure-specific signal, we define four primary feature configurations: RV-only, MYO-only, LV-only, and ALL-structures, where the latter concatenates descriptors from all three anatomical compartments; refer Figure 2 for a representative visual.

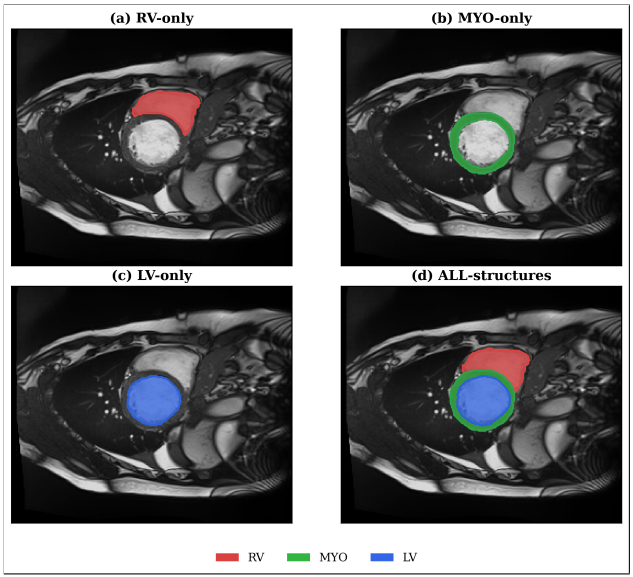

Figure 2. **A representative visual for the primary anatomical feature configurations used in the benchmark.**

Figure 2 is constructed from a representative ACDC subject by selecting a labeled cardiac phase and displaying the slice with the largest total annotated area. The same underlying MRI slice is then shown in four structure-selection settings: RV-only, MYO-only, LV-only, and ALL-structures. For each labeled frame and each anatomical structure, we extract simple shape descriptors from the binary segmentation mask, including area, area fraction, aspect ratio, principal-axis statistics, elongation, compactness, circularity, extent, and radial distance summaries. Slice-wise features are aggregated to the patient level by mean and standard deviation across the slices, together with the number of slices containing the structure.

2.3. Evaluation Protocol

We evaluated models using accuracy, balanced accuracy and macro-F1 under 5-fold stratified cross-validation following standard multiclass classification metrics (Sokolova & Lapalme, 2009; Powers, 2020); refer Appendix A for their expression. All preprocessing steps, including median imputation and feature standardization when required, are

fit within each training fold and then applied to the corresponding validation fold. For label-fraction experiments, we additionally repeat random subsampling at each fraction and report mean and standard deviation across repeated runs.

2.4. Models

We compare three lightweight model families: multinomial logistic regression as a linear baseline (Hosmer Jr et al., 2013; Hastie, 2009), RBF-SVM as a nonlinear kernel baseline (Cortes & Vapnik, 1995), and random forest as a tree-based nonlinear baseline (Breiman, 2001). Additional details of the model that we considered are given in Table 1, Sub-section A.2. Beyond aggregate predictive performance, we evaluated label efficiency, anatomy ablation, dynamic-feature enhancement, sanity checks, robustness, and feature-level interpretability.

3. Results

Our results address four complementary questions: whether the benchmark remains informative under limited labels, which anatomical structure carries the strongest predictive signal, whether simple inter-phase dynamic summaries add value beyond static anatomy-aware features, and whether the observed gains survive basic sanity checks.

3.1. Label Efficiency

We begin by evaluating whether segmentation-derived anatomical descriptors carry a meaningful pathological signal in the low-data regime. We observe that across repeated label-fraction sweeps, performance remains above chance ($1/5 = 0.2$, shown by the blue dashed line in Figure 3) and improves gradually as additional labels are incorporated. Hence, this shows that the benchmark is neither trivial nor noise-dominated, and that simple patient-level anatomical features already support nontrivial multiclass prediction. A lightweight end-to-end ResNet-18 (He et al., 2016) baseline trained on representative raw MRI slices performed substantially worse than all three anatomy-aware baselines (Appendix E-Table 2), reinforcing the value of explicit anatomical representation in the low-label regime.

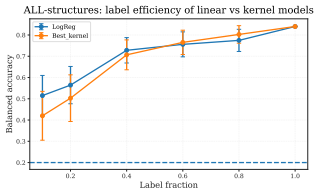

Figure 3. **Balanced accuracy across label fractions** for 5-class ACDC pathology prediction using the all-structure representation.

3.2. Which Anatomy Matters?

Now, we analyze the anatomy ablation study where among single-structure feature sets, MYO-only performs best, substantially outperforming both LV-only and RV-only representations, while the full multi-structure representation performs best overall. This suggests that the pathology signal is not distributed uniformly throughout the anatomy. Instead, myocardial morphology appears to concentrate the strongest single-structure information in this benchmark.

From Figure 4, we see that the myocardial descriptors are the strongest single-structure feature set, while the combined multi-structure representation performs best overall. In particular, the gain from moving from RV-only to MYO-only is much larger than the gain from switching among linear, kernel, and tree-based classifiers once the representation is fixed. A finer-grained decomposition of feature importance by anatomical structure and descriptor family is provided in Figure 13, Appendix D, further supporting the dominant role of myocardial descriptors.

We highlight that this result is important for exactly two reasons. First, on scientific grounds, our result suggests that the benchmark is driven less by generic whole-heart geometry and more by structure-specific morphology, with the myocardium serving as the most informative individual component. Second and lastly, from an empirical ground, our result provides an informative lesson for low-resource medical ML: when building simplified or compute-efficient pipelines under limited labels, the myocardium may be the most valuable single anatomical target to prioritize. All-structure representation still performs best, but the MYO result identifies where much of the predictive signal is already concentrated.

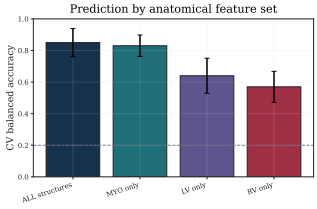

Figure 4. **Cross-validated balanced accuracy across anatomical feature** sets for 5-class ACDC pathology prediction. Myocardium is the strongest single-structure representation, while combining RV, myocardium, and LV yields the best overall performance.

3.3. Do Explicit Dynamic Features Help?

To assess whether simple cardiac phase dynamics add information beyond static anatomical representation, we augment the full feature set with explicit inter-phase delta and ratio descriptors. These additions do not materially improve over the static multi-structure representation. We address this negative result cautiously.

One possibility is that ACDC pathology groups are already strongly reflected in static morphology, particularly in myocardial structure, so that simple phase-difference summaries contribute little additional information. Another possibility is that our handcrafted dynamic descriptors are too compressed to preserve the richer spatial deformation patterns present across phases. Thus, our result should not be read as evidence that dynamics are uninformative in general; rather, it shows that simple low-dimensional inter-phase summaries do not outperform already-strong anatomy-aware static representations in this benchmark.

For sanity check, under random label permutation, the balanced accuracy drops from 0.870 ± 0.057 to 0.230 ± 0.057, which is close to chance for a balanced 5-class task. This supports the interpretation that the observed gains arise from a genuine anatomical signal rather than leakage or spurious shortcut cues in the dataset.

3.4. Why Does MYO Matter?

We provide a plausible explanation for the MYO result as follows: several ACDC pathologies are expressed strongly through the morphology of the myocardial wall rather than the geometry of the chamber geometry. Therefore, MYO descriptors capture clinically meaningful variation in shape, extent, circularity, elongation, and radial-distance structure, making the myocardium the most informative single anatomical compartment in the benchmark. Figure 5 quantitatively helps explain the result of anatomy ablation by aggregating feature importance at the structure level.

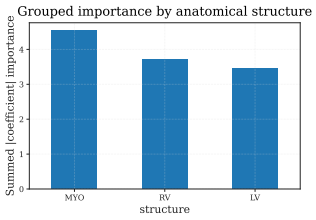

Figure 5. **Grouped feature importance by anatomical structure.** Summed absolute logistic-regression coefficients are highest for myocardium, reinforcing the quantitative ablation result that MYO is the strongest single-structure source of predictive signal.

4. Discussion and Limitations

Our key finding is that, in this low-data cardiac pathology benchmark, the dominant factor is the anatomical representation rather than the complexity of the classifier. Our reproducible benchmark experiments suggest that kernel and tree-based models provide limited gains beyond a strong anatomy-aware representation, while explicit handcrafted dynamic summaries add little further improvement. We note that our study also has limitations as we consider a single public dataset, rely on handcrafted segmentation-derived descriptors rather than raw-image end-to-end learning, and study dynamics only through simple inter-phase summaries. Our future work can extend this framework to additional datasets, uncertainty-aware analysis, richer temporal descriptors, and external validation across institutions.

5. Conclusion

We empirically show that across anatomy ablations, model comparisons, dynamic-feature tests, and sanity checks, the representation choice matters more than classifier complexity. These results suggest that in low-label structured medical learning, identifying the right anatomical factor may be more important than choosing a more expressive classifier.

Acknowledgments

The author acknowledges the kind effort of PAYAL GHOSH (PG) in making Figure 1. Additionally, we also acknowledge PG for helpful discussion and suggestions that were incorporated in the paper.

Impact Statement

Our work has a broader impact in the design of practical medical AI systems.

Disease-dependent anatomy-aware modeling. More generally, different cardiac conditions can shift the anatomical priority, suggesting that future low-label benchmarks should identify and emphasize structures that carry the most clinically meaningful information for the task at hand. This perspective may be relevant for the Global South settings, where data analysis capacity are limited and where anatomically focused representations may offer a more practical path toward deployable medical AI.

Data-efficient medical AI. Our central message suggests that practical gains in medical AI may come not only from larger architectures but also from choosing the right anatomical representation. Recent learning from radiology and medical imaging has emphasized that limited data, limited expert labeling, and limited annotation resources present challenges for real-world model development (Candemir et al., 2021; Willemink et al., 2020).

Resource-constrained healthcare settings. Our results support a complementary actionable principle in resource-constrained healthcare settings: instead of assuming that better performance requires heavier end-to-end models, it may be more effective to identify the anatomical structures that carry the clinically meaningful signal and represent them explicitly.

References

Adomat, F., Schaub, C., Hoh, T., Fischer, X., Guggenberger, R., Manka, R., Eberhard, M., and Weber, L. Cardiac MR function analysis with DL-based super resolution reconstruction: application in the clinical setting. *The International Journal of Cardiovascular Imaging*, pp. 1–11, 2026. URL https://link.springer.com/article/10.1007/s10554-026-03642-8.

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M. A. G., et al. 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, 2018. URL https://ieeexplore.ieee.org/document/8360453/.

Blagec, K., Kraiger, J., Frühwirt, W., and Samwald, M. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. *Journal of Biomedical Informatics*, 137:104274, 2023. URL https://www.sciencedirect.com/science/article/pii/S1532046422002799.

Breiman, L. Random forests. *Machine learning*, 45(1): 5–32, 2001.

Buja, L. M. and Butany, J. *Cardiovascular pathology*. Academic Press, 2022.

Candemir, S., Nguyen, X. V., Folio, L. R., and Prevedello, L. M. Training Strategies for Radiology Deep Learning Models in Data-limited Scenarios. *Radiology: Artificial Intelligence*, 3(6):e210014, 2021. URL https://pubs.rsna.org/doi/10.1148/ryai.2021210014.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. URL https://link.springer.com/article/10.1023/A:1022627411411.

Counseller, Q. and Aboelkassem, Y. Recent technologies in cardiac imaging. *Frontiers in medical technology*, 4:984492, 2023. URL https://www.frontiersin.org/journals/medical-technology/articles/10.3389/fmedt.2022.984492/full.

Flachskampf, F. A., Biering-Sørensen, T., Solomon, S. D., Duvernoy, O., Bjerner, T., and Smiseth, O. A. Cardiac Imaging to Evaluate Left Ventricular Diastolic Function. *JACC: Cardiovascular Imaging*, 8(9):1071–1093, 2015. URL https://www.jacc.org/doi/epdf/10.1016/j.jcmg.2015.07.004.

Hastie, T. The elements of statistical learning: data mining, inference, and prediction, 2009.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. *Applied logistic regression*. John Wiley & Sons, 2013.

Isensee, F., Jaeger, P. F., Full, P. M., Wolf, I., Engelhardt, S., and Maier-Hein, K. H. Automatic Cardiac Disease Assessment on cine-MRI via Time-Series Segmentation and Domain Specific Features. In *International workshop on statistical atlases and computational models of the heart*, pp. 120–129. Springer, 2017. URL https://link.springer.com/chapter/10.1007/978-3-319-75541-0_13.

Jin, C., Guo, Z., Lin, Y., Luo, L., and Chen, H. Learning with less supervision: A survey of label-efficient learning for medical image analysis. *Medical Image Analysis*, pp. 104062, 2026. URL https://www.sciencedirect.com/science/article/abs/pii/S1361841526001301?via%3Dihub.

Khened, M., Kollerathu, V. A., and Krishnamurthi, G. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Medical Image Analysis*, 51:21–45, 2019. URL https://www.sciencedirect.com/science/article/abs/pii/S136184151830848X?via%3Dihub.

Powers, D. M. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *arXiv preprint arXiv:2010.16061*, 2020. URL https://arxiv.org/pdf/2010.16061.

Sokolova, M. and Lapalme, G. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

Varoquaux, G. and Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5 (1):48, 2022. URL https://www.nature.com/articles/s41746-022-00592-y.

Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L. R., Summers, R. M., Rubin, D. L., and Lungren, M. P. Preparing Medical Imaging Data for Machine Learning. *Radiology*, 295 (1):4–15, 2020. URL https://pubs.rsna.org/doi/10.1148/radiol.2020192224.

Yousef, K. and Schmollgruber, S. Artificial Intelligence in Low- and Middle-Income Countries: Reducing the Gaps in Health Care, Research, and Education. *International Journal of Critical Care*, 18(2):1–3, 2024. URL https://wfccn-ijcc.com/index.php/ijcc/article/view/981.

Zheng, Q., Delingette, H., and Ayache, N. Explainable cardiac pathology classification on cine mri with motion characterization by semi-supervised learning of apparent flow. *Medical Image Analysis*, 56:80–95, 2019. URL https://www.sciencedirect.com/science/article/abs/pii/S1361841519300519?via%3Dihub.

Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., Prince, J. L., Rueckert, D., and Summers, R. M. A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises. *Proceedings of the IEEE*, 109(5):820–838, 2021. URL https://ieeexplore.ieee.org/document/9363915.

A. Evaluation and Model Details

A.1. Metric Definitions

Let C_{ij} denote the confusion matrix entry counting examples whose true class is i and predicted class is j. For class i, the recall is defined as follows: $\mathrm{Recall}_i = \frac{C_{ii}}{\sum_j C_{ij}}$. Then, *balanced accuracy* BA is defined as the average recall across classes: $\mathrm{BA} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{Recall}_i$, where K is the number of classes. This metric is appropriate here because the benchmark is multiclass and class-conditional performance is of primary interest. In the usual manner, *accuracy* is defined as follows: $\mathrm{Acc} = \frac{\sum_i C_{ii}}{\sum_{i,j} C_{ij}}$. For macro-F1, let the precision for class i be $\mathrm{Precision}_i = \frac{C_{ii}}{\sum_j C_{ji}}$, and define the classwise F1 score by $\mathrm{F1}_i = \frac{2\,\mathrm{Precision}_i\,\mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i}$. The macro-F1 score is then $\text{Macro-F1} = \frac{1}{K}\sum_{i=1}^{K} \mathrm{F1}_i$.

A.2. Model Hyperparameters

Other important details related to the models that we consider in the paper.

Table 1. Model families and representative hyperparameter settings used in the main experiments.

Model	Role in study	Main settings
Multinomial logistic regression	Linear baseline for testing whether simple decision boundaries suffice once anatomical representation is fixed	Multinomial formulation; median imputation; feature standardization; default ℓ_2 regularization; trained with standard convex optimization
RBF-SVM	Nonlinear kernel baseline for testing whether modestly more expressive decision boundaries improve over the linear baseline	RBF kernel; $C = 10$; `gamma=scale`; one-vs-rest multiclass decision function; median imputation; feature standardization
Random forest	Tree-based nonlinear baseline for testing whether ensemble partitioning improves over linear and kernel models	300 trees; median imputation; no feature standardization; default impurity-based splits and bagging-style ensemble aggregation

B. Additional Qualitative Analysis

To visually complement the anatomical ablation study, we align and average structure-specific masks within each pathology class and plot the resulting class prototypes; this is visually shown in Figure 6.

B.1. Representative Failure Cases

To complement the confusion analysis, we visualize representative misclassified patients to illustrate the kinds of anatomical ambiguity that remain under the all-structure representation; this is shown in Figure 7.

C. Additional Quantitative Analysis

C.1. Normalized Confusion Matrix

To characterize residual class-wise ambiguity, we compute the row-normalized confusion matrix. Let C_{ij} denote the number of examples whose true class is i and predicted class is j. We normalize each row by the total number of examples in the corresponding true class:

$$\widetilde{C}_{ij} = \frac{C_{ij}}{\sum_k C_{ik}}.$$

Thus, $\widetilde{C}_{ij}$ represents the fraction of class-i examples assigned to class j, and each row sums to one. Figure 8 shows that the remaining errors are structured rather than random. DCM and RV are especially recovered well, while MINF and NOR account for most of the residual ambiguity. This pattern suggests that the benchmark is meaningful but not saturated, with failures concentrated in a small number of clinically plausible class pairs.

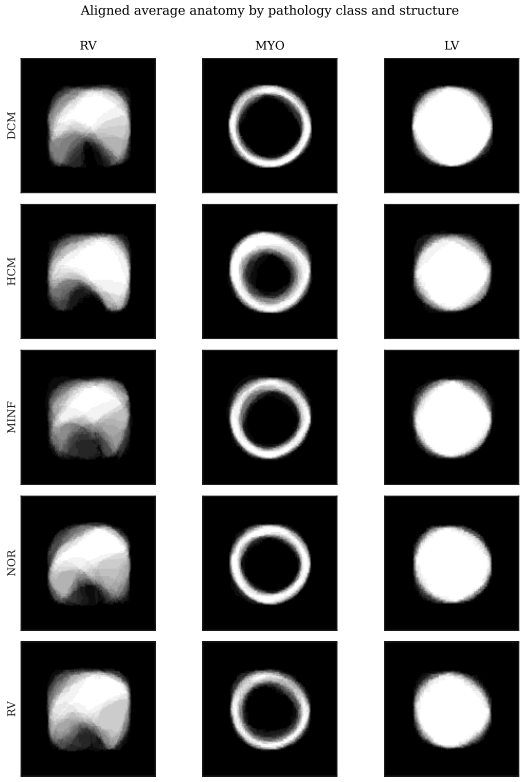

Figure 6. Aligned class prototypes for RV, myocardium, and LV across ACDC pathologies. Masks were centered and size-normalized before averaging. Myocardial contours exhibit the clearest class-dependent variation, consistent with the quantitative finding that myocardium is the strongest single-structure feature set.

C.2. Robustness to Imperfect Segmentations

To assess whether the anatomy-aware pipeline remains reliable under realistic segmentation imperfections, we simulate mild boundary perturbations through mask erosion and dilation and re-evaluate the benchmark under the same cross-validation protocol. Figure 9 tests whether the pipeline remains reliable under imperfect segmentations. Performance remains relatively stable under mild erosion and dilation perturbations, suggesting that the anatomy-aware representation is not unduly brittle to realistic contour variability. This is particularly relevant for deployment in resource-constrained settings, where segmentation quality may vary due to annotation differences or heterogeneous imaging conditions.

C.3. Feature-Level and Representation Analysis

Figure 10 refines the MYO story by showing which individual descriptors drive the classification. Many of the top-ranked features are myocardial in origin, with only a smaller contribution from RV-derived quantities. This suggests that the predictive value of MYO is not a coarse artifact of grouping, but is expressed through specific shape statistics extracted from the myocardium itself.

C.4. Patient embedding

Figure 11 shows that the resulting patient-level feature space has visible class structure, supporting the interpretation that the representation captures clinically meaningful anatomical organization rather than random variation.

C.5. Most important myocardial descriptor families

Figure 12 further decomposes the myocardial contribution into descriptor families. Radial-distance variability, extent, circularity, elongation, and compactness dominate, indicating that the MYO signal arises from geometry and morphological variation rather than from a single scalar quantity. This deepens the interpretability of the benchmark and suggests which

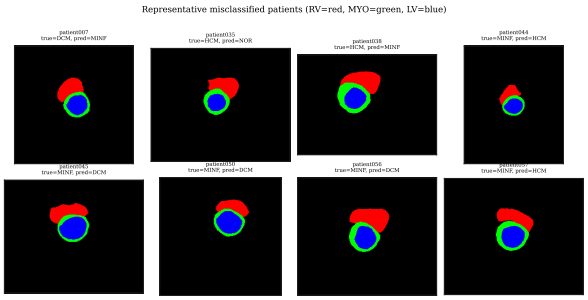

Figure 7. Representative misclassified patients under the all-structure model. RV, myocardium, and LV are shown in red, green, and blue, respectively. The errors are concentrated in anatomically ambiguous cases rather than random failures, especially for MINF-related confusions.

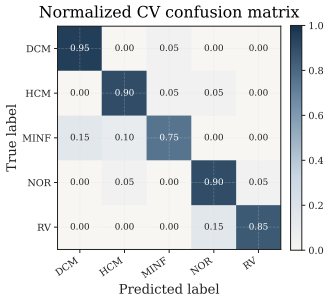

Figure 8. Normalized cross-validated confusion matrix for 5-class ACDC pathology prediction using the all-structure anatomical representation. Most classes are well separated, while residual errors are concentrated in a small number of structured class pairs, particularly NOR–RV and MINF–DCM, indicating a meaningful but nontrivial benchmark.

types of myocardial structure are most informative with limited labels.

D. Structure-by-Descriptor-Family Importance

In Figure 13, the heatmap provides a finer-grained interpretability view of the anatomy-aware benchmark and further highlights the dominant contribution of myocardial descriptors.

D.1. Robustness

E. End-to-End Image Baseline

To contextualize anatomy-aware baselines, we additionally compared with a lightweight end-to-end image classifier based on ResNet-18 (parameter counts 11,175,941) trained on representative raw MRI slices under the same low-label protocol.

Table 2. Comparison of lightweight anatomy-aware baselines and a small end-to-end image baseline (ResNet-18). The anatomy-aware models operate on segmentation-derived patient descriptors, whereas ResNet-18 operates on representative raw MRI slices. This comparison contextualizes the tradeoff between predictive performance and computational cost in low-resource settings.

Model	Input Type	Balanced Acc.	Macro-F1	Train time/fold (s)	Infer time/fold (s)
LogReg	anatomy features	0.85 ± 0.089	0.849 ± 0.089	-	-
RBF-SVM	anatomy features	0.87 ± 0.051	0.870 ± 0.051	-	-
RandomForest	anatomy features	0.86 ± 0.097	0.853 ± 0.102	-	-
ResNet18 (2-channel)	raw MRI slices	0.41 ± 0.074	0.380 ± 0.083	5.45	0.02

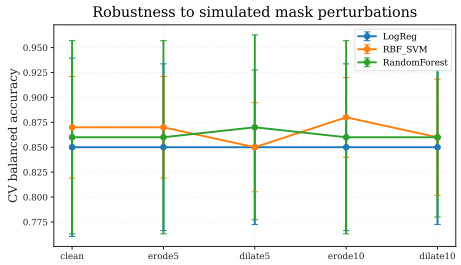

Figure 9. Robustness to simulated mask perturbations. We erode and dilate segmentation masks by small amounts to mimic annotation disagreement or lower-quality imaging conditions. Performance remains relatively stable across mild perturbations, suggesting that the anatomy-aware pipeline is robust to realistic segmentation noise.

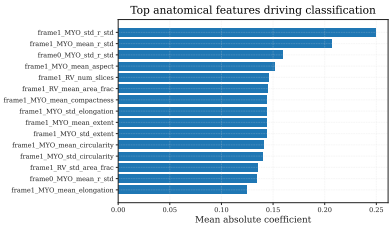

Figure 10. Top anatomical features driving classification under the logistic-regression model, ranked by mean absolute coefficient magnitude. Many of the most influential descriptors arise from myocardial morphology, with additional contribution from a smaller set of RV-derived features.

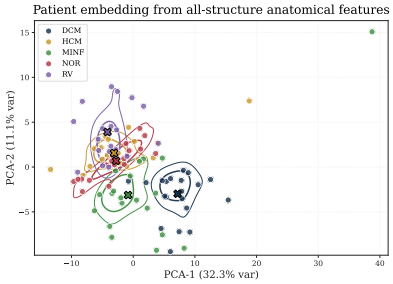

Figure 11. PCA embedding of patients in the all-structure anatomical feature space. The classes exhibit visible organization without becoming trivially separable, indicating that the anatomy-aware representation captures meaningful pathology structure while preserving nontrivial class overlap.

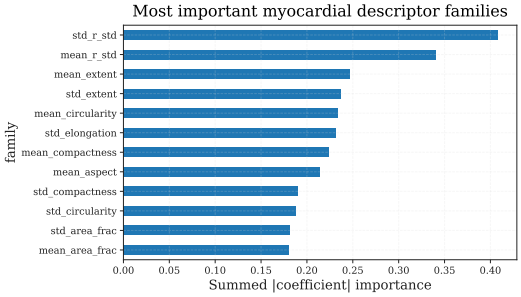

Figure 12. Most important myocardial descriptor families under the logistic-regression model. Radial-distance variability, extent, circularity, elongation, and compactness emerge as the most influential myocardial descriptor groups, suggesting that the predictive value of MYO arises from geometry rather than a single scalar measurement alone.

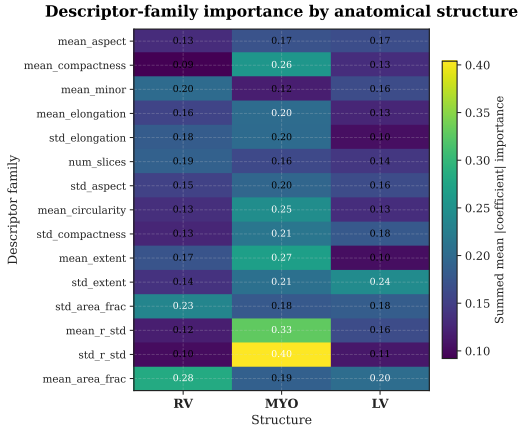

Figure 13. Structure-by-descriptor-family importance analysis using multinomial logistic regression coefficients. Each cell reports the summed mean absolute coefficient importance for a descriptor family within a given anatomical structure, averaged across cross-validation folds.

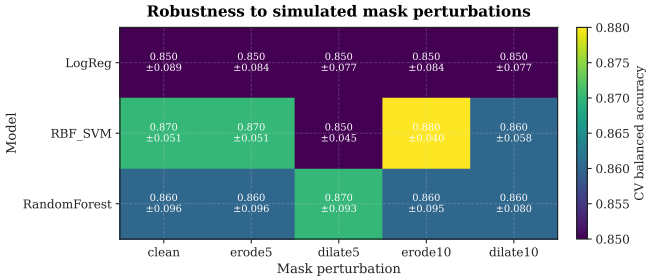

Figure 14. Robustness to simulated mask perturbations. Each cell reports mean cross-validation balanced accuracy ± standard deviation under mild erosion and dilation of the segmentation masks. Performance remains stable across perturbation settings for logistic regression, RBF-SVM, and random forest, suggesting that the anatomy-aware pipeline is robust to modest contour variability.