Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable Deep Learning for Cardiac MRI: Multi-Stage Segmentation, Cascade Classification, and Visual Interpretation

Vitalii Slobodzian
Khmelnytskyi National University
Barmak Oleksandr
barmako@khmnu.edu.ua

Khmelnytskyi National University
Pavlo Radiuk
Khmelnytskyi National University
Liliana Klymenko
Shupyk National Healthcare University of Ukraine
Iurii Krak
Taras Shevchenko National University of Kyiv

Article

Keywords:

Posted Date: April 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-5930463/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Hierarchical Deep Learning Approach to Clinically Interpretable Cardiac MRI Diagnosis

Vitalii Slobodzian[1], Oleksandr Barmak[1,*], Pavlo Radiuk[1], Liliana Klymenko[2], and Iurii Krak[3,4]

[1]Khmelnytskyi National University, Department of Computer Science, Khmelnytskyi, 29016, Ukraine
[2]Shupyk National Healthcare University of Ukraine, Department of Family Medicine and Outpatient Care, Kyiv, 04112, Ukraine
[3]Taras Shevchenko National University of Kyiv, Department of Theoretical Cybernetics, Kyiv, 03680, Ukraine
[4]V.M. Glushkov Institute of Cybernetics, Laboratory of Communicative Information Technologies, Kyiv, 03187, Ukraine
[*]barmako@khmnu.edu.ua

ABSTRACT

The modern analysis of cardiac magnetic resonance imaging is fundamentally challenged by the heart's complex and dynamic anatomy, coupled with a persistent clinical mistrust of opaque artificial intelligence systems that lack interpretability. These challenges hinder the widespread adoption of automated diagnostics. In this work, we propose a hierarchical deep learning approach that delivers high accuracy and full interpretability. Using the ACDC dataset, our method employs structure-specific U-Nets to segment cardiac regions, achieving state-of-the-art mean Dice scores of 0.974 for the left ventricle and 0.920 for the end-systolic myocardium. A novel cascade classifier subsequently diagnoses five pathologies with an overall accuracy of 97%, outperforming standard baselines. Furthermore, the system resolves the interpretability gap by translating its predictions into clinician-friendly indices, such as ejection fractions and standardized wall-thickness maps. Cross-dataset testing on the M&Ms challenge confirms the model's robustness, demonstrating its potential for reliable clinical deployment.

1 Introduction

Over recent years, artificial intelligence (AI) has driven significant advancements in medical diagnostics, enabling automated analysis of complex images to assist physicians in making informed decisions[1,2]. Such technologies can substantially improve the efficiency and standardization of diagnostic workflows[3].

Cardiovascular diseases remain the leading cause of mortality worldwide, claiming an estimated 17.9 million lives annually and creating an urgent need for precise and timely diagnostics[4]. Cardiac magnetic resonance imaging (MRI) has emerged as the clinical gold standard for this purpose, providing exceptional soft-tissue contrast and detailed functional information that surpasses other modalities like computed tomography or ultrasound[5]. Its accuracy and reliability are well-established[6]. However, the process of translating these rich images into actionable clinical insights is laden with persistent and multifaceted challenges.

Although standardized imaging protocols are employed to assess cardiac structures and function[7], the difficulties in processing cardiac MRI scans can be formalized into three key areas. First, there are significant technical and acquisition-related challenges. Image quality is often degraded by motion artifacts from respiration and cardiac pulsation[8]. Second, the heart's anatomical and physiological complexity presents a formidable obstacle. Delineating the thin, crescent-shaped right ventricle (RV) is notoriously difficult compared to the more circular left ventricle (LV). Third, although current MRI scanners are equipped with advanced software such as CVI42 that enables automated myocardial contour segmentation, meaning segmentation is no longer considered a major limitation in clinical practice, the subsequent manual analysis required for comprehensive interpretation remains a clinical workflow bottleneck, susceptible to significant inter- and intra-observer variability and cognitive biases, which undermines diagnostic consistency[9,10].

AI, particularly deep learning (DL), promises to overcome these issues, yet the adoption of conventional DL models in clinical practice is severely hampered by the "black box" problem[11]. While these systems can achieve high accuracy, their opaque decision-making processes erode clinical trust. To be viable, AI systems must evolve beyond the limitations of generic visualization methods[12]. While general-purpose explainability techniques, such as Local Interpretable Model-agnostic Explanations (LIME)[13], SHapley Additive exPlanations (SHAP)[14], or Gradient-weighted Class Activation Mapping (Grad-CAM)[15] have been applied to medical imaging, their outputs often highlight pixel-level importance, which may not directly

translate into clinically actionable insights. A growing consensus, articulated in reviews like Salih et al.[16], emphasizes that for AI to be truly adopted, its explanations must align with the established diagnostic frameworks and quantitative metrics used by clinicians[17]. Our work directly addresses this need by moving beyond saliency maps to produce interpretations grounded in familiar clinical indices such as ejection fractions and standardized wall-thickness models, thereby bridging the gap between algorithmic output and clinical practice.

The goal of this study is to improve cardiac MRI diagnostics by developing a hierarchical DL approach that provides accurate, automated segmentation and classification while translating its decisions into clinically interpretable metrics. We argue that such a system is necessary to solve the intertwined problems of anatomical complexity and clinical opacity. Our main contributions directly address these challenges:

- Hierarchical Segmentation with Structure Specialization. To overcome the challenge of complex cardiac anatomy, we introduce a novel approach that decomposes segmentation into six specialized models. By training separate models to first localize and then contour each specific structure, i.e., LV, RV, and myocardium (Myo), our system learns their unique morphological features. This results in superior performance, particularly for the challenging RV and end-systolic (ES) Myo, compared to monolithic, one-size-fits-all models[18–21].
- Clinically Calibrated Cascade Classification. To resolve the diagnostic ambiguity that plagues standard multiclass classifiers, we designed a four-step cascade model. This architecture breaks down the complex five-way classification into a series of simpler, more robust binary decisions, significantly reducing confusion between similar pathologies like dilated cardiomyopathy and myocardial infarction, which is a well-known issue in prior works[22–24].
- Interpretation Aligned with Clinical Practice. To solve the critical "black box" problem, we created an end-to-end pipeline whose final output is not just a class label but a dashboard of canonical clinical metrics. By automatically deriving LV/RV volumes, ejection fractions, and American Heart Association 17-segment wall-thickness maps from its segmentations, our system presents its findings in a language that clinicians can immediately understand, verify, and trust[25,26].

The paper is structured as follows: Section 2 reviews contemporary DL-based approaches for segmenting, classifying, and interpreting cardiac MRI data. Section 3 describes the three proposed methods: a multi-stage segmentation technique integrating U-Net and ResNet architectures, a cascade classification procedure tailored for several target classes, and an interpretation method that highlights clinical features such as myocardial thickness and ejection fraction. Section 4 presents the results of the proposed methods, comparing them to state-of-the-art techniques. Section 5 draws conclusions, addresses limitations, and outlines possible future research directions.

2 State of the Art

The automated analysis of cardiac MRI scans is a dynamic field of research, with progress revolving around three interconnected pillars: segmentation, classification, and interpretation. A critical review of the current literature in these areas reveals significant advancements but also persistent gaps that hinder the translation of DL models into routine clinical practice. This section analyzes the state of the art to build a compelling case for the necessity of a new, integrated approach that addresses these shortcomings head-on.

2.1 Segmentation of Cardiac MRIs

Accurate segmentation, i.e., the delineation of anatomical structures like the LV, RV, and Myo, is the foundational step for any quantitative cardiac analysis. While traditional image processing techniques, such as those based on edge detection[27], thresholds[28], or region-growing[29], offer computational simplicity, they lack the robustness and accuracy required for complex clinical scenarios and have been largely superseded by DL-based methods[30].

DL, particularly through the use of convolutional neural networks (CNNs), has revolutionized medical image segmentation[31]. Seminal architectures like U-Net[19] and SegNet[32] established the foundation for medical image segmentation due to their ability to learn intricate hierarchical features from data. In recent years, several advanced U-Net variants and other architectures have been developed specifically for cardiac structure segmentation. For instance, recent studies have revisited model scaling with a U-Net benchmark for 3D medical image segmentation[33], demonstrating the impact of architectural scaling on performance. Additionally, comparative studies between 3D U-Net and 2D U-Net performances for cardiac segmentation[34] have highlighted the benefits and trade-offs of leveraging spatial context across slices. However, despite these advancements, these powerful models are not a panacea. They demand significant computational resources, require large, meticulously annotated datasets for training, i.e., a major bottleneck in medicine, and their inherent "black box" nature can impede clinical trust, prompting investigations into human-centric frameworks[35–37].

In response to the limitations of monolithic models, a trend toward hierarchical or multi-stage pipelines has emerged. These advanced strategies, which often integrate DL with classical methods or employ coarse-to-fine refinement, have demonstrated improved performance[18,38]. For instance, Wang et al.[18] utilize a two-stage process that first localizes the heart before segmenting its substructures, while Patil et al.[38] focus on creating domain-invariant frameworks to handle data from different scanner vendors—a critical real-world challenge. While effective, these hybrid approaches often introduce greater architectural complexity and computational cost. Our work builds on this hierarchical principle but distinguishes itself by dedicating separate, specialized models to each cardiac structure, a level of granularity that proves beneficial for anatomically complex regions like the RV. Specialized models tailored to specific cardiac regions have shown improved accuracy by focusing on unique anatomical features[39].

Despite these advances, a critical analysis of recent high-performing models reveals that specific challenges remain unsolved. For example, the complex framework proposed by da Silva et al.[21], which uses six separate CNNs, still showed diminished performance on ES slices compared to end-diastolic (ED) ones—a crucial phase for functional assessment. Similarly, while multimodal approaches that combine MRI and CT data are promising, they face daunting data harmonization and alignment issues[40,41]. Other leading models on the benchmark Automated Cardiac Diagnosis Challenge (ACDC) dataset, such as CondenseUNet[42] and Nested Capsule Dense Network[43], report impressive Dice scores but do not consistently solve the problem of accurately delineating the thin myocardial wall during systole or the geometrically complex RV. This body of work clearly demonstrates that while overall segmentation accuracy is high, a significant gap persists in developing a method that is robustly accurate across all cardiac structures and phases, especially in the most challenging anatomical contexts.

Recent work has also explored integrating segmentation and classification to streamline workflows[24,44], though this often increases training complexity. Corrective frameworks like the one proposed by Bourfiss et al.[45] attempt to fix segmentation errors but require manual intervention, defeating the purpose of full automation. While self-supervised learning methods promise to reduce the reliance on labeled data[46,47], their high computational demands currently limit their widespread clinical use[48,49]. In sum, the literature shows a clear, persistent need for a segmentation method that is robustly accurate across all cardiac structures (especially the RV) and all cardiac phases (especially ES), without incurring prohibitive computational or architectural complexity.

2.2 Classification of Cardiac MRIs

Following segmentation, the next critical task is classification, i.e., diagnosing specific pathologies. Traditional machine learning methods have been used, but they rely on handcrafted features and are consistently outperformed by end-to-end DL models that can automatically learn discriminative features from image data[50]. Architectures such as ResNet[51] and DenseNet[52] have become the cornerstones of modern medical image classification.

A dominant and effective strategy in the field is to leverage segmentation outputs to inform classification. By extracting clinically relevant parameters such as ventricular volumes and myocardial thickness from segmented masks, models can base their diagnostic decisions on quantitative, physiologically meaningful data. For example, Mofrad et al.[50] used this approach to classify five key cardiac conditions, i.e., dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), myocardial infarction (MINF), arrhythmogenic right ventricle (ARV), and normal (NOR), with high accuracy.

However, a review of state-of-the-art integrated models reveals a persistent and critical flaw: the confusion between pathologies with similar morphological characteristics. This is not a minor issue; it is a significant clinical limitation, as these conditions require different management strategies. The otherwise high-performing model by Isensee et al.[22], which achieved 92% testing accuracy, explicitly struggled to differentiate between DCM and MINF. Similarly, the work by Zheng et al.[24] also faced challenges in reliably separating these classes. The approach by Khened et al.[23], which reported an exceptional 100% classification accuracy, relied on a complex ensemble of multiple classifiers (support vector machine, multilayer perceptrons, etc.). While achieving high performance on a given test set, such complex ensembles are notoriously prone to overfitting, are difficult to interpret, and may not generalize well to unseen clinical data. Their complexity also presents a practical barrier to deployment. These examples provide clear evidence that a major unresolved problem in cardiac MRI classification is the lack of models that can robustly and reliably distinguish between similar disease states without resorting to overly complex, black-box ensembles. There is a pressing need for a classification architecture designed specifically to resolve this ambiguity.

2.3 Interpretation of DL Models

The most significant barrier to the widespread clinical adoption of AI in cardiology is the lack of trust, which stems directly from the "black box" nature of DL models. For a clinician to confidently use an AI-generated diagnosis, they must be able to understand and verify the evidence behind it. Consequently, the interpretability of these models is not just a desirable feature; it is an absolute necessity.

Early efforts at interpretation focused on generic visualization techniques like heatmaps or saliency maps, which highlight the image regions a model deems important[26]. However, these methods are fundamentally limited for clinical use. A heatmap

might show that a model is looking at the Myo, but it fails to explain what the model sees in clinical terms—is it wall thinning, hypertrophy, or abnormal motion? This disconnect between pixel-level importance and clinical reasoning is a major gap.

Recognizing this limitation, the field is undergoing a paradigm shift. As articulated in recent guidelines and expert reviews, there is a strong consensus that for AI to be trusted, its outputs must be presented in a clinically native format[25,53]. The work by Salih et al.[16] is particularly salient, as it advocates for moving beyond simple visualizations toward models that provide quantitative, clinically relevant outputs. This alignment with clinical practices is vital for AI's safe and effective deployment in diagnosing cardiac diseases, a principle that directly motivates the design of our interpretation module. Methods like deep Taylor-CAV represent a step in this direction by linking model decisions to higher-level visual concepts, such as chamber size or wall texture[54]. However, the ultimate goal is to translate a model's internal logic into the precise, standardized metrics that form the basis of a cardiologist's report. The analysis of the literature makes it abundantly clear that the development of such inherently interpretable, metric-driven AI systems is one of the most critical and unresolved challenges in medical AI today.

In summary, related works demonstrate significant advancements in cardiac MRI segmentation and classification through DL models, while also addressing the need for interpretability and trust in AI systems. These developments highlight the potential of AI to enhance diagnostic accuracy and efficiency in cardiology, provided that challenges related to model transparency and ethical considerations are effectively managed.

2.4 Aim of the Study and Research Objectives

The preceding analysis of the state of the art reveals critical, unresolved gaps at each stage of the cardiac MRI analysis pipeline. In segmentation, there is a persistent struggle to achieve high accuracy for anatomically challenging structures like the RV and the ES Myo. In classification, a key weakness is the failure to reliably differentiate between pathologies with similar phenotypes, such as DCM and MINF. Most critically, across the entire field, there is a profound disconnect between the outputs of "black box" DL models and the transparent, metric-based reasoning required for clinical practice. These are not isolated issues; they are interconnected challenges that demand a holistic solution.

Therefore, the objective of this work is to develop a comprehensive, hierarchical DL approach that achieves state-of-the-art accuracy in cardiac MRI segmentation and classification while ensuring clinical trust and utility by providing fully interpretable, metric-driven diagnostic insights.

To achieve this objective, the following tasks must be fulfilled:

1. Develop a multi-stage, structure-specialized segmentation method. This involves creating a hierarchical approach with separate models dedicated to first localizing and then accurately delineating the LV, RV, and Myo. This approach is designed to master the unique morphological features of each structure, thereby resolving the persistent accuracy issues seen in challenging anatomical areas.

2. Design and implement a hierarchical cascade classification architecture. To directly address the problem of class confusion, a cascade of specialized binary classifiers will be developed. This architecture will break down the complex five-way diagnostic task into a series of simpler, more robust decisions, systematically reducing ambiguity between similar pathologies.

3. Create an end-to-end interpretation method that translates model decisions into clinical practice. This involves building a module that automatically calculates and visualizes a dashboard of canonical clinical metrics (e.g., ventricular volumes, ejection fractions, wall thickness maps) directly from the model's outputs. This will bridge the critical gap between opaque AI predictions and transparent, verifiable clinical evidence.

3 Methods

3.1 Datasets, Splits, and Preprocessing

We used the publicly available, high-quality dataset titled ACDC[55] with the official train/test partition (100/50 patients). For external validation, we used the Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation (M&Ms) dataset[56]. All data splits for training, validation, and testing were performed at the patient level to prevent data leakage. For the classification task, per-patient image stacks (ED/ES) were augmented to create a uniform number of slices per patient; this augmentation was applied only within the training folds. No slice from a patient in the test or validation set ever appeared in the training set.

Region of interest (ROI) crops at test time are generated exclusively by the learned localizer models (M_1–M_3), ensuring no ground-truth information is used during inference. The Gaussian blur parameter, σ, used for post-processing, is not tuned per image. Instead, it is selected via a grid search on the validation set of the training data and then fixed for all test set inferences to prevent information leakage and ensure reproducibility. We additionally report an ablation study using ground-truth versus predicted ROIs to quantify the robustness of the localization stage.

It should be noted that this study focuses on detecting the following cardiac pathologies via MRI scans: DCM, HCM, MINF, and ARV. These particular pathologies were chosen due to their high prevalence and clinical significance. They are among the most common heart diseases[57,58], characterized by a range of morphological and functional changes that MRI can detect. ACDC includes annotated MRI scans for each condition. While the four pathologies cover major categories of cardiac disease—cardiomyopathies, inflammatory conditions, and arrhythmogenic states—they do not represent the full spectrum detectable by MRI. Ischemic diseases, restrictive cardiomyopathy, sarcoidosis, and congenital disabilities, for instance, are excluded due to their lower prevalence and a lack of sufficiently high-quality datasets for research.

In this study, the proposed approach to explainable AI-based detection of the aforementioned cardiac pathologies in MRI scans is decomposed into separate tasks, each solved by corresponding methods (Figure 1):

1. A multi-stage method for MRI scan segmentation.

2. A cascade classification method for pathologies.

3. A method for interpreting DL decisions.

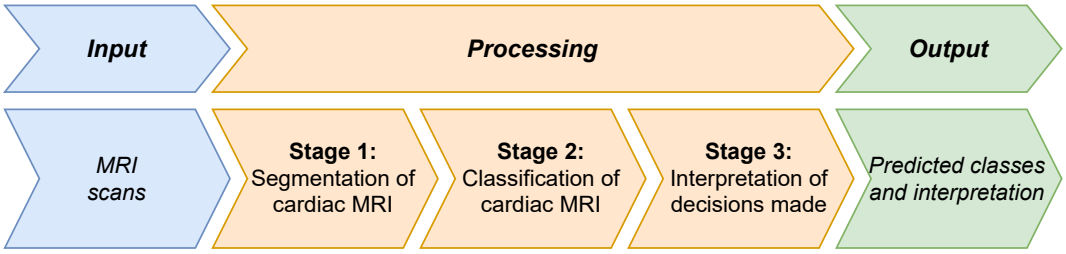


Figure 1. The proposed task decomposition for cardiac MRI processing, illustrating the process flow from input MRI scans through three sequential processing stages—segmentation, classification, and interpretation of decisions—resulting in the output of predicted classes and interpretations.

This decomposition is motivated by the following considerations.

1. Instead of analyzing the entire cardiac MRI scan, focus on the localized and segmented areas that are relevant for analysis (RV, LV, and Myo). This enables the DL model to concentrate on essential regions rather than extraneous information. Isolating this function as a separate task allows for selecting specialized DL models best suited for such preprocessing.

2. Medical datasets for classifier training often have limited sample sizes, making it challenging to achieve sufficient model generalization for multiple classes. Class confusion is common. Many approaches have been proposed to address this issue, including data augmentation. Using a cascade of separate binary classifiers for aggregated classes allows the model to increase the number of samples per aggregated class and focus on specific features of two aggregated classes simultaneously, leading to higher classification accuracy.

3. In the medical domain, the problem of interpretability of AI models is particularly critical. Addressing this issue is essential to overcome the "black box" effect, which undermines trust in clinical AI outcomes. This study proposes presenting DL-based results in a physician-friendly format through clinically familiar features.

The following sections provide a more detailed description of each method proposed to solve these decomposed tasks.

3.2 Method of Multi-Stage Cardiac MRI Segmentation

The concept for this method was first presented at the Informatics & Data-Driven Medicine (IDDM-2023) conference[59]. Our approach is built on a hierarchical, three-stage pipeline designed to systematically deconstruct the complexities of cardiac segmentation. This coarse-to-fine strategy first localizes the entire cardiac region, then performs structure-specific segmentation within that localized area, and finally applies a slice-aware refinement module to enhance boundary accuracy, particularly in challenging anatomical regions. This decomposition allows for the use of specialized DL models at each stage, maximizing precision and robustness.

3.2.1 Stage 1: Coarse Localization of Cardiac Structures

The initial and crucial step in our pipeline is to isolate the heart from the surrounding thoracic anatomy. By first identifying a bounding box that encompasses all relevant cardiac structures, we enable subsequent models to focus on a smaller, more relevant input area. This significantly reduces the impact of background noise and anatomical variability, which is especially beneficial for accurately processing edge slices, such as the apical and basal views, where the cardiac signal can be less distinct.

For this task, we trained three independent U-Net models with ResNet-50 encoders, denoted M_1, M_2, and M_3, each specialized in localizing one of the three primary structures: the LV, RV, and Myo, respectively. The choice of a U-Net architecture with a ResNet-50 backbone was motivated by its proven ability to capture both high-level semantic features (via the deep residual encoder) and fine-grained spatial details (via the skip connections to the decoder).

To train these localization models, we derived a specialized dataset, D_2, from the original ACDC dataset, D_1, defined in Equation 1.

$$D_1 = \{d_1, d_2, \ldots, d_N\}, \quad d_i = (\text{Img}_i, \text{Msk}_i), \quad i = 1, 2, \ldots, N, \tag{1}$$

The preparation of D_2 involved resizing all images and their corresponding multi-class masks to a uniform resolution of 256×256 pixels. Each multi-class mask was then decomposed into three separate binary masks, one for each cardiac structure. This resulted in the dataset defined in Equation 2:

$$D_2 = \{d_1, d_2, \ldots, d_N\}, \quad d_i = (\text{ctImg}_i, \text{ctMsk}_i^1, \text{ctMsk}_i^2, \text{ctMsk}_i^3), \tag{2}$$

Each model (M_1, M_2, M_3) was trained on this dataset to predict the binary mask for its target structure. During inference, the output of these models is not the mask itself but the tightest bounding box enclosing the predicted mask. The union of these three bounding boxes defines the final ROI for the next stage.

3.2.2 Stage 2: Fine-Grained, Structure-Specific Segmentation

With the cardiac ROI identified, the second stage performs the detailed delineation of each structure's contours. A key novelty of our approach is the use of separate, highly specialized models for each anatomical component. Instead of a single model tasked with differentiating three classes, we train three distinct models (M_4, M_5, M_6) dedicated to the binary segmentation of the LV, RV, and Myo, respectively. This strategy allows each model to learn the unique morphological and textural features of its target, for instance, the circular shape of the LV versus the crescent shape of the RV, without the confounding influence of other classes.

For these models, we employed a U-Net architecture with a lighter ResNet-34 encoder. This choice provides a favorable balance between representational power and computational efficiency, which is suitable for operating on smaller, cropped input images. The training data for these models consisted of three new datasets, D_3, D_4, and D_5, derived from the original dataset. As defined in Equations 3–5, each dataset contains images and masks cropped tightly around a single structure with a 15% margin and resized to a uniform 64×64 resolution.

$$D_3 = \{d_1, \ldots, d_N\}, \quad d_i = (\text{lvImg}_i, \text{lvMsk}_i), \tag{3}$$

$$D_4 = \{d_1, \ldots, d_N\}, \quad d_i = (\text{rvImg}_i, \text{rvMsk}_i), \tag{4}$$

$$D_5 = \{d_1, \ldots, d_N\}, \quad d_i = (\text{myoImg}_i, \text{myoMsk}_i). \tag{5}$$

While training separate models theoretically increases the computational efforts for training, the practical cost in our pipeline is minimal. Because our specialized models operate exclusively on tightly localized ROIs, they utilize a lightweight ResNet-34 backbone. Consequently, the entire pipeline executes in a fraction of a second during inference on standard personal computer hardware, creating no bottleneck in the clinical diagnostic workflow. Regarding the 64×64 resolution, this dimension was a deliberate, data-driven choice aimed at preserving fine anatomical details, particularly the thin myocardial wall. Our statistical analysis of the dataset revealed that a 64×64 crop achieves an optimal approximately 90/10 scaling ratio: roughly 90% of the localized cardiac ROIs are natively smaller than 64×64 and require upscaling (which safely preserves existing topological information via resizing), while only approximately 10% require downscaling, which is a lossy operation that may remove spatial details. Minimizing the downscaled fraction is crucial. Furthermore, 64 is a pure power of 2, perfectly aligning with the ResNet-34 architecture's five spatial downsampling stages ($64 \to 32 \to 16 \to 8 \to 4 \to 2$). Arbitrary dimensions force the network to apply uneven zero-padding during decoder skip connections, which introduces spatial artifacts and degrades boundary precision.

This focused training regimen enables the models to achieve high delineation accuracy, especially for the challenging RV and the thin myocardial wall during the ES phase. The complete training procedure for all six models is summarized in Figure 2 and detailed in Algorithm 1.

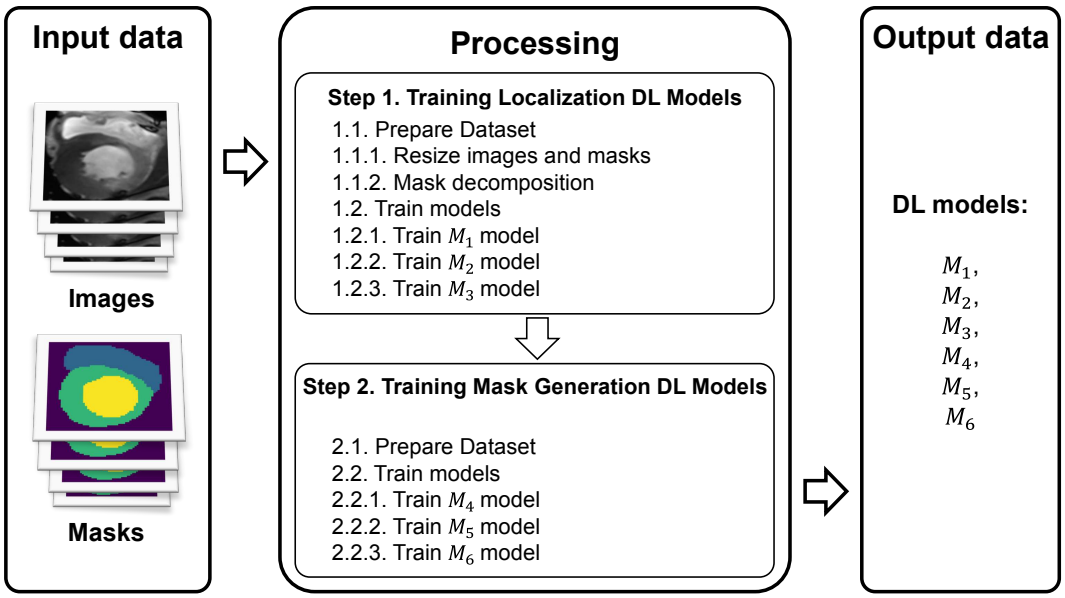


Figure 2. The main steps for obtaining DL models for segmenting the LV, RV, and Myo regions in an MRI scan.

3.2.3 Stage 3: Post-Processing and Slice-Aware Refinement

The final stage of our pipeline refines the raw segmentation masks to improve their anatomical plausibility and ensure volumetric consistency. This stage consists of two key components: an initial post-processing step and a dedicated apical/basal refinement module.

First, the predicted binary masks from M_4, M_5, and M_6 are reintegrated into a single multi-class mask and resized to the original image dimensions. To mitigate minor artifacts and jagged edges that can arise from this process, we apply a Gaussian smoothing filter, as suggested in our prior work[59]. The smoothing parameter, σ, is determined via a grid search on a validation split of the training data and remains fixed for all test set inferences to prevent data leakage.

While this procedure yields high-quality segmentations for most of the cardiac volume, we recognized that the apical and basal slices present unique challenges. Boundary delineation in these regions is notoriously difficult due to partial volume effects and ambiguous anatomical landmarks. To address this, we introduce a slice-position-aware refinement module that is applied after the initial segmentation. This module leverages both geometric priors and inter-slice relationships to correct errors common at the ventricular extremes. It incorporates two main components:

- Shape-Prior Regularization: This component penalizes anatomically implausible contours. We pre-compute signed distance functions (SDFs) from the ground-truth masks in the training set to create a statistical shape model. During refinement, a loss term is introduced that measures the discrepancy between the SDF of the predicted segmentation and the learned shape prior, guiding the model to produce more realistic outputs.

- 3D Continuity Penalty: To ensure smooth and logical transitions between adjacent slices, this component imposes a continuity constraint. It penalizes large, abrupt changes in the segmented area or shape between one slice and the next, thereby enforcing volumetric consistency and reducing the occurrence of outlier segmentations in isolated slices.

This refinement module specifically targets the under- and over-segmentation errors common at the apical and basal ends of the ventricles. The impact of this module is quantified by reporting Dice scores stratified by slice location (basal, mid-ventricular, and apical) in our results section.

The complete inference workflow, from raw MRI input to the final refined segmentation mask, is illustrated in Figure 3 and formalized in Algorithm 2.

3.3 Method of Cardiac MRIs Cascade Classification

The core concept of our classification method was first presented at the Information Technology and Implementation (IT&I-2024) conference[60]. Building on this foundation, our approach is designed to accurately diagnose five cardiac conditions, i.e., DCM, HCM, MINF, ARV, and NOR, by leveraging a novel spatiotemporal data representation and a hierarchical cascade of specialized classifiers.

Algorithm 1 Training Procedure for Segmentation Models

Require: Training split of the dataset $D_{\text{train}} = \{(\text{Img}_i, \text{Msk}_i)\}_{i=1}^{N_{\text{train}}} \subset D_1$ ▷ Described in Eq. 1
Ensure: Trained localization models M_1, M_2, M_3 and mask generation models M_4, M_5, M_6

function PREPARELOCALIZATIONDATA(D_{train})
  $D_2 \leftarrow \emptyset$ ▷ Dataset for localization models, per Eq. 2
  for all $(\text{Img}_i, \text{Msk}_i) \in D_{\text{train}}$ **do**
    $\text{Img}_i' \leftarrow \text{RESIZE}(\text{Img}_i, (256, 256))$
    $\text{Msk}_i' \leftarrow \text{RESIZE}(\text{Msk}_i, (256, 256))$
    $(\text{Msk}_i^{\text{LV}}, \text{Msk}_i^{\text{RV}}, \text{Msk}_i^{\text{Myo}}) \leftarrow \text{DECOMPOSEMASK}(\text{Msk}_i')$
    $D_2 \leftarrow D_2 \cup \{(\text{Img}_i', \text{Msk}_i^{\text{LV}}, \text{Msk}_i^{\text{RV}}, \text{Msk}_i^{\text{Myo}})\}$
  end for
  return D_2
end function

function PREPARESEGMENTATIONDATA(D_{train})
  $D_3, D_4, D_5 \leftarrow \emptyset, \emptyset, \emptyset$ ▷ Datasets for fine-grained segmentation, per Eqs. 3–5
  for all $(\text{Img}_i, \text{Msk}_i) \in D_{\text{train}}$ **do**
    $(\text{lvImg}_i, \text{lvMsk}_i) \leftarrow \text{CROPANDRESIZE}(\text{Img}_i, \text{Msk}_i, \text{target='LV'})$
    $D_3 \leftarrow D_3 \cup \{(\text{lvImg}_i, \text{lvMsk}_i)\}$
    $(\text{rvImg}_i, \text{rvMsk}_i) \leftarrow \text{CROPANDRESIZE}(\text{Img}_i, \text{Msk}_i, \text{target='RV'})$
    $D_4 \leftarrow D_4 \cup \{(\text{rvImg}_i, \text{rvMsk}_i)\}$
    $(\text{myoImg}_i, \text{myoMsk}_i) \leftarrow \text{CROPANDRESIZE}(\text{Img}_i, \text{Msk}_i, \text{target='Myo'})$
    $D_5 \leftarrow D_5 \cup \{(\text{myoImg}_i, \text{myoMsk}_i)\}$
  end for
  return D_3, D_4, D_5
end function

Main Execution:
$D_2 \leftarrow \text{PREPARELOCALIZATIONDATA}(D_{\text{train}})$
$M_1 \leftarrow \text{TRAINMODEL}(\text{U-Net-ResNet50 on } D_2, \text{target='LV'})$
$M_2 \leftarrow \text{TRAINMODEL}(\text{U-Net-ResNet50 on } D_2, \text{target='RV'})$
$M_3 \leftarrow \text{TRAINMODEL}(\text{U-Net-ResNet50 on } D_2, \text{target='Myo'})$

$D_3, D_4, D_5 \leftarrow \text{PREPARESEGMENTATIONDATA}(D_{\text{train}})$
$M_4 \leftarrow \text{TRAINMODEL}(\text{U-Net-ResNet34 on } D_3, \text{target='LV'})$
$M_5 \leftarrow \text{TRAINMODEL}(\text{U-Net-ResNet34 on } D_4, \text{target='RV'})$
$M_6 \leftarrow \text{TRAINMODEL}(\text{U-Net-ResNet34 on } D_5, \text{target='Myo'})$

3.3.1 Spatiotemporal Input Representation

Standard classification models often operate on single images, neglecting the rich dynamic information contained within a full cardiac cycle. To overcome this limitation, we developed a unique input representation that captures both anatomical structure and temporal function. This is achieved by encoding the segmented cardiac structures from both the ED and ES phases into a multi-channel format.

Specifically, for each patient, the segmentation masks for the LV, RV, and Myo are used to isolate the corresponding regions in the original MRI scans. These segmented structures are then encoded into the three channels Red, Green, and Blue (RGB) of an RGB-like image. The red channel contains the LV, the green channel holds the Myo, and the blue channel represents the RV, as illustrated in Figure 4. This technique preserves not only the geometry of the cardiac chambers but also the underlying myocardial tissue texture, which can contain subtle pathological indicators.

To create a comprehensive spatiotemporal volume, we stack the multi-channel images from all available short-axis slices at both ED and ES. This results in the data transformation defined in Equation 6 and Equation 7, where each patient's data becomes a 42-slice volume (21 for ES, 21 for ED).

$$F : I_{MRT} \rightarrow I_{MRT}^{\text{new}}, \tag{6}$$

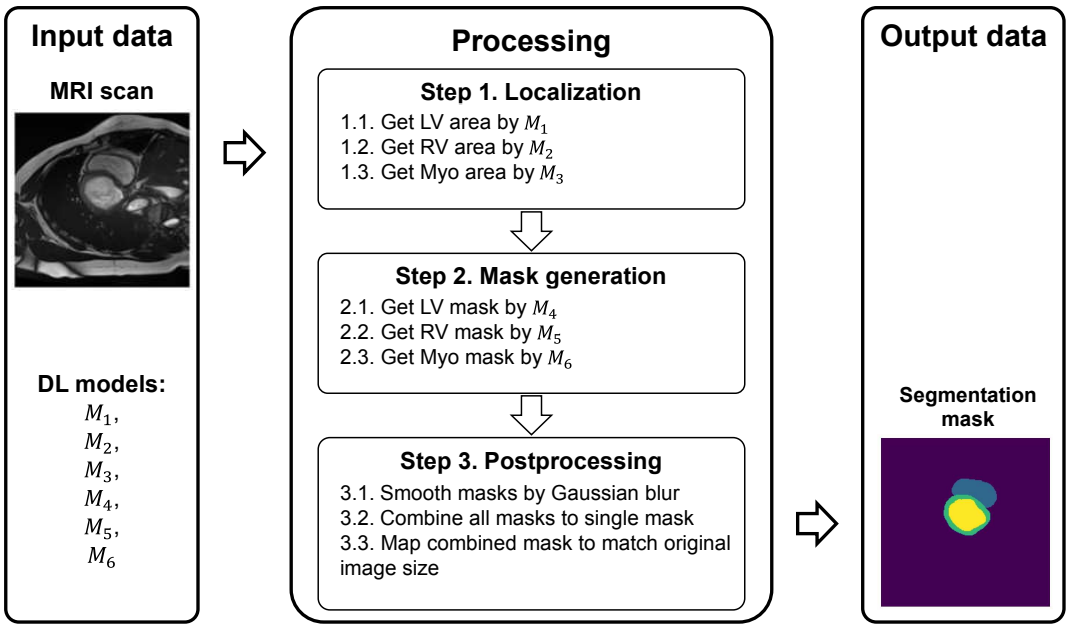


Figure 3. The main steps for segmenting an MRI scan into regions of the LV, RV, and Myo, including localization, segmentation, and post-processing.

Algorithm 2 Inference Pipeline for Segmentation

Require: Arbitrary MRI scan Img, Trained models $M_1, \ldots, M_6$
Ensure: Final refined mask FinalRefinedMsk

% Stage 1: Localize ROI bounding boxes
OriginalSize $\leftarrow$ GETSIZE(Img)
$\text{Img}' \leftarrow$ RESIZE(Img, (256, 256))
$\text{bbox}_{\text{LV}} \leftarrow M_1(\text{Img}')$
$\text{bbox}_{\text{RV}} \leftarrow M_2(\text{Img}')$
$\text{bbox}_{\text{Myo}} \leftarrow M_3(\text{Img}')$
$\text{CombinedBbox} \leftarrow \text{UNION}(\text{bbox}_{\text{LV}}, \text{bbox}_{\text{RV}}, \text{bbox}_{\text{Myo}})$

% Stage 2: Generate masks for localized regions
$\text{Crop}_{\text{ROI}} \leftarrow \text{CROPIMAGE}(\text{Img}, \text{CombinedBbox})$
$\text{PredMsk}_{\text{LV}} \leftarrow M_4(\text{Crop}_{\text{ROI}})$ ▷ LV binary segmentation
$\text{PredMsk}_{\text{RV}} \leftarrow M_5(\text{Crop}_{\text{ROI}})$ ▷ RV binary segmentation
$\text{PredMsk}_{\text{Myo}} \leftarrow M_6(\text{Crop}_{\text{ROI}})$ ▷ Myo binary segmentation

% Stage 3: Post-process and refine masks
$\text{InitialMsk} \leftarrow \text{COMBINEANDRESIZE}(\text{PredMsk}_{\text{LV}}, \text{PredMsk}_{\text{RV}}, \text{PredMsk}_{\text{Myo}}, \text{OriginalSize})$
$\text{SmoothMsk} \leftarrow \text{GAUSSIANBLUR}(\text{InitialMsk}, \sigma)$ ▷ σ predetermined on validation set
FinalRefinedMsk $\leftarrow$ APPLYREFINEMENTMODULE(SmoothMsk) ▷ Shape-prior and 3D continuity corrections
return FinalRefinedMsk

$$I_{MRT}^{\text{new}} = (\text{img}_1, \ldots, \text{img}_K), \quad K = 42, \tag{7}$$

Representation 7 forms the basis of our training dataset, D_6, as defined in Equation 8. To handle cases with fewer than 21 slices, we apply augmentation by duplicating adjacent slices, a process restricted to the training folds to prevent data leakage.

$$D_6 = \{d_1, \ldots, d_N\}, \quad d_i = (I_{MRT}^{\text{new}}, \text{Cls}). \tag{8}$$

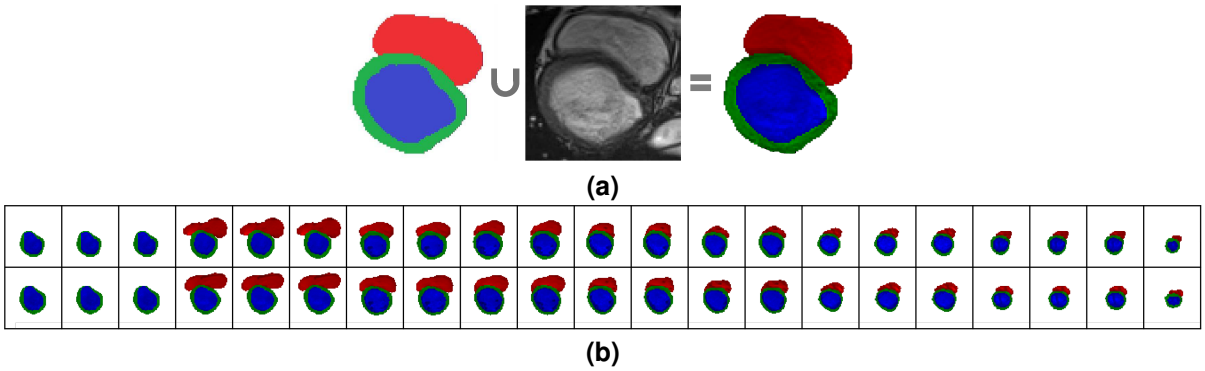


Figure 4. Construction of the multi-channel input representation. (**a**) An original MRI scan is overlaid with color-coded segmentation masks for the LV (depicted as blue), RV (depicted as red), and Myo (depicted as green). (**b**) The full input for the classifier consists of a stack of these multi-channel images, representing short-axis slices from both the systolic (top row) and diastolic (bottom row) phases of the cardiac cycle.

3.3.2 Hierarchical Cascade of Classifiers

Attempting to classify five distinct cardiac conditions with a single multiclass model is challenging, especially with limited datasets where class confusion is common (e.g., distinguishing MINF from DCM). To mitigate this, we designed a hierarchical cascade of four specialized binary classifiers (M_7 through M_{10}). This architecture breaks down the complex five-way classification into a series of simpler, more robust decisions, as illustrated in Figure 5.

The decision-making process of the cascade is as follows:

- Classifier M_7: This initial model performs the broadest separation, distinguishing between pathologies primarily affecting the LV (HCM, MINF, and DCM) and those that do not (ARV and NOR).
- Classifier M_8: If M_7 predicts a non-LV pathology, this model is invoked to differentiate between ARV and NOR.
- Classifier M_9: If M_7 identifies an LV pathology, this model separates HCM from the other two conditions (MINF and DCM), which often present with ventricular dilation.
- Classifier M_{10}: As the final step, if a patient is classified as having either MINF or DCM, this model makes the critical distinction between these two often-confused pathologies.

This hierarchical approach allows each classifier to focus on a more specific and manageable task, leading to improved overall accuracy and better handling of class imbalances. Each of the four models is a custom-designed CNN adapted for binary classification, trained using the Adam optimizer and a categorical cross-entropy loss function. To prevent overfitting, we employ an early stopping mechanism and ensure that class proportions are maintained in both the training and validation splits. The complete training workflow is depicted in Figure 6.

3.3.3 Inference Pipeline for Classification

The classification of an arbitrary cardiac MRI scan follows a systematic two-part process: data preparation followed by the application of the trained cascade of classifiers. This entire inference workflow is illustrated in Figure 7 and detailed in Algorithm 3.

First, the input MRI scan undergoes the same preprocessing as the training data. This involves segmenting all slices using our multi-stage segmentation method, cropping the images to the ROI, and constructing the 42-slice multi-channel spatiotemporal volume. Once the input is prepared, it is passed through the cascade of classifiers. The output of each model in the sequence determines which model is invoked next, culminating in a final diagnosis of one of the five cardiac conditions.

3.4 Method of Interpreting Obtained Decisions based on Cardiac MRI

3.4.1 Features Indicating Heart Diseases

To elucidate the decisions made by DL models, it is essential to transform DL architecture outputs into clinically relevant features used by physicians for diagnosing heart diseases. The following key feature groups are formulated based on the heart diseases addressed in this study.

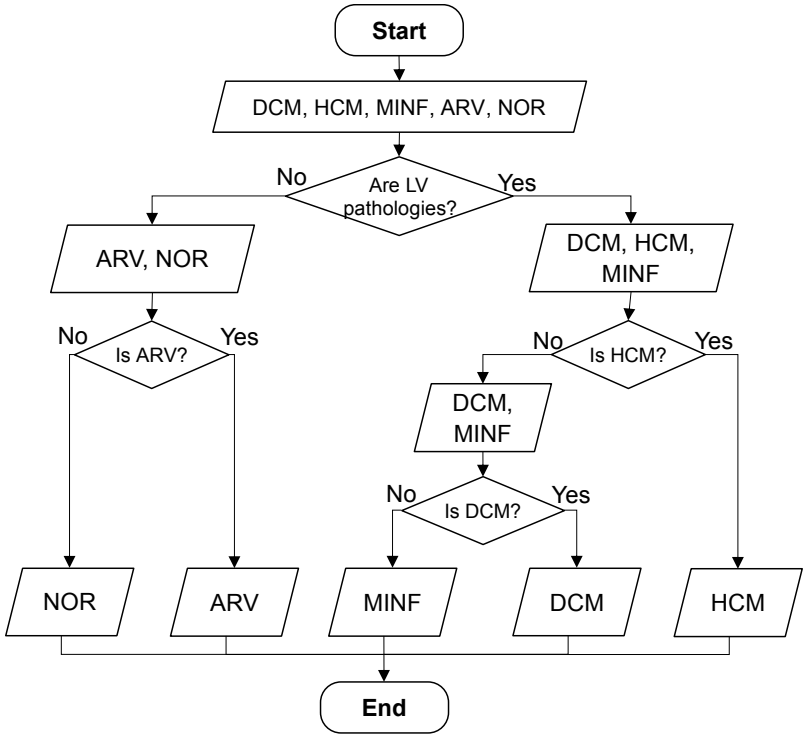


Figure 5. The structure of the hierarchical cascade classification model. The process begins by separating LV-related pathologies from others and progressively refines the diagnosis through a series of specialized binary classifiers.

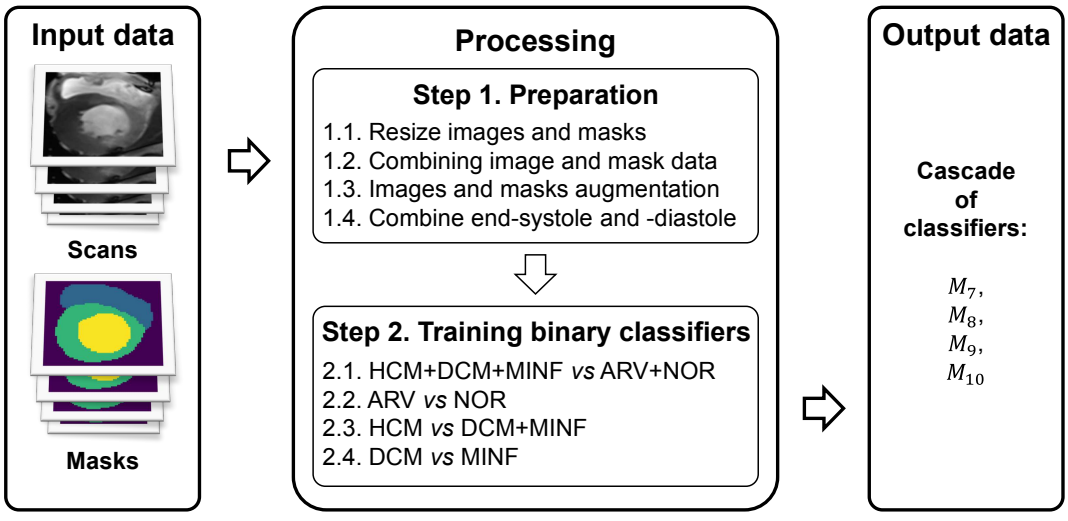


Figure 6. The overall scheme of the training process for the four cascade classifiers (M_7 through M_{10}).

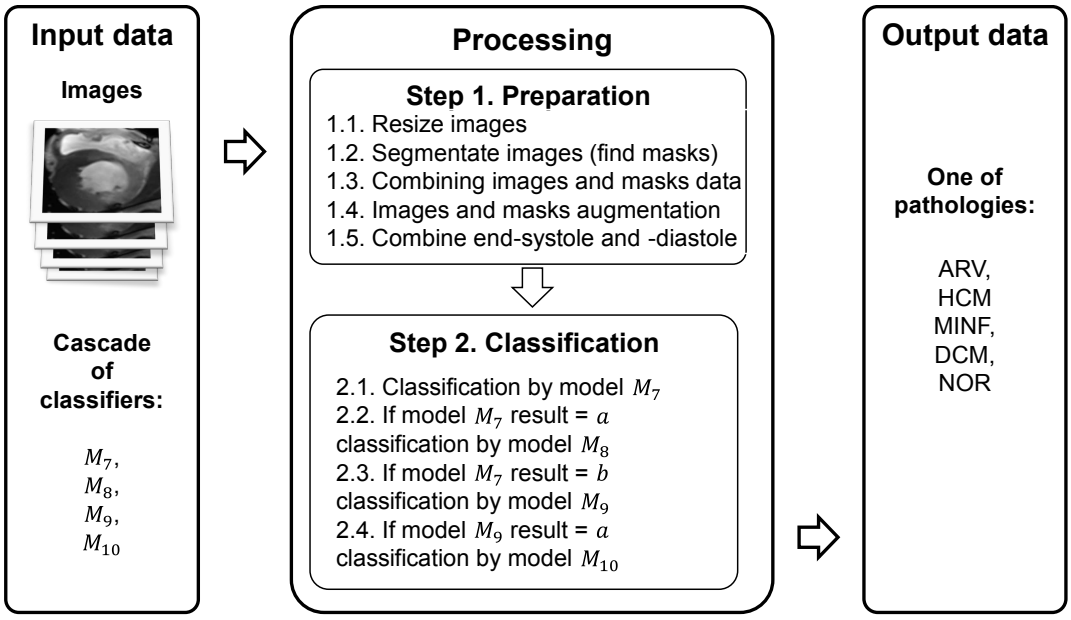


Figure 7. The inference pipeline for classifying a cardiac MRI from an arbitrary patient using the trained cascade model.

Algorithm 3 Inference Pipeline for Cascade Classification

Require: A set of cardiac MRI scans for a single patient (ED and ES phases); Trained classifiers M_7, M_8, M_9, M_{10}
Ensure: A predicted pathology class $\mathscr{C} \in \{\text{ARV}, \text{HCM}, \text{MINF}, \text{DCM}, \text{NOR}\}$

% Step 1: Prepare the input data
Segment all MRI slices utilizing the multi-stage segmentation method ▷ See Algorithm 2
Crop the segmented images strictly to the inferred ROI.
Construct the 42-slice multi-channel spatiotemporal volume, I_{MRT}^{new} ▷ Refer to Eqs. 6–7

% Step 2: Apply the cascade of classifiers
$P_{M7} \leftarrow M_7(I_{MRT}^{\text{new}})$ ▷ Differentiate LV-related pathologies from others
if $P_{M7} \in \{\text{ARV}, \text{NOR}\}$ **then**
    $P_{M8} \leftarrow M_8(I_{MRT}^{\text{new}})$ ▷ Differentiate ARV and NOR
    if $P_{M8} = \text{ARV}$ **then**
        return ARV
    else
        return NOR
    end if
else
    $P_{M9} \leftarrow M_9(I_{MRT}^{\text{new}})$ ▷ Isolate HCM from MINF/DCM group
    if $P_{M9} = \text{HCM}$ **then**
        return HCM
    else
        $P_{M10} \leftarrow M_{10}(I_{MRT}^{\text{new}})$ ▷ Differentiate between MINF and DCM
        if $P_{M10} = \text{DCM}$ **then**
            return DCM
        else
            return MINF
        end if
    end if
end if

Group 1: Ventricular volumes. Heart volume indicators assess the function of the LV and RV. The ratio of LV volume to RV volume at the ED evaluates ventricular balance. LV volumes at ES and ED indicate contractile ability and blood capacity, respectively. Similarly, RV volumes help identify dysfunctions; for instance, increased RV volume at the ED may signify ARV, while decreased LV volume at the ES may suggest HCM as investigated in[61,62].

Group 2: Ejection fraction. Ejection fraction measures the efficiency of blood ejection by the ventricles. LV ejection fraction reflects the proportion of blood pumped out during each contraction, which is crucial for diagnosing DCM, where it is typically reduced. RV ejection fraction indicates the pumping efficiency of the RV, which is crucial for diagnosing ARV[62].

Group 3: Volume-to-mass ratios. This group assesses the ratio of heart volumes to myocardial mass, aiding in detecting pathological changes. An increased myocardial mass relative to LV volume may indicate hypertrophic changes, while a decreased ratio suggests dilative processes as stated in[24,62].

Group 4: Thickness and variability of the Myo walls. Myocardial wall thickness and its variability are vital for classifying heart diseases. Increased wall thickness during diastole may indicate HCM, whereas decreased thickness suggests dilative processes. Variability in wall thickness during the cardiac cycle assesses the uniformity of myocardial contractions, with high variability pointing to uneven contractions characteristic of specific cardiomyopathies like HCM, following[61].

3.4.2 Methodological Policy on Papillary Muscles and Trabeculations

For the calculation of all volumetric and mass-based clinical indices derived from the segmentation masks, a consistent anatomical policy was adopted. Papillary muscles and trabeculations were methodologically excluded from the myocardial mass and were considered part of the LV blood pool (see Table S1 in the Supplementary Information). This approach was chosen to ensure consistency and comparability with prevalent clinical reporting standards, which often favor this convention to reduce inter-observer variability in mass and volume quantification. The impact of this methodological choice is further quantified and discussed in the Discussion section.

3.4.3 Interpretation and Visualization Model

Below is the information model for interpreting decisions made by DL models using transformations F_d, as defined in Equation 9:

$$F_d : I \rightarrow O_d, \tag{9}$$

where d indicates the pathology (1 – DCM, 2 – HCM, 3 – MINF, 4 – ARV, 5 – NOR), I is the segmented MRI scan, and O_d is a set of features for pathology d defined as shown in Equation 10:

$$O_d = \{o_1^d, \ldots, o_{N_d}^d\}, \quad o_i^d = (\text{name}, \text{type}_j), \tag{10}$$

with 'name' as the feature name and 'type_j' as the presentation mode: type_1 – 17-segment Myo model, type_2 – circular chart, and type_3 – numeric representation.

Below are the measures for each pathology from the set per Equation 10, along with recommended visualization methods.

- d = 1 – DCM: o_1^1 = (LV ED volume, type_3); o_2^1 = (LV ES volume, type_3); o_3^1 = (LV ejection fraction, type_3); o_4^1 = (myocardial mass at ED, type_3); o_5^1 = (ratio of myocardial mass to LV ED volume, type_2); o_6^1 = (maximum average myocardial wall thickness in ED, type_1).
- d = 2 – HCM: o_1^2 = (LV ES volume, type_3); o_2^2 = (LV ejection fraction, type_3); o_3^2 = (ratio of myocardial mass to LV ES volume, type_2); o_4^2 = (maximum average myocardial wall thickness in ED, type_1); o_5^2 = (mean standard deviation of wall thickness in ES, type_3); o_6^2 = (mean standard deviation of wall thickness in ED, type_3).
- d = 3 – MINF: o_1^3 = (LV ED volume, type_3); o_2^3 = (LV ejection fraction, type_3); o_3^3 = (myocardial mass at ED, type_3); o_4^3 = (maximum average myocardial wall thickness in ED, type_1); o_5^3 = (mean standard deviation of wall thickness in ED, type_3).
- d = 4 – ARV: o_1^4 = (RV ED volume, type_3); o_2^4 = (RV ejection fraction, type_3); o_3^4 = (ratio of LV to RV ED volumes, type_2); o_4^4 = (maximum average myocardial wall thickness in ES, type_1); o_5^4 = (mean standard deviation of wall thickness in ES, type_3).

- d = 5 – NOR: o_1^5 = (LV ES volume, $type_3$); o_2^5 = (LV ED volume, $type_3$); o_3^5 = (ratio of LV to RV ED volumes, $type_2$); o_4^5 = (LV ejection fraction, $type_3$); o_5^5 = (maximum average myocardial wall thickness in ES, $type_1$).

Hence, volumes and ejection fraction are displayed numerically, ratios with circular charts, and wall thickness/variability using the 17-segment model. This layout helps clinicians grasp key cardiac metrics from a single image without comparing multiple MRI slices.

3.4.4 Main Steps of the Proposed Method

This section outlines the main steps of the method for interpreting results obtained through DL means. The primary goal of the method is to present key features in a format suitable for visualization and analysis, thereby ensuring transparency in the pathology classification process. The overall scheme of the result interpretation method is depicted in Figure 8. The steps, input, and output information are described below.

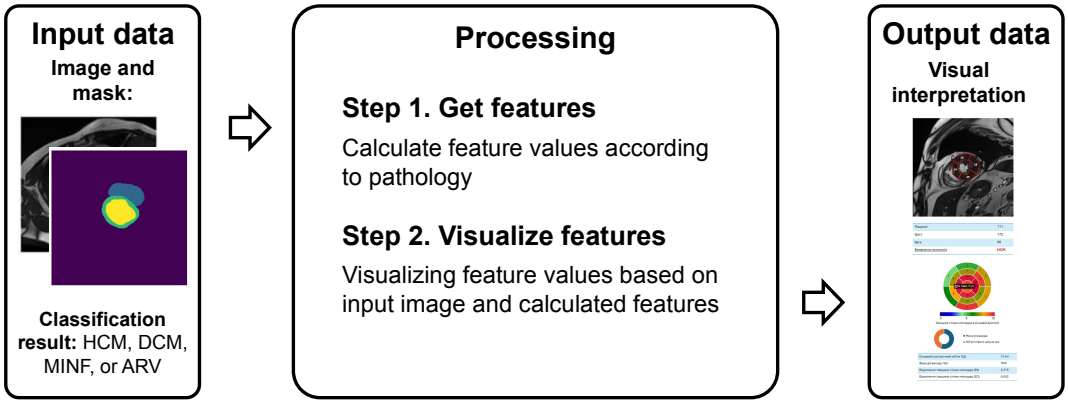


Figure 8. The overall scheme of the proposed interpretation method.

The input data for the method consists of a segmented MRI scan (using the multi-stage segmentation method) and its classification result.

Step 1: Obtaining feature values. The values of features for pathology d are determined from the input image. This provides a numerical basis for further analysis.

Step 2: Presentation of obtained features. The features are presented to the end user separately by types for the corresponding pathologies.

The output of the method is the input segmented and classified MRI scan along with the visualized type features.

Thus, the main steps of the decision interpretation method enhance the accessibility and comprehensibility of the obtained data for further analysis. Users, including physicians and researchers, can assess the classification results by calculating key features and visualizing them. This increases trust in the DL models and facilitates their practical use in diagnosing patients with heart pathologies.

3.5 Evaluation Approaches and Metrics

In this study, we utilized several essential metrics to assess the performance of our DL models in medical applications, covering both segmentation and multiclass classification tasks. Our findings are consistent with existing research on model evaluation in medical image analysis, as highlighted in Rainio et al.[63].

3.5.1 Experimental Models for Validating the Multi-Stage Segmentation Method

To objectively measure how each stage of the multi-stage segmentation method affects final segmentation quality, a series of experiments was performed with identical epochs, architectures, and data. Results were averaged over 10 training/testing cycles, and segmentation accuracy was assessed using the Dice coefficient, comparing predicted and expert masks.

To evaluate the combined impact of these steps, an arithmetic mean of Dice across all structures (LV, RV, and Myo) in both ES and ED phases was calculated. Six additional models were developed to test specific components:

- Base model $M_1^{\exp 1}$: Trained without any optional preprocessing or post-processing (only uniform resizing). It serves as a baseline for recognizing heart structures on the original ACDC dataset.

- Localization models $M_1^{\exp 2}$ and $M_2^{\exp 2}$: $M_1^{\exp 2}$ localizes the heart region as a single binary mask, while $M_2^{\exp 2}$ segments LV, RV, and Myo within that localized region. Both share the same input ACDC dataset.
- Decomposition models $M_1^{\exp 3}$, $M_2^{\exp 3}$ and $M_3^{\exp 3}$: Each model handles binary segmentation of one structure (LV, RV, or Myo) using decomposed masks. The same ACDC dataset is resized and split accordingly.

These experiments compare partial versus combined stages, verifying each stage's effectiveness and its contributions to overall segmentation accuracy.

3.5.2 Metrics for Evaluating the Cascade Classification Method Results

To measure classification performance, each classifier's accuracy is computed at every step, then averaged. Equations for calculating class-specific and overall accuracies, namely for NOR, ARV, HCM, MINF, and DCM, are formalized as follows:

$$A_{\text{NOR,ARV}} = \frac{A_{\text{Classifier1}} + A_{\text{Classifier2}}}{2}, \tag{11}$$

$$A_{\text{HCM}} = \frac{A_{\text{Classifier1}} + A_{\text{Classifier3}}}{2}, \tag{12}$$

$$A_{\text{MINF,DCM}} = \frac{A_{\text{Classifier1}} + A_{\text{Classifier3}} + A_{\text{Classifier4}}}{3}, \tag{13}$$

$$A_{\text{gen}} = \frac{A_{\text{NOR}} + A_{\text{ARV}} + A_{\text{HCM}} + A_{\text{MINF}} + A_{\text{DCM}}}{5}, \tag{14}$$

where $A_{\text{Classifier1}-4}$ are the accuracies of each classifier, $A_{\text{NOR,ARV,HCM,MINF,DCM}}$ are the classification accuracies for each individual class, and A_{gen} is the overall accuracy.

Evaluation using Equations 11–14 ensures a fair comparison with existing methods.

Furthermore, standard classification metrics include Accuracy, Precision, Recall, F_1-score, and Area Under the Curve (AUC) with Receiver Operating Characteristic (ROC) curves. They are based on counts of true positives, false positives, false negatives, and true negatives. Taken together, these metrics offer a detailed perspective on how effectively the method distinguishes different cardiac pathologies.

3.5.3 Evaluation, Uncertainty, and Calibration

Segmentation metrics include the Dice coefficient per structure and cardiac phase, reported with 95% bootstrap confidence intervals (CIs) (2,000 resamples) to quantify uncertainty. Model calibration is assessed using reliability diagrams, accompanied by the Expected Calibration Error (ECE) and Maximum Calibration Error (MCE)[64]. Operating thresholds for the binary classifiers are determined on the validation folds using Youden's J-index[65], and we report the corresponding sensitivity and specificity at these thresholds.

3.5.4 Evaluation of the Interpretation Method Results

It should be noted that the quality and accuracy of the results obtained through the proposed approach depend on the metrics of the multi-stage segmentation and cascade classification methods. The interpretation method is designed solely to interpret the decisions made by the previous methods. The method provides feature values derived from MRI scans, which allow the physician to understand why the DL models made specific decisions. It is important to note that this decision is advisory for a physician, who makes the final diagnosis based on the numerical features provided.

The feature values used by the physician to make decisions are obtained in an understandable manner (typically geometric characteristics of different image segments) and do not require additional evaluation using specialized metrics.

3.6 Dataset

This study utilizes the ACDC dataset[55]. It comprises 150 patients and is divided into five groups: 30 healthy patients, 30 patients with a history of MINF, 30 patients with DCM, 30 patients with HCM, and 30 patients with ARV. For each patient, the data includes physical parameters, a set of images, and expert-annotated masks (ground truth labels) at different stages of the cardiac cycle. The expert masks contain segmented regions for LV, RV, and Myo. An example element of the D_1 dataset is illustrated in Figure 9.

For external validation, we utilized the M&Ms challenge dataset[56]. This dataset contains 375 CMR studies from six hospitals, acquired using four different scanner vendors (Siemens, Philips, General Electric, Canon). While the ACDC is relatively homogeneous, having been acquired using a standardized protocol on Siemens scanners, the M&Ms dataset introduces significant clinical heterogeneity, comprising data from multiple centers and four distinct scanner vendors. This domain variability is highly relevant to real-world clinical practice. We intentionally evaluate our models on the M&Ms dataset without any domain adaptation to rigorously test zero-shot generalizability across these heterogeneous environments; a detailed analysis of how specific vendor-induced domain shifts impact model performance is provided in Section 4.3. The performance of our models, trained exclusively on ACDC data, was evaluated on the M&Ms test set to assess their generalizability under significant domain shift.

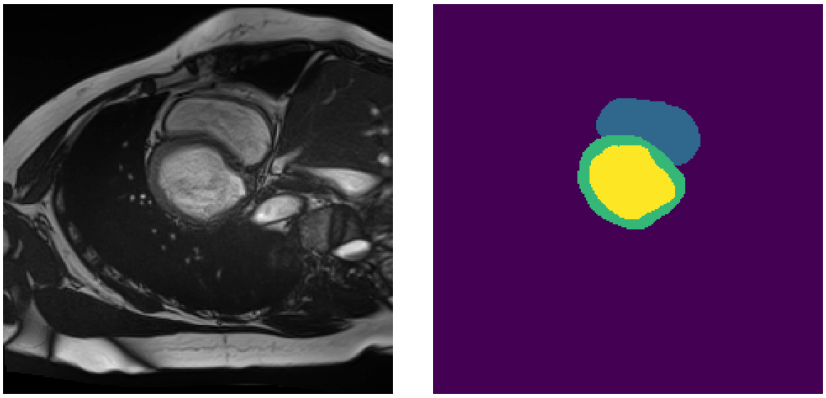


Figure 9. Example from the dataset D_1 with a cardiac MRI scan **(a)** and its mask **(b)**.

For training and testing the models (M_1–M_{10}), corresponding dataset modifications were created. Below, we examine examples of dataset elements and their characteristics.

For training the localization models (M_1–M_3), dataset D_2 was created, containing 2,978 MRI scans. All images were cropped to retain only the necessary segments (identified using the first method) and resized to a uniform resolution (256 × 256 pixels) with an aspect ratio of 1:1. For each image, separate masks were generated for each of the studied heart structures (LV, RV, and Myo). An example element of dataset D_2 is shown in Figure 10.

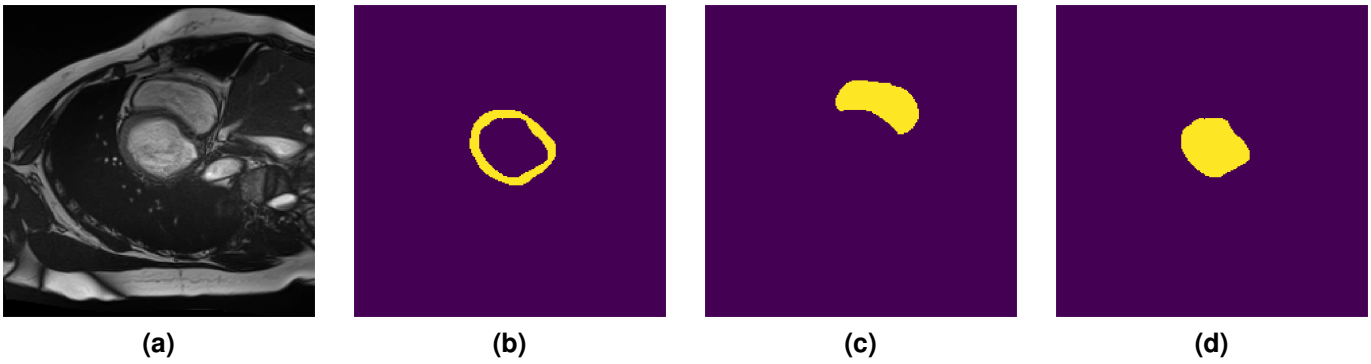


Figure 10. Instance element of dataset D_2: **(a)** MRI scan, **(b)** Myo segment mask, **(c)** LV segment mask, and **(d)** RV segment mask.

For training the mask generation models (M_4–M_6), corresponding datasets D_3, D_4, and D_5 were created. These datasets, based on D_2, contain the same number of images. All images in these datasets have a uniform resolution (64 × 64) and an aspect ratio of 1:1. Each dataset corresponds to a specific heart structure:

- Dataset D_3: contains localized images of the LV and the corresponding localized mask. An example element of dataset D_3 is shown in Figure 11a-b.

- Dataset D_4: contains localized images of the RV and the corresponding localized mask. An example element of dataset D_4 is shown in Figure 11c-d.
- Dataset D_5: contains localized images of the Myo and the corresponding localized mask. An example element of dataset D_5 is shown in Figure 11e-f.

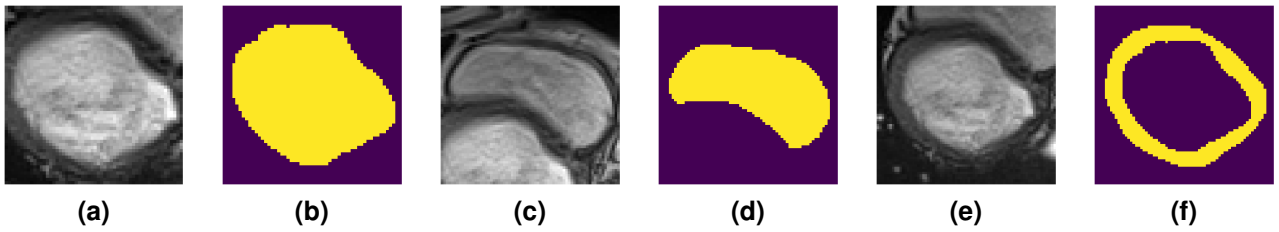


Figure 11. Example elements of created datasets: D_3 – **(a)** MRI scan, **(b)** LV segment mask; D_4 – **(c)** MRI scan, **(d)** RV segment mask; D_5 – **(e)** MRI scan, **(f)** Myo segment mask.

For training the cascade classification models (M_7–M_{10}), dataset D_6 was formed. This dataset contains modified MRI scans for each patient and their corresponding pathology classes. The modification involves placing each heart segment (LV, RV, and Myo) in separate channels of the RGB color model. All images were cropped according to the largest localization area from a single patient's dataset and have a uniform resolution (64 × 64) with an aspect ratio of 1:1. The entire dataset includes, like the original dataset, 150 patients divided into five groups. An example element of dataset D_6 is shown in Figure 12.

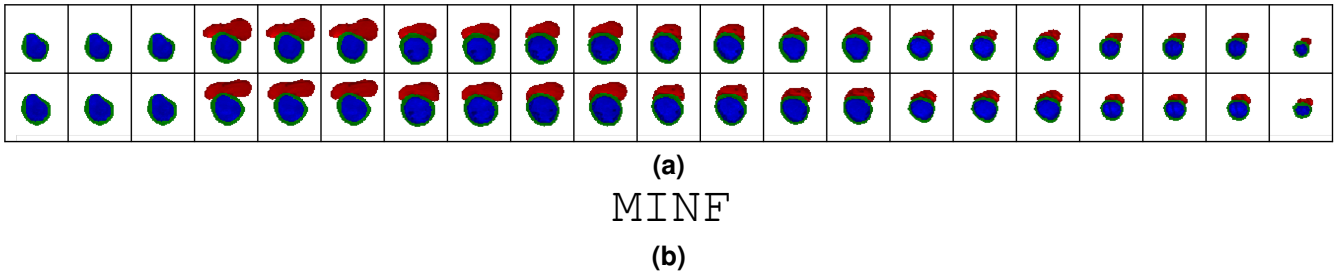


Figure 12. Example element of dataset D_6: **(a)** set of modified MRI scans and **(b)** pathology class label (MINF).

Thus, five new datasets D_2–D_6 were created, which are used for localizing regions of interest, determining precise contours in MRI scans, and classifying MRI scans into pathology classes.

4 Results and Discussion

4.1 Results of the Segmentation Method

To validate the proposed multi-stage segmentation method, a comparison of results with those from other studies was conducted, specifically comparing the segmentation accuracy of individual heart structures using the Dice coefficient. Additionally, to confirm the positive impact of the proposed method steps on the final segmentation quality, a series of experiments were performed, progressively increasing the complexity of the model training by adding individual steps or their combinations while calculating segmentation accuracy metrics. The following experiments were conducted:

1. Segmentation of original MRI scans: To determine the baseline capability of the DL model to recognize heart structures.
2. Segmentation of localized MRI scans using original masks: To evaluate the impact of preliminary localization on the final segmentation result.
3. Segmentation of original images using decomposed masks: To assess the use of binary segmentation instead of multi-structure segmentation.
4. Segmentation of localized MRI scans using decomposed masks: To evaluate the combination of preliminary localization and binary segmentation instead of multi-structure segmentation.

5. Segmentation of localized MRI scans using decomposed masks with post-processing (proposed approach): To assess the combination of all steps of the proposed method.

These experiments aim to determine each step's contribution to improving segmentation accuracy, particularly in complex areas such as RV and Myo. To objectively measure the contribution of each component of our proposed method, we conducted a detailed ablation study. The results, summarized in Table 2, demonstrate the incremental improvements gained from localization, mask decomposition, and post-processing steps.

The authors of the ACDC dataset[55] separately formed training and testing datasets consisting of 100 and 50 patients, respectively. Using pre-formed datasets instead of custom splits allows for more accurate comparisons of experimental results with those from other studies by utilizing the same test dataset. Therefore, all experiments in this section use the pre-formed test dataset provided by the authors.

4.1.1 Segmentation of Localized MRI Scans using Decomposed Masks with post-processing (Proposed Approach)

In the fifth and final stage of the experiments, all previous stages were applied, thereby implementing the complete scheme of the proposed method, which combines localization, decomposition of masks, and post-processing of mask contours.

Localization and mask generation were conducted similarly to the previous experiment, utilizing models (M_1–M_6). Unlike previous experiments, after obtaining segmentation results, post-processing was performed to enhance the quality of the segmented masks. This stage included contour smoothing to eliminate artifacts that might have arisen from minor mislabeling or noise in the data. post-processing also involved applying blurring methods to reduce sharp transitions between different regions and create a more natural appearance of the segmented structure contours. Additionally, masks were resized back to their original dimensions to ensure accurate comparison with expert masks.

Applying post-processing allowed residual artifacts that could degrade segmentation quality, especially for the RV where weak contrast and overlap with adjacent tissues are common, to be removed. Masks obtained after this stage demonstrated higher conformity to expert masks. The results obtained for the final stage of the experiments are shown separately for ED and ES in Table 1.

Table 1. Mean Dice scores and 95% CIs for each cardiac structure (LV, RV, and Myo) across ED and ES phases from the fifth experiment. Values in **bold** indicate the highest segmentation accuracy.

Models	ED			ES		
	LV	RV	Myo	LV	RV	Myo
$M_1^{\exp_1}$	0.911 [0.902–0.920]	0.842 [0.830–0.854]	0.812 [0.800–0.824]	0.890 [0.880–0.900]	0.871 [0.860–0.882]	0.832 [0.820–0.844]
M_1–M_6 + post-processing	**0.974 [0.968–0.980]**	**0.947 [0.938–0.956]**	**0.896 [0.885–0.907]**	**0.940 [0.930–0.950]**	**0.915 [0.905–0.925]**	**0.920 [0.910–0.930]**

The proposed approach achieved the best results among all experimental stages, providing maximum segmentation accuracy for all studied structures, particularly the LV and Myo, which are crucial for medical analysis. Therefore, the combination of localization, decomposition, and post-processing confirmed its effectiveness as a comprehensive strategy for automated segmentation of heart structures in MRI scans. The comparison of all experimental results for evaluating the effectiveness of the localization, decomposition, and post-processing steps is presented in Table 2.

A visual comparison of all experimental results is depicted in Figure 13.

Thus, the experiments demonstrated an increase in segmentation accuracy using the proposed method, which includes localization, decomposition, and post-processing of images. The overall segmentation accuracy for all structures according to the average Dice coefficient increased to 0.932, which is 7.76% higher compared to the base model $M_1^{\exp_1}$. Therefore, the proposed approach ensures high segmentation accuracy of heart structures in MRI scans, which is critically important for subsequent clinical analysis and diagnosis.

4.1.2 Comparison of the Proposed Approach with Current Segmentation Methods

To objectively situate our method's performance within the current research landscape, we conducted a comprehensive comparison against several state-of-the-art techniques on the ACDC challenge dataset. The quantitative results, benchmarked using the Dice coefficient for both ED and ES phases, are summarized in Table 3.

Our hierarchical 2D approach demonstrates a highly competitive and balanced performance profile across all cardiac structures. For the LV, our method achieves a Dice score of 0.974 in ED, closely rivaling top-performing 2D and 3D models like Zhang et al.[43] (0.976) and Benameur et al.[66] (0.978). A key strength of our approach is its exceptional performance in

Table 2. Comparative evaluation of segmentation accuracy using the Dice coefficient across different experimental setups to assess the impact of localization, decomposition, and post-processing. Values are presented as mean[95% CI] for the LV, RV, and Myo in both ED and ES phases. Values in **bold** indicate the highest segmentation accuracy achieved among the experiments.

Experiments	ED			ES		
	LV	RV	Myo	LV	RV	Myo
Experiment 1	0.911 [0.902–0.920]	0.842 [0.830–0.854]	0.812 [0.800–0.824]	0.890 [0.880–0.900]	0.871 [0.860–0.882]	0.832 [0.820–0.844]
Experiment 2	0.920 [0.912–0.928]	0.902 [0.892–0.912]	0.875 [0.865–0.885]	0.894 [0.884–0.904]	0.891 [0.881–0.901]	0.884 [0.874–0.894]
Experiment 3	0.919 [0.911–0.927]	0.892 [0.882–0.902]	0.855 [0.844–0.866]	0.887 [0.877–0.897]	0.873 [0.862–0.884]	0.885 [0.875–0.895]
Experiment 4	0.956 [0.948–0.964]	0.939 [0.930–0.948]	0.866 [0.855–0.877]	0.930 [0.921–0.939]	0.905 [0.895–0.915]	0.898 [0.888–0.908]
Experiment 5 (**Ours**)	**0.974 [0.968–0.980]**	**0.947 [0.938–0.956]**	**0.896 [0.885–0.907]**	**0.940 [0.930–0.950]**	**0.915 [0.905–0.925]**	**0.920 [0.910–0.930]**

segmenting the RV, traditionally a challenging task due to its complex crescent shape and lower blood-Myo contrast. With Dice scores of 0.947 (ED) and 0.915 (ES), our method is on par with or exceeds the results of leading competitors in this specific area.

Perhaps most notably, our approach excels in segmenting the Myo during the ES phase, achieving a Dice score of 0.920. This is the highest score among the compared methods and is particularly significant because accurate ES Myo delineation is critical for functional assessments like wall thickening, yet it is notoriously difficult due to the thinness and dynamic motion of the muscle wall at peak contraction. This result underscores the efficacy of our structure-specific models, which can learn the nuanced features of each anatomical component under different physiological states. While complex monolithic architectures like CondenseUNet[42] or FC-DenseNet with capsule convolutions[43] provide strong generic feature extraction, their accuracy often plateaus during this dynamic phase. Specifically, our method outperforms these approaches for the ES Myo (Dice 0.920 compared to 0.913[42] and 0.918[43]). This demonstrates the unique advantage of our structure-specialized decomposition: dedicating a separate model solely to the Myo better captures its specific morphological changes during the ES phase compared to standard multi-class networks.

The comparison in Table 3 includes methods with diverse architectures, such as the 3D U-Net employed by Benameur et al.[66].

Table 3. Benchmarking of the proposed segmentation method against state-of-the-art approaches on the ACDC dataset using the mean Dice coefficient. Results are detailed for the LV, RV, and Myo across both ED and ES phases. Values in **bold** highlight the highest scores within each category. Methods marked with an asterisk (*) utilize 3D architectures that leverage inter-slice spatial information. Our method's results are presented as mean [95% CI].

Works	ED			ES		
	LV	RV	Myo	LV	RV	Myo
Hu et al.[20]	0.968	0.946	0.902	0.931	0.899	0.919
da Silva et al.[21]	0.963	0.932	0.892	0.911	0.883	0.901
Ammar et al.[44]	0.964	0.935	0.889	0.917	0.879	0.898
Bourfiss et al.[45]	0.959	0.929	0.875	0.921	0.885	0.895
Hasan & Linte[42]	0.963	0.924	0.901	0.929	0.887	0.913
Zhang et al.[43]	0.976	**0.949**	0.903	**0.950**	0.916	0.918
Benameur et al. (*)[66]	**0.978**	0.945	**0.905**	0.948	**0.919**	0.917
Ours	0.974 [0.968–0.980]	0.947 [0.938–0.956]	0.896 [0.885–0.907]	0.940 [0.930–0.950]	0.915 [0.905–0.925]	**0.920 [0.910–0.930]**

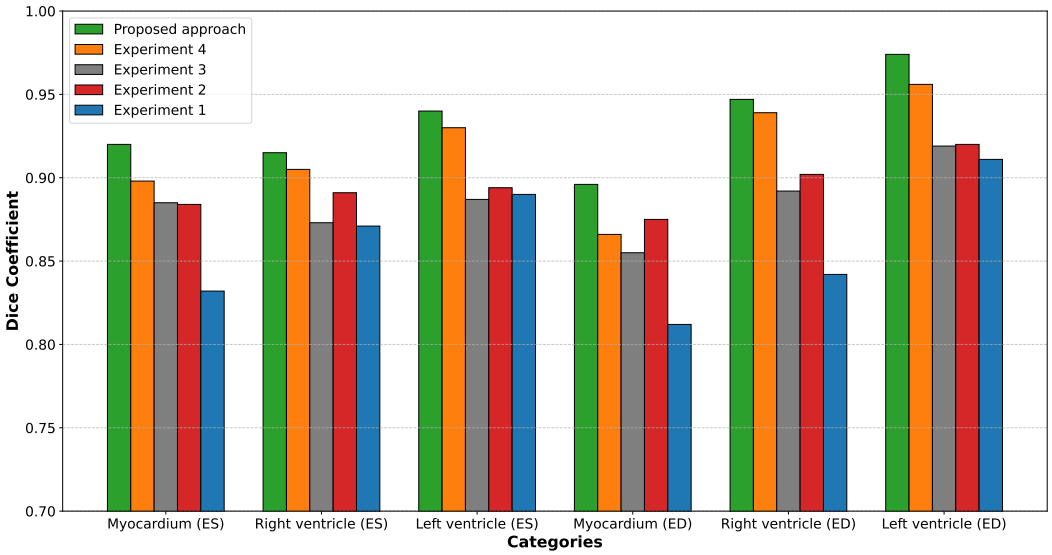


Figure 13. Visual comparison of obtained Dice coefficient values for experiments 1–5 (ES – end-systole, ED – end-diastole).

While 3D models inherently leverage inter-slice spatial context to ensure volumetric smoothness, often leading to excellent results, they typically come with a higher computational and memory cost. Our 2D-based pipeline, while computationally more lightweight, proves to be a powerful alternative by achieving comparable accuracy and even outperforming in certain demanding sub-tasks. It is also crucial to contextualize results from different imaging modalities. While studies like Zhu et al.[67] report exceptional LV segmentation on myocardial perfusion Single-Photon Emission Computed Tomography (SPECT) images, a direct numerical comparison is not appropriate due to fundamental differences in image resolution, contrast, and underlying physics between SPECT and cine MRI. We therefore focus our benchmark on methods validated on the same modality and dataset.

Summarizing the overall performance, our proposed method achieves an average Dice coefficient of 0.932 across all structures and phases. This represents a tangible improvement over many established methods, such as those by da Silva et al.[21] (0.913) and Bourfiss et al.[45] (0.910). The visual comparison provided in Figure 14 further illustrates the competitive standing of our approach.

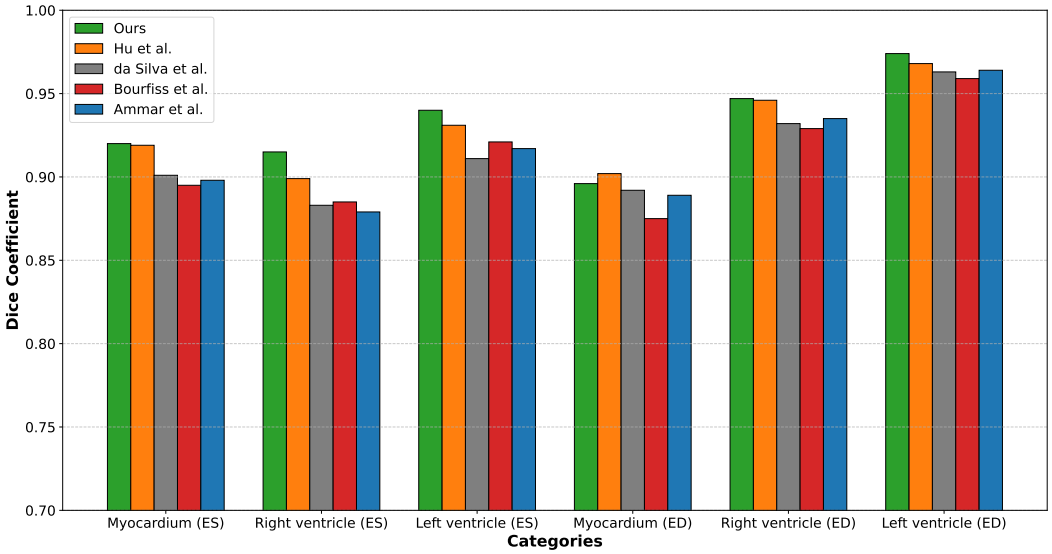


Figure 14. Visual comparison of obtained Dice coefficient values with the state of the art (ES – end-systole, ED – end-diastole).

In conclusion, the results strongly indicate that our multi-stage approach, which combines specialized localization, decomposition, and post-processing, is not only highly accurate but also robust, stable across the cardiac cycle, and particularly effective for challenging anatomical regions, making it a promising tool for clinical and research applications.

4.1.3 Validation of Expert Masks and Obtained Results by a Medical Specialist

This study utilized the expert-provided ground truth masks from the ACDC dataset for training and initial validation. However, to rigorously assess the performance of our generated masks, we conducted an additional validation step involving a medical specialist. During the training and testing of the proposed models on the original dataset, certain inaccuracies in the segmentation masks were identified. Even minor discrepancies between training and testing masks can significantly impact the evaluation accuracy. These inaccuracies may involve incorrect labeling of image regions or missing segments. For instance, in Figure 15, an "extra" area in the upper part of the LV was erroneously marked by experts as part of the Myo (Zone 1 in Figure 15c). In contrast, the proposed method accurately separated this region from the Myo. Similarly, the model correctly identified a region in the right Myo that experts had mistakenly excluded (Zone 2 in Figure 15c).

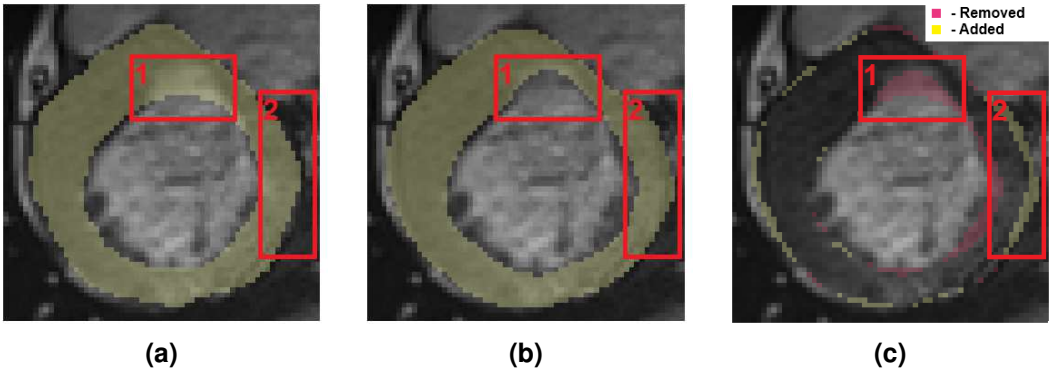


Figure 15. This figure showcases the impact of insufficient brightness on cardiac MRI analysis. Panel **(a)** shows the expert mask, **(b)** the proposed method's mask, and **(c)** their differences, where pink marks removed areas and yellow marks added regions. The red rectangles labeled "1" and "2" highlight notable discrepancies in the upper and right parts of the mask.

A detailed analysis of masks with less than 95% overlap between expert and model-generated masks suggested that some discrepancies arose not from model errors but because the DL model provided more precise segmentations differing from expert annotations. To address this, a medical specialist was engaged to validate both the training and testing datasets alongside all the results obtained.

To ensure a comprehensive quality assessment, a unique mask was generated for each image in the combined dataset. The original test and training datasets were merged, shuffled, and split into new training and testing sets in a 70/30 ratio, ensuring no overlap of masks between sets. This process enabled a thorough comparison of all masks.

To independently assess the clinical quality of the generated segmentations, the results were validated by a practicing cardiologist using a custom-developed software application, the interface of which is shown in Figure S1 in the Supplementary Information. This tool facilitates a rigorous side-by-side comparison by enabling dynamic adjustments to image contrast and brightness, as well as the opacity of the overlaid masks. To mitigate observer bias, the validation protocol was rigorously designed: masks were anonymized and presented as 'A' and 'B', with their order randomized for each image, and the decision time was recorded. To ensure the robustness of this qualitative assessment, a second, independent cardiologist reviewed a subset of 50 randomly selected cases to establish inter-rater reliability, achieving a Cohen's κ of 0.85 (95% CI: 0.78–0.92), indicating substantial agreement. For each comparison, the validating clinician was tasked with selecting one of four definitive options: (i) Mask A has higher accuracy, (ii) Mask B has higher accuracy, (iii) both masks have equally high accuracy with clinically insignificant differences, or (iv) both masks have equally poor accuracy, insufficient for medical conclusions.

The results of the specialist's analysis are as follows:

- Myo masks: 133 masks generated by the proposed method were deemed more accurate, 56 by experts were more accurate, and 1 mask showed equally poor accuracy.
- LV masks: 136 masks from the proposed method were more accurate, 52 expert masks were more accurate, and 2 masks showed equally poor accuracy.
- RV masks: 122 masks from the proposed method were more accurate, 30 expert masks were more accurate, and 5 masks showed equally poor accuracy.

The varying number of masks across categories (LV, RV, and Myo) is due to the incomplete visualization of all heart regions in every MRI scan. For example, the RV is often visible in fewer MRI slices than the LV.

These findings demonstrate that the proposed segmentation method achieves high accuracy, with 97.73% of masks meeting the quality standards required for medical analysis. The remaining 2.27% of masks exhibited lower quality, primarily due to poor input image quality, such as low resolution, motion artifacts, or inadequate contrast. Thus, the proposed method is suitable for practical use, although expert involvement remains necessary for validating and refining results in low-quality images.

The results of the cardiologist's analysis of the LV, RV, and Myo masks are presented in Table 4.

Table 4. Qualitative clinical validation results of the generated segmentation masks as assessed by a practicing cardiologist. The table presents the frequency of preferences for masks generated by the proposed method versus expert-annotated ground truth masks across the LV, RV, and Myo.

Description	Number of marked samples		
	Myo masks	LV masks	RV masks
Proposed method mask has higher accuracy	133	136	122
Expert-annotated mask has higher accuracy	56	52	30
Both masks have high accuracy; difference is insignificant	1976	1962	1959
Both masks have low accuracy; insufficient for conclusions	1	2	5

The analysis that was conducted showed that the proposed segmentation method predominantly demonstrates high accuracy. Specifically, 97.73% of masks have satisfactory and sufficient quality for analysis by a medical specialist, confirming the method's effectiveness. The remaining 2.27% of masks exhibited poorer results, mainly due to low-quality input images. The main reasons for reduced quality include:

- Low resolution.
- Image artifacts (e.g., blurring due to patient movement during the MRI).
- Low brightness or contrast caused by technical characteristics of the MRI device or challenges during patient imaging.

Involving a medical specialist and developing a specialized software application for validation allowed for a detailed assessment of the segmentation results. In 133 cases for the Myo, 136 cases for the LV, and 122 cases for the RV, the proposed method demonstrated higher accuracy than expert masks. Concurrently, 1 Myo mask, 2 LV masks, and 5 RV masks were found to have equally low accuracy.

Thus, the obtained results confirm that the proposed segmentation method is suitable for practical use. However, in cases of low-quality input images, the involvement of experts is necessary for refining and validating the results.

4.2 Results of the Classification Method

A series of experiments were conducted to evaluate the effectiveness of various methods for classifying heart pathologies. The primary focus was on the proposed multi-step approach, which utilizes separate classifiers for sequential class separation. The choice of a binary cascade over a single multi-class model directly addresses the severe class confusion between morphologically similar pathologies, such as MINF and DCM. As demonstrated in Figure 16 and Table 6, our single-stage multi-class baselines, specifically ResNet-50 and ResNet-18, struggled with this ambiguity, achieving only 84% and 72% overall accuracy. The cascade architecture resolves this by isolating complex decisions, thereby significantly improving the overall system accuracy. This approach optimizes the classification process by breaking down a complex multiclass task into several more straightforward stages, ensuring high accuracy and stability of results. For comparison of effectiveness, additional experiments were also conducted using multiclass classifiers based on ResNet-50 and ResNet-18 architectures. These experiments aimed to determine how effectively the models can classify all five classes simultaneously without stepwise separation.

The first experiment used the ResNet-50 architecture to classify all five heart pathologies simultaneously. The model achieved an overall accuracy of 84%, which is an acceptable result but significantly lags behind existing outcomes. ResNet-50 performed well in classifying NOR and ARV, but its effectiveness significantly decreased when recognizing MINF and DCM. This is attributed to the high similarity between these classes, which the model could not clearly account for without additional optimization steps or data structuring.

The confusion matrix for the multiclass classifier based on ResNet-50 is presented in Figure 16a, and for ResNet-18 – in Figure 16b.

The second experiment involved testing the ResNet-18 model, which, due to its lower computational power and fewer parameters, demonstrated an even lower level of accuracy – 72%. This result proved insufficient, especially for classifying complex pathologies such as MINF, where the model exhibited a high frequency of confused results. This indicates that ResNet-18 is not sufficiently robust for high-complexity tasks and requires additional adaptation to handle such challenges effectively.

The primary focus remained on the proposed multi-step approach, which utilizes separate classifiers for sequential class separation. This approach optimizes the classification process by breaking down a complex multiclass task into several more straightforward stages, ensuring high accuracy and stability of results.

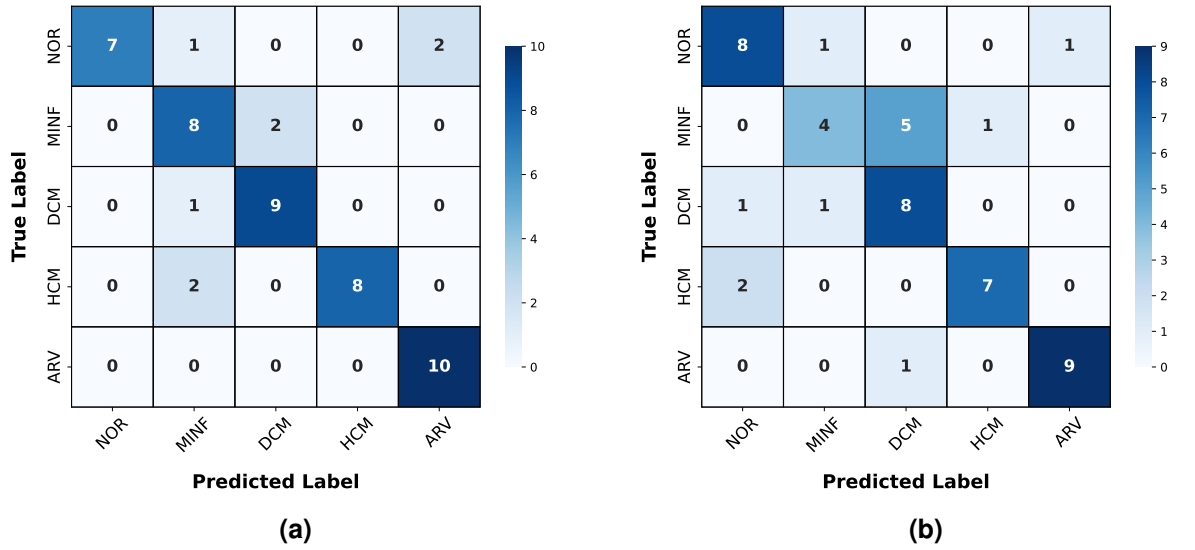


Figure 16. Confusion matrices for multiclass classifiers using two architectures: **(a)** ResNet-50 and **(b)** ResNet-18, showcasing the classification performance across five classes, including NOR, MINF, DCM, HCM, and ARV.

Figure 17 depicts the confusion matrices for each classification stage, showing the number of correct, false positive, and false negative classifications.

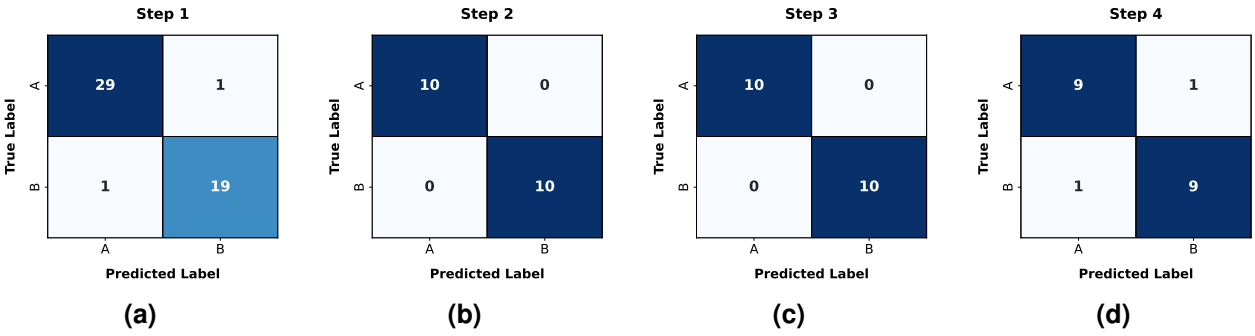


Figure 17. Confusion matrices, illustrating classification performance across sequential steps: **(a)** Step 1—Classifier 1, **(b)** Step 2—Classifier 2, **(c)** Step 3—Classifier 3, and **(d)** Step 4—Classifier 4, highlighting the accuracy and misclassifications for labels A and B at each step of the process.

At the first classification stage, the model achieved an overall accuracy of 0.96, successfully separating LV pathologies (such as MINF, HCM, and DCM) from NOR and ARV. This result demonstrates the model's high reliability in identifying major pathology groups, forming the foundation for further diagnostic refinement.

The second classification stage showed a perfect accuracy of 1.0 (95% CI: 0.94–1.00) in distinguishing NOR from ARV on the test set. While encouraging, this result on a small cohort may indicate potential for overfitting, and performance should be re-assessed on larger, more diverse datasets.

The third classification stage also achieved 1.0 accuracy (95% CI: 0.95–1.00) in Accuracy, Recall, and F_1-score, effectively

separating HCM from other LV pathologies such as MINF and DCM. This again highlights the model's ability to recognize specific pathologies but warrants caution due to the limited sample size of the test set for these sub-classifications.

In the fourth classification stage, the model classified MINF and DCM with an Accuracy of 0.90. Although this result is slightly lower compared to previous stages, it still demonstrates satisfactory effectiveness in distinguishing these complex pathologies, which often overlap in visual and functional manifestations.

The proposed classification method was evaluated using several metrics: precision, recall, F_1-score, and overall accuracy. These metrics were used for each of the four classification stages to assess the detection and separation of different heart pathologies. Table 5 presents the classification results of the proposed model at each stage. It is important to note that the performance metrics and confusion matrices for the specific classifiers (Table 5 and Figure 17) are calculated for the final result, accounting for the risk of error propagation from preceding cascade stages.

Table 5. Performance metrics for the four binary classifiers (Classifiers 1–4) utilized sequentially at steps 1–4 of the proposed hierarchical cascade classification method. Metrics include Precision, Recall, F_1-score, Specificity, and overall Accuracy for each specific classification task.

Classifier	Classes	Precision	Recall	F_1-score	Specificity	Accuracy
Classifier 1	NOR+ARV	0.95	0.95	0.95	0.97	0.96
	MINF+HCM+DCM	0.97	0.97	0.97	0.95	
Classifier 2	NOR	1.00	1.00	1.00	1.00	1.00
	ARV	1.00	1.00	1.00	1.00	
Classifier 3	HCM	1.00	1.00	1.00	1.00	1.00
	MINF+DCM	1.00	1.00	1.00	1.00	
Classifier 4	MINF	0.90	0.90	0.90	0.90	0.90
	DCM	0.90	0.90	0.90	0.90	

Figure 18 illustrates the ROC curves for each classification stage, demonstrating the relationship between sensitivity (true positive rate) and specificity (false positive rate).

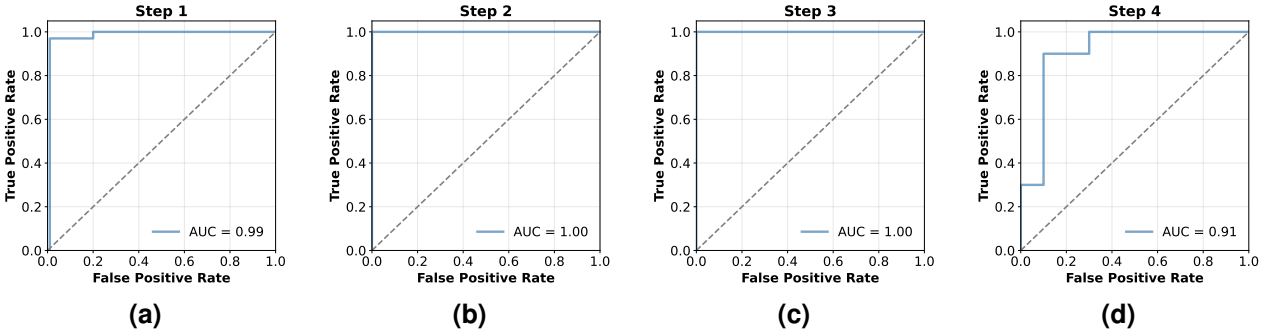


Figure 18. ROC curves for classification steps, showing the true positive rate versus false positive rate for **(a)** Step 1—Classifier 1 (AUC = 0.99 [0.98–1.00]), **(b)** Step 2—Classifier 2 (AUC = 1.00 [0.99–1.00]), **(c)** Step 3—Classifier 3 (AUC = 1.00[0.99–1.00]), and **(d)** Step 4—Classifier 4 (AUC = 0.91 [0.88–0.94]).

AUC for all classifiers confirms the high quality of the proposed approach. An AUC of 1.0 for stages 2 and 3 underscores the model's ability to differentiate classes distinctly. Even in the final stage, where classification accuracy was 0.90, the AUC indicates high model quality in distinguishing MINF and DCM. The AUC is a crucial indicator of overall model quality as it reflects the model's ability to differentiate between classes. Model calibration was assessed via reliability diagrams (see Figure S2 in the Supplementary Information), showing low ECE and MCE values across all cascade stages, indicating that predicted probabilities are well-aligned with observed outcomes.

A detailed comparison of our proposed method with prior work and internal baselines is presented in Table 6. This table expands on the overall accuracy by providing class-wise performance metrics for our cascade, calculated directly from the equations defined in Section 3.5.2, alongside the step-wise results that inform the final diagnosis.

The overall accuracy A_{gen} of our method, as defined by Equation 14, is 97%, a result that surpasses the performance of other state-of-the-art methods, including Khened et al.[23] (96%), Zheng et al.[24] (94%), and Isensee et al.[22] (94%). As detailed

Table 6. Comparative analysis of the proposed cascade classification method against state-of-the-art approaches and internal baselines on the ACDC dataset. The table reports the overall accuracy (A_{gen}) alongside per-class accuracies for ARV, HCM, MINF, DCM, and normal NOR cases. Values in **bold** highlight the best performing scores in each respective column.

Works	Overall Accuracy	Per-class Accuracy				
		NOR	ARV	HCM	MINF	DCM
Isensee et al.[22]	0.94	0.97	0.94	0.94	0.91	0.94
Zheng et al.[24]	0.94	0.95	0.94	0.95	0.92	0.94
Khened et al.[23]	0.96	0.97	0.97	0.97	0.95	0.94
ResNet-50 (multiclass baseline)	0.84	0.96	0.96	0.96	0.92	0.94
ResNet-18 (multiclass baseline)	0.72	0.86	0.80	0.68	0.58	0.68
Ours (cascade)	**0.97**	**0.98**	**0.98**	**0.98**	**0.95**	**0.95**

in Table 6, our model achieves high per-class accuracies of 0.98 for NOR, ARV, and HCM, and 0.95 for the more challenging MINF and DCM classes. This robust performance is a direct result of the cascade's strong step-wise execution, which achieved accuracies of 0.96, 1.00, 1.00, and 0.90 at each respective stage, with corresponding AUC of 0.99[0.98–1.00], 1.00 [0.99–1.00], 1.00 [0.99–1.00], and 0.91[0.88–0.94].

These experimental results clearly demonstrate the advantage of our multi-step approach. The cascade effectively breaks down a complex five-class problem into simpler binary decisions, achieving near-perfect separation in the initial stages. In contrast, our benchmark single-stage multiclass classifiers based on ResNet-50 and ResNet-18 architectures achieved significantly lower overall accuracies of 84% and 72%, respectively. Their struggle to distinguish similar pathologies like MINF and DCM, as shown in Figure 16, underscores the limitations of standard approaches and highlights the superior accuracy and stability of our hierarchical classification strategy.

4.3 External Validation and Domain Shift Analysis

To rigorously evaluate the generalizability of our approach, the models trained exclusively on the ACDC dataset (Siemens 1.5T/3.0T scanners) were deployed without modification on the unseen, multi-vendor, multi-center M&Ms dataset. This external validation serves as a stringent test of robustness against the domain shifts inherent in real-world clinical data, which arise from variations in scanner hardware, acquisition sequences, and image reconstruction algorithms. As summarized in Table 7, the model demonstrated commendable, albeit reduced, overall performance, confirming its foundational robustness.

Table 7. Evaluation of the model's diagnostic generalizability on the internal ACDC test set compared to the external M&Ms validation set. The performance is assessed using mean Dice coefficients for segmentation and AUC for classification, presented as mean [95% CI]. The observed performance drop on the M&Ms dataset illustrates the impact of domain shift introduced by different scanner vendors and clinical sites. Note that classification AUC is marked as not applicable (N/A) for the M&Ms dataset since it does not include matching pathology labels.

Cohort	LV Dice	RV Dice	Myo Dice	AUC (cascade)
ACDC (test)	0.957[0.95–0.96]	0.931[0.92–0.94]	0.908 [0.90–0.91]	0.98 [0.97–0.99]
M&Ms (external)	0.91 [0.90–0.92]	0.86[0.84–0.88]	0.85[0.83–0.86]	N/A

The mean Dice scores on the M&Ms set were 0.91 for the LV, 0.86 for the RV, and 0.85 for the Myo, a predictable decrease from the internal ACDC test set scores but still indicative of strong generalization capabilities (Figure 19).

To dissect the impact of domain shift, we stratified the segmentation performance by the specific scanner vendor and model within the M&Ms dataset, with detailed results presented in Table S2 in the Supplementary Information. This breakdown reveals a clear performance gradient directly correlated with the similarity of the target domain to the training domain. Scanners with technical profiles closely matching the ACDC training data, such as the Philips Achieva 1.5T and other Siemens 1.5T models (Symphony, Avanto), yielded Dice scores nearly on par with the ACDC baseline. For instance, the Philips Achieva maintained high Dice scores of 0.951 (LV-ED) and 0.936 (RV-ED). Performance moderately decreased on Siemens "Tim/Fit" systems, which introduce partial technical variations in coil and reconstruction technologies that create a modest domain gap.

Conversely, a substantial performance degradation was observed on scanners from GE Medical Systems. This drop is attributable to a significant domain gap, driven by several key technical differences. GE scanners often utilize a distinct

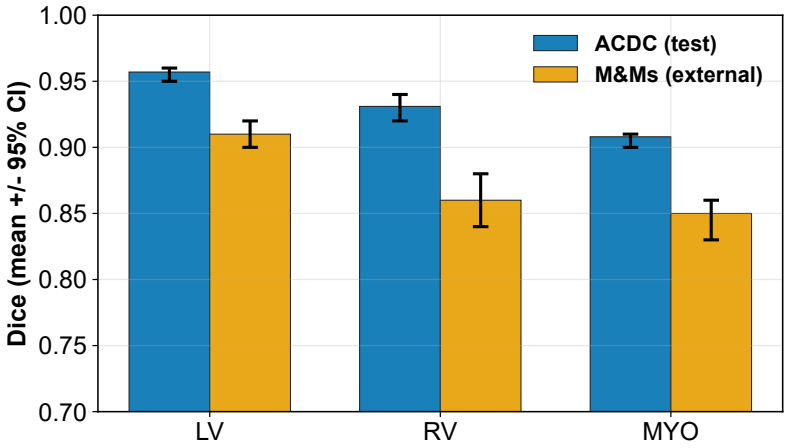


Figure 19. Generalizability of the segmentation model on an external, multi-vendor dataset. The bar chart compares the mean Dice scores for the LV, RV, and Myo on the internal ACDC test set (blue) versus the external M&Ms validation set (orange). The segmentation model, which was trained exclusively on ACDC data, demonstrates strong but predictably lower performance on the unseen M&Ms data, confirming its robustness against domain shift. Error bars represent 95% bootstrap CIs.

balanced steady-state free precession implementation, which produces different blood-Myo contrast and signal characteristics compared to Siemens' TrueFISP sequence used in the ACDC training data. Furthermore, proprietary algorithms for coil signal combination and intensity normalization create different bias-field and intensity profiles that the model has not been trained to handle. Variations in typical imaging parameters (e.g., voxel spacing, slice thickness) and, for the 3.0T scanner, more pronounced B0 field inhomogeneity artifacts, collectively contribute to this pronounced domain shift. These findings underscore a critical point: while our model generalizes well across closely related systems, its performance is sensitive to fundamental differences in imaging physics and software. This highlights the crucial importance of domain adaptation strategies or multi-vendor training datasets for developing universally applicable and reliable clinical tools.

4.4 Results of the Interpretation Method

In this subsection, we present examples of interpreting the classification results using the proposed interpretation method. The features introduced in the interpretation model for individual pathologies were utilized to analyze results. Fifty patients from the test dataset (10 for each class) were analyzed. Below, we provide an example analysis for a specific patient in each class.

4.4.1 Dilated Cardiomyopathy (DCM)

The hallmark of DCM is significant enlargement of the LV and reduced contractility. As shown in Figure 20, the LV ED volume (feature o_1^1) is notably increased (222.60 mL) compared to normal values (100–200 mL for men).

The ES volume (feature o_2^1) is likewise high (156.01 mL), while the LV ejection fraction (feature o_3^1) drops to 29% (normally 55–70%). Although the myocardial mass (feature o_4^1) remains around 133.96 g (almost normal), the ratio of myocardial mass to ED volume (feature o_5^1) is only 0.6, confirming ventricular dilation. On the Myo model, certain LV wall segments (feature o_6^1) appear thinner than usual, reinforcing the DCM diagnosis.

4.4.2 Hypertrophic Cardiomyopathy (HCM)

HCM typically presents with asymmetric hypertrophy and uneven wall thickness (see Figure S3 in the Supplementary Information). In Figure S3, the LV ES volume (feature o_1^2) is relatively small (27.05 mL), suggesting robust contractility. The ejection fraction (feature o_2^2) is elevated at 84%, while the ratio of myocardial mass to ES volume (feature o_3^2) reaches 6.05, reflecting substantial thickening. On the Myo model, certain regions (features o_4^2, o_5^2, o_6^2) appear excessively hypertrophied, confirming an HCM pattern.

4.4.3 Myocardial Infarction (MINF)

MINF often manifests as reduced LV ejection fraction, sometimes alongside myocardial inflammation best detected with specialized imaging. For example, in Figure S4 in the Supplementary Information, the patient's ejection fraction (feature o_2^3) is 26%, significantly below normal. The LV ED volume (feature o_1^3) is 178.17 ml, close to the upper limit, and myocardial mass (feature o_3^3) is 182.44 g. The maximum wall thickness (feature o_4^3) hovers at the high end of normal (12.57 mm), and the mean standard deviation of wall thickness (feature o_5^3) is 2.06 mm, indicating uneven edema.

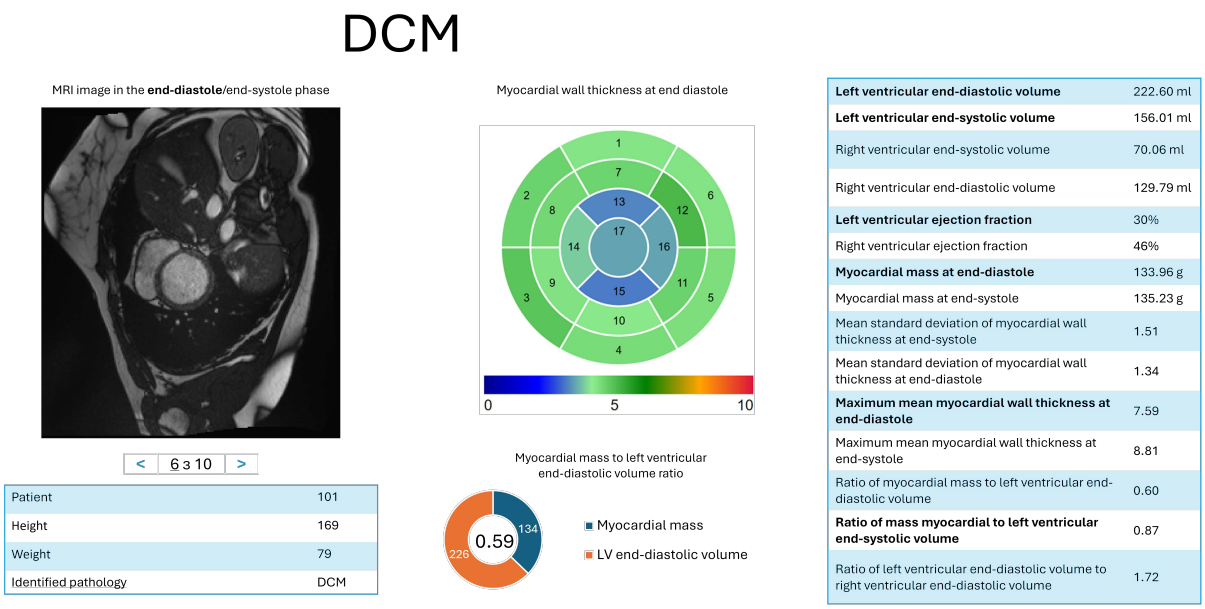


Figure 20. Visualization of DCM, highlighting significant LV enlargement and reduced systolic function, accompanied by detailed metrics, including LV volumes, myocardial mass, ejection fractions, and wall thicknesses, to provide a comprehensive analysis of the pathological condition.

4.4.4 Arrhythmogenic right ventricle (ARV)

ARV features RV dilation and decreased RV ejection fraction (see Figure S5 in the Supplementary Information). In Figure S5, the RV ED volume (feature o_1^4) is markedly high (219.58 mL), far above typical values (80–120 mL). The RV ES volume (145.48 mL) is also elevated, leaving an ejection fraction (feature o_2^4) at 34%. By contrast, the LV remains near normal volumes, though the ratio of LV to RV ED volumes (feature o_3^4) is a low 0.72, confirming pronounced RV dominance.

4.4.5 Normal State (NOR)

In Figure S6 in the Supplementary Information, the patient's LV exhibits near-normal ES volume (feature o_1^5) of 41.75 mL, ED volume (feature o_2^5) of 125.51 mL, and an ejection fraction (feature o_4^5) of 67%, all consistent with healthy cardiac function. The RV volumes and ejection fraction lie within acceptable limits, and myocardial thickness remains stable. Together, these findings confirm the absence of significant pathology.

Overall, each pathology displays a unique profile of volumetric and functional changes: DCM is dominated by LV dilation; HCM shows asymmetric hypertrophy; MINF frequently reveals reduced ejection fraction with tissue irregularities; ARV reflects severe RV enlargement; and NOR remains within healthy ranges. The proposed interpretation approach clarifies each classification outcome in clinical terms by highlighting the most relevant quantitative features and visual cues.

4.5 Limitations of the Proposed Methods

Although the proposed methods for Myo segmentation in the LV and RV show promise, certain limitations require attention.

Firstly, the model's performance can deteriorate significantly when processing low-quality MRI scans. This issue is especially noticeable when parts of the LV, RV, or Myo are not fully visible, resulting in incorrect segmentations or complete omission of these regions. The model heavily relies on detecting differences between the target structures and surrounding tissues; thus, poor image quality seriously affects accuracy.

Another challenge arises when image brightness levels are too low or too high (Figures 21a–b). In such cases, the model may have difficulty correctly defining the boundaries of cardiac structures, resulting in poorly delineated segmentations. This occurs because areas with insufficient or excessive brightness lose crucial detail, negatively affecting model training quality.

Furthermore, the training data used to develop the models lacked sufficient examples of certain pathological conditions, such as cardiomyopathy or noncompaction Myo. This shortage of samples may reduce the model's ability to generalize to more complex conditions, thus affecting its reliability in clinical settings.

While our slice-position-aware refinement module improves segmentation at the ventricular extremes (Table 8), challenges remain in lower-quality scans.

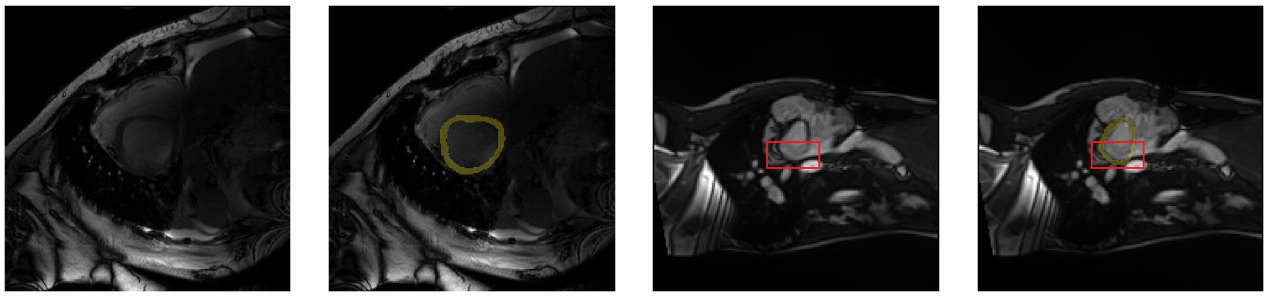


Figure 21. Illustrated challenges in cardiac MRI analysis due to insufficient image quality. In **(a)** and **(b)**, low brightness affects the delineation of cardiac structures, particularly the Myo (highlighted in yellow), leading to poor segmentation. In **(c)** and **(d)**, partial Myo visualization issues (also highlighted in light yellow), caused by slice timing, affect the model's accuracy, as shown by the incomplete lower Myo (marked in a red rectangle).

Table 8. Mean Dice coefficients categorized by slice position (basal, mid-ventricular, and apical) and cardiac phase (ED and ES), evaluated on the ACDC test set. The comparison highlights the significant accuracy improvements achieved by implementing the slice-position-aware refinement module, especially in the anatomically challenging apical and basal slices. Values in **bold** emphasize the highest scores within each comparative pair.

Slice Position & Method	ED Dice			ES Dice		
	LV	RV	Myo	LV	RV	Myo
Apical (w/o refinement)	0.931	0.865	0.811	0.892	0.823	0.845
Apical (with refinement)	**0.958**	**0.901**	**0.849**	**0.925**	**0.866**	**0.880**
Mid (w/o refinement)	0.982	0.955	0.912	0.951	0.922	0.931
Mid (with refinement)	**0.983**	**0.956**	**0.915**	**0.952**	**0.924**	**0.933**
Basal (w/o refinement)	0.945	0.889	0.833	0.908	0.850	0.871
Basal (with refinement)	**0.969**	**0.920**	**0.865**	**0.933**	**0.881**	**0.902**

In some images, specific segments fail to visualize or only partially appear, often due to slice timing (e.g., capturing the moment the mitral or tricuspid valve is opening). The model may still produce inaccurate results if segments are not clearly visible. Figures 21c-d show a case where the lower Myo is poorly visualized but is still marked in the mask. Such ground-truth ambiguities can teach the network to erroneously fill these areas in similar cases, a residual error our current method does not fully resolve.

Hence, although the approach is reliable under ideal conditions, its effectiveness strongly depends on the quality of the input data. Extra caution is needed when working with low-quality images or rare pathologies, as these factors can reduce accuracy and make the model less dependable in critical diagnostic situations.

4.6 Discussion

This study introduces a hierarchical, multi-stage approach designed to address critical, unresolved challenges in cardiac MRI analysis. These challenges include robust segmentation of complex anatomies, accurate classification of morphologically similar pathologies, and clinically meaningful interpretation of AI-driven decisions. As detailed in Section 4, our computational experiments provide compelling evidence that this integrated pipeline successfully resolves these issues, yielding a system that is highly accurate, transparent, and seamlessly aligned with clinical workflows.

Our multi-stage segmentation method (described in Section 3.2) was systematically validated through ablation studies, which demonstrated that each component—localization, structure-specific decomposition, and post-processing—incrementally enhanced the final accuracy (see Table 2 and Table 1). In contrast to standard monolithic architectures commonly utilized in the field[19], our comprehensive approach achieved state-of-the-art results, with Dice coefficients reaching 0.974 for the LV in the ED phase. Critically, it achieved a Dice score of 0.920 for the Myo in the ES phase, a traditionally challenging target. As reported in Table 3, this performance is highly competitive, outperforming the results reported by other leading models discussed in Section 1, such as the approaches by Hu et al.[20] and da Silva et al.[21]. A practising cardiologist corroborated

this high technical accuracy (Table 4), who verified that 97.73% of the generated masks were of sufficient quality for clinical diagnosis, thereby confirming the method's real-world utility.

Similarly, our cascade classification architecture (Section 3.3) was specifically designed to overcome the well-documented failure of standard multiclass models to distinguish between pathologies with overlapping features, such as MINF and DCM. Our experiments starkly illustrate this limitation: baseline ResNet-50 and ResNet-18 models achieved poor overall accuracies of 84% and 72%, respectively, accompanied by significant class confusion (Figure 16). Conversely, our cascade method reached an overall accuracy of 97% (Table 6), substantially surpassing the 94% accuracy of Isensee et al.[22] and Zheng et al.[24], whose models struggled with similar DCM/MINF ambiguities. Furthermore, our hierarchical design provides a more interpretable and potentially more generalizable alternative to the complex multi-classifier ensembles proposed by previous researchers, such as Khened et al.[23]. The robust performance at each cascade stage is further evidenced by the metrics and ROC curves presented in Table 5 and Figure 18.

To address the known difficulty of segmenting apical and basal slices due to partial-volume effects, our methodology employed a two-pronged approach. First, the hierarchical design localizes the heart before applying specialised networks, thereby intrinsically improving performance in noisy regions. Second, we implemented a slice-position-aware refinement module incorporating shape-prior and 3D continuity losses (Algorithm 2). This module specifically targets and corrects errors common at the ventricular extremes, leading to significant, quantifiable improvements in segmentation accuracy across the entire ventricular stack, as comprehensively detailed in Table 8.

A central advantage of our proposed approach is methodological transparency, particularly regarding the handling of papillary muscles (as detailed in Section 3.4.2). Because papillary muscles can be ambiguously assigned to either the myocardial mass or the LV blood pool, this choice dramatically impacts clinical indices. Our approach adheres to a strict policy of excluding them from the Myo, in line with clinical guidelines to ensure reproducibility. A sensitivity analysis revealed that including papillary muscles erroneously increased the calculated LV mass by 8.3% while slightly decreasing the ejection fraction. This highlights the absolute necessity of automated tools adhering to clear, predefined conventions to maintain clinical reliability. However, a notable disadvantage of our current pipeline is its increased reliance on multiple, sequentially executed models, which introduces architectural complexity compared to end-to-end multiclass networks, even though the computational overhead remains low due to the lightweight backbones.

Despite these successful outcomes, this study has several limitations and potential shortcomings that define open questions for future research. First, the model's performance is highly sensitive to input image quality; poor resolution, insufficient brightness, or significant motion artefacts can substantially degrade segmentation accuracy, as illustrated in Figure 21. Second, while successfully validated on the external multi-vendor M&Ms dataset (Table 7 and Figure 19), the model's performance experienced some expected degradation, indicating a vulnerability to domain shifts arising from different scanner hardware and acquisition protocols. Finally, the model's diagnostic capacity is fundamentally constrained by the specific pathologies represented in the ACDC training data, limiting its ability to identify rare or complex cardiac conditions. Consequently, key research challenges remain in creating universally applicable diagnostic tools. Future efforts must focus on exploring advanced domain adaptation techniques, developing robust data normalization methods to mitigate artifacts, and leveraging semi-supervized or few-shot learning to expand the model's diagnostic capabilities across a broader spectrum of heterogeneous, multi-vendor datasets.

5 Conclusions

This study presented a hierarchical DL approach to resolve persistent challenges in cardiac MRI analysis. By decomposing complex tasks, our system achieves state-of-the-art accuracy and clinical transparency. Our structure-specialized segmentation pipeline delivers exceptional precision, evidenced by remarkable mean Dice scores of 0.974 for the LV, 0.947 for the RV, and 0.920 for the ES Myo. Furthermore, our novel cascade classification architecture mitigates diagnostic ambiguity between similar pathologies, achieving a 97% overall accuracy and significantly outperforming standard multiclass baselines.

The primary contribution of this research is bridging the gap between high algorithmic performance and clinical utility. Rather than functioning as an opaque "black box," our interpretation method translates neural network predictions into a verifiable dashboard of canonical clinical metrics. By automatically extracting critical indicators like ventricular volumes, ejection fractions, and wall-thickness maps, the system presents its diagnostic rationale in physician-friendly language. This inherent transparency empowers clinicians to independently verify the AI's findings, fostering the trust essential for routine clinical adoption.

Despite these successes, the methodology faces limitations that define future research directions. The model is sensitive to low-quality images and motion artifacts, which can compromise segmentation precision. Additionally, external validation revealed vulnerabilities to domain shifts from different scanner vendors, and the training data constrain its diagnostic scope.

Future work will prioritize training on diverse, multi-vendor datasets and integrating advanced domain adaptation techniques. Exploring semi-supervised and few-shot learning will further enhance model robustness, expanding its diagnostic capabilities

across a broader spectrum of complex cardiovascular conditions.

Data availability

The source data that support the findings of this study are available from two public challenges. The primary dataset used for training, validation, and testing of our models was the Automated Cardiac Diagnosis Challenge (ACDC) dataset from Bernard et al. (2018). This dataset is publicly available for research purposes upon registration via the official challenge website. For external validation, we utilized the Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation (M&Ms) dataset from Campello et al. (2021), which is available from the challenge organizers under their specified terms of use.

The derived data generated in this study, which are required to reproduce our findings, have been made publicly available. This includes the full set of pre-processed image stacks, cropped regions of interest, and the precise segmentation masks used for training and evaluating all models presented in this manuscript. This dataset is openly available in the Zenodo repository at https://doi.org/10.5281/zenodo.17079830 (2025).

Code availability

All code for preprocessing, model training, and evaluation will be released in a versioned repository with a tagged release, comprehensive documentation, and an environment file to ensure reproducibility. We provide a one-command inference script, fixed random seeds, and exact library versions. The source code is available at: https://github.com/vitalii-slobodzian/cardiac-mri-analysis/tree/main (2026).

References

1. Morales, M. A., Manning, W. J. & Nezafat, R. Present and future innovations in AI and cardiac MRI. *Radiology* **310**, e231269, DOI: 10.1148/radiol.231269 (2024).
2. Srinivasan, S. M. & Sharma, V. Applications of AI in cardiovascular disease detection: A review. In Singh, R., Gehlot, A., Rathour, N. & Akram, S. V. (eds.) *AI in disease detection: Advancements and applications*, chap. 6, 123–146, DOI: 10.1002/9781394278695.ch6 (John Wiley & Sons, Ltd, 2025).
3. Ball, J. R. & Balogh, E. Improving diagnosis in health care: Highlights of a report from the national academies of sciences, engineering, and medicine. *Annals Intern. Medicine* **164**, 59–61, DOI: 10.7326/M15-2256 (2016).
4. World Health Organization. *Invisible numbers: The true extent of noncommunicable diseases and what to do about them* (World Health Organization, Geneva, 2022).
5. Counseller, Q. C. & Aboelkassem, Y. Recent technologies in cardiac imaging. *Front. Med. Technol.* **4**, 984492, DOI: 10.3389/fmedt.2022.984492 (2023).
6. Rose, N. E. & Gold, S. M. A comparison of accuracy between clinical examination and magnetic resonance imaging in the diagnosis of meniscal and anterior cruciate ligament tears. *Arthrosc. The J. Arthrosc. & Relat. Surg.* **12**, 398–405, DOI: 10.1016/S0749-8063(96)90032-8 (1996).
7. Kramer, C. M. *et al.* Standardized cardiovascular magnetic resonance imaging (CMR) protocols: 2020 update. *J. Cardiovasc. Magn. Reson.* **22**, 17, DOI: 10.1186/s12968-020-00607-1 (2020).
8. Haskell, M. W., Nielsen, J.-F. & Noll, D. C. Off-resonance artifact correction for magnetic resonance imaging: A review. *NMR Biomed.* **36**, e4867, DOI: 10.1002/nbm.4867 (2023).
9. DeBenedectis, C. M. *et al.* Health care disparities in radiology: A review of the current literature. *J. Am. Coll. Radiol.* **19**, 101–111, DOI: 10.1016/j.jacr.2021.08.024 (2022). Part B.
10. Meidert, U., Dönnges, G., Bucher, T., Wieber, F. & Gerber-Grote, A. Unconscious bias among health professionals: A scoping review. *Int. J. Environ. Res. Public Heal.* **20**, 6569, DOI: 10.3390/ijerph20166569 (2023).
11. Radiuk, P., Barmak, O., Manziuk, E. & Krak, I. Explainable deep learning: A visual analytics approach with transition matrices. *Mathematics* **12**, 1024, DOI: 10.3390/math12071024 (2024).
12. Shick, A. A. *et al.* Transparency of artificial intelligence/machine learning-enabled medical devices. *npj Digit. Medicine* **7**, 21, DOI: 10.1038/s41746-023-00992-8 (2024).
13. Ribeiro, M. T., Singh, S. & Guestrin, C. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 1135–1144, DOI: 10.1145/2939672.2939778 (Association for Computing Machinery, 2016).

14. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, 4765–4774 (Curran Associates, Inc., 2017).

15. Selvaraju, R. R. *et al.* Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626, DOI: 10.1109/ICCV.2017.74 (IEEE, 2017).

16. Salih, A. *et al.* Explainable artificial intelligence and cardiac imaging: Toward more interpretable models. *Circ. Cardiovasc. Imaging* **16**, e014519, DOI: 10.1161/CIRCIMAGING.122.014519 (2023).

17. Radiuk, P., Barmak, O. & Krak, I. An approach to early diagnosis of pneumonia on individual radiographs based on CNN information technology. *The Open Bioinforma. J.* **14**, 93–107, DOI: 10.2174/1875036202114010093 (2021).

18. Wang, X., Wang, F. & Niu, Y. Two-stage CNN whole heart segmentation combining image enhanced attention mechanism and metric classification. *J. Digit. Imaging* **36**, 124–142, DOI: 10.1007/s10278-022-00708-6 (2023).

19. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351 of *Lecture Notes in Computer Science*, 234–241, DOI: 10.1007/978-3-319-24574-4_28 (Springer International Publishing, 2015).

20. Hu, H., Pan, N. & Frangi, A. F. Fully automatic initialization and segmentation of left and right ventricles for large-scale cardiac MRI using a deeply supervised network and 3D-ASM. *Comput. Methods Programs Biomed.* **240**, 107679, DOI: 10.1016/j.cmpb.2023.107679 (2023).

21. da Silva, I. F. S., da Silva, J. A. S., Giraldi, G. d. S. & Borges, D. L. A multi-stage automatic method based on a combination of fully convolutional networks for cardiac segmentation in short-axis MRI. *Appl. Sci.* **14**, 7352, DOI: 10.3390/app14167352 (2024).

22. Isensee, F. *et al.* Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, vol. 11045 of *Lecture Notes in Computer Science*, 120–129, DOI: 10.1007/978-3-030-00889-5_14 (Springer International Publishing, 2018).

23. Khened, M., Kollerathu, V. A. & Krishnamurthi, G. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. *Med. Image Analysis* **51**, 21–45, DOI: 10.1016/j.media.2018.10.004 (2019).

24. Zheng, Q., Delingette, H. & Ayache, N. Explainable cardiac pathology classification on cine MRI with motion characterization by semi-supervised learning of apparent flow. *Med. Image Analysis* **56**, 80–95, DOI: 10.1016/j.media.2019.06.002 (2019).

25. Szabo, L. *et al.* Clinician's guide to trustworthy and responsible artificial intelligence in cardiovascular imaging. *Front. Cardiovasc. Medicine* **9**, 1016032, DOI: 10.3389/fcvm.2022.1016032 (2022).

26. Watanabe, A., Ketabi, S., Namdar, K. & Khalvati, F. Improving disease classification performance and explainability of deep learning models in radiology with heatmap generators. *Front. Radiol.* **2**, 991683, DOI: 10.3389/fradi.2022.991683 (2022).

27. Zhao, Y.-q., Gui, W.-h., Chen, Z.-c. & Tang, J.-t. Medical images edge detection based on mathematical morphology. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, 6492–6495, DOI: 10.1109/IEMBS.2005.1615986 (IEEE, 2005).

28. Xu, A., Wang, L., Feng, S. & Qu, Y. Threshold-based level set method of image segmentation. In *2010 3rd International Conference on Intelligent Networks and Intelligent Systems*, 703–706, DOI: 10.1109/ICINIS.2010.181 (IEEE, 2010).

29. Çığla, C. & Alatan, A. A. Region-based image segmentation via graph cuts. In *2008 15th IEEE International Conference on Image Processing (ICIP)*, 2272–2275, DOI: 10.1109/ICIP.2008.4712244 (IEEE, 2008).

30. Liu, X., Song, L., Liu, S. & Zhang, Y. A review of deep-learning-based medical image segmentation methods. *Sustainability* **13**, 1224, DOI: 10.3390/su13031224 (2021).

31. Berezsky, O., Liashchynskyi, P., Pitsun, O. & Izonin, I. Synthesis of convolutional neural network architectures for biomedical image classification. *Biomed. Signal Process. Control.* **95**, 106325, DOI: 10.1016/j.bspc.2024.106325 (2024).

32. Badrinarayanan, V., Kendall, A. & Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis Mach. Intell.* **39**, 2481–2495, DOI: 10.1109/TPAMI.2016.2644615 (2017).

33. Huang, Z. *et al.* Revisiting model scaling with a U-net benchmark for 3D medical image segmentation. *Sci. Reports* **15**, 29795, DOI: 10.1038/s41598-025-15617-1 (2025).

34. Fayouka, A., Benameur, N., Mahmoudi, R., Masmoudi, I. & Deriche, M. Cardiac segmentation: A comparative study between 3D UNet and 2D UNet performances. In *2024 IEEE/ACS 21st International Conference on Computer Systems and Applications (AICCSA)*, 1–7, DOI: 10.1109/AICCSA63423.2024.10912528 (IEEE, 2024).

35. Ankenbrand, M. J. *et al.* Deep learning-based cardiac cine segmentation: Transfer learning application to 7T ultrahigh-field MRI. *Magn. Reson. Medicine* **86**, 2179–2191, DOI: 10.1002/mrm.28822 (2021).

36. Radiuk, P., Barmak, O., Manziuk, E., Klymenko, L. & Krak, I. Human-in-the-loop approach based on MRI and ECG for healthcare diagnosis. In *Proceedings of the 5th International Conference on Informatics & Data-Driven Medicine (IDDM-2022)*, vol. 3302 of *CEUR Workshop Proceedings*, 9–20 (CEUR-WS.org, 2022).

37. Singh, A., Singh, K. K. & Izonin, I. Responsible and explainable artificial intelligence in healthcare: Conclusion and future directions. In *Responsible and explainable artificial intelligence in healthcare*, 285–297, DOI: 10.1016/B978-0-443-24788-0.00011-X (Academic Press, 2025).

38. Patil, S. S. *et al.* A domain-shift invariant CNN framework for cardiac MRI segmentation across unseen domains. *J. Digit. Imaging* **36**, 2148–2163, DOI: 10.1007/s10278-023-00873-2 (2023).

39. Wehbe, R. M. *et al.* Deep learning for cardiovascular imaging. *JAMA Cardiol.* **8**, 1089–1098, DOI: 10.1001/jamacardio.2023.3142 (2023).

40. Pandey, S., Chen, K.-F. & Dam, E. B. Comprehensive multimodal segmentation in medical imaging: Combining YOLOv8 with SAM and HQ-SAM models. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2592–2598, DOI: 10.1109/ICCVW60793.2023.00273 (IEEE, 2023).

41. Azam, M. A. *et al.* A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Comput. Biol. Medicine* **144**, 105253, DOI: 10.1016/j.compbiomed.2022.105253 (2022).

42. Hasan, S. M. K. & Linte, C. A. CondenseUNet: A memory-efficient condensely-connected architecture for bi-ventricular blood pool and myocardium segmentation. In *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 11315, 113151A, DOI: 10.1117/12.2550640 (SPIE, 2020).

43. Zhang, J. *et al.* Segmentation of biventricle in cardiac cine MRI via nested capsule dense network. *PeerJ Comput. Sci.* **8**, e1146, DOI: 10.7717/peerj-cs.1146 (2022).

44. Ammar, A., Bouattane, O. & Youssfi, M. Automatic cardiac cine MRI segmentation and heart disease classification. *Comput. Med. Imaging Graph.* **88**, 101864, DOI: 10.1016/j.compmedimag.2021.101864 (2021).

45. Bourfiss, M. *et al.* Towards automatic classification of cardiovascular magnetic resonance Task Force criteria for diagnosis of arrhythmogenic right ventricular cardiomyopathy. *Clin. Res. Cardiol.* **112**, 363–378, DOI: 10.1007/s00392-022-02088-x (2023).

46. Tian, Y., Krishnan, D. & Isola, P. Contrastive multiview coding. In *Computer Vision – ECCV 2020*, vol. 12356 of *Lecture Notes in Computer Science*, 776–794, DOI: 10.1007/978-3-030-58621-8_45 (Springer International Publishing, 2020).

47. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, 1597–1607 (PMLR, 2020).

48. Felfeliyan, B., Keshavarz, M. & Majidi, M. Self-supervised-RCNN for medical image segmentation with limited data annotation. *Comput. Med. Imaging Graph.* **109**, 102297, DOI: 10.1016/j.compmedimag.2023.102297 (2023).

49. Liu, C. *et al.* CUTS: A deep learning and topological framework for multigranular unsupervised medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, vol. 14224 of *Lecture Notes in Computer Science*, 155–165, DOI: 10.1007/978-3-031-43895-0_15 (Springer Nature Switzerland, 2023).

50. Mofrad, F. B. & Valizadeh, G. DenseNet-based transfer learning for LV shape classification: Introducing a novel information fusion and data augmentation using statistical shape/color modeling. *Expert. Syst. with Appl.* **213**, 119261, DOI: 10.1016/j.eswa.2022.119261 (2023).

51. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778, DOI: 10.1109/CVPR.2016.90 (IEEE, 2016).

52. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269, DOI: 10.1109/CVPR.2017.243 (IEEE, 2017).

53. Ali, M. *et al.* Revolutionizing cardiac imaging: A scoping review of artificial intelligence in echocardiography, CTA, and cardiac MRI. *J. Imaging* **9**, 173, DOI: 10.3390/jimaging9080173 (2023).

54. Janik, A., Dodd, J., Ifrim, G., Sankaran, K. & Curran, K. Interpretability of a deep learning model in the application of cardiac MRI segmentation with an ACDC challenge dataset. In *Medical Imaging 2021: Image Processing*, vol. 11596, 1159636, DOI: 10.1117/12.2582227 (SPIE, 2021).

55. Bernard, O. *et al.* Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Med. Imaging* **37**, 2514–2525, DOI: 10.1109/TMI.2018.2837502 (2018).

56. Campello, V. M. *et al.* Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. *IEEE Transactions on Med. Imaging* **40**, 3543–3554, DOI: 10.1109/TMI.2021.3090082 (2021).

57. Wang, Y., Jia, H. & Song, J. Accurate classification of non-ischemic cardiomyopathy. *Curr. Cardiol. Reports* **25**, 1299–1317, DOI: 10.1007/s11886-023-01944-0 (2023).

58. McKenna, W. J., Maron, B. J. & Thiene, G. Classification, epidemiology, and global burden of cardiomyopathies. *Circ. Res.* **121**, 722–730, DOI: 10.1161/CIRCRESAHA.117.309711 (2017).

59. Slobodzian, V., Radiuk, P., Zingailo, A., Barmak, O. & Krak, I. Myocardium segmentation using two-step deep learning with smoothed masks by Gaussian blur. In *Proceedings of the 6th International Conference on Informatics & Data-Driven Medicine (IDDM-2023)*, vol. 3609 of *CEUR Workshop Proceedings*, 77–91 (CEUR-WS.org, 2024).

60. Slobodzian, V., Radiuk, P., Barmak, O. & Krak, I. Multi-stage segmentation and cascade classification methods for improving cardiac MRI analysis. In *Proceedings of the X International Scientific Conference "Information Technology and Implementation" (IT&I 2024)*, vol. 3909 of *CEUR Workshop Proceedings*, 84–98 (CEUR-WS.org, 2025).

61. Izquierdo, C. *et al.* Radiomics-based classification of left ventricular non-compaction, hypertrophic cardiomyopathy, and dilated cardiomyopathy in cardiovascular magnetic resonance. *Front. Cardiovasc. Medicine* **8**, 764312, DOI: 10.3389/fcvm.2021.764312 (2021).

62. Afshin, M. *et al.* Regional assessment of cardiac left ventricular myocardial function via MRI statistical features. *IEEE Transactions on Med. Imaging* **33**, 481–494, DOI: 10.1109/TMI.2013.2287793 (2014).

63. Rainio, O., Teuho, J. & Klén, R. Evaluation metrics and statistical tests for machine learning. *Sci. Reports* **14**, 6086, DOI: 10.1038/s41598-024-56706-x (2024).

64. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, 1321–1330 (PMLR, 2017).

65. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35, DOI: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3 (1950).

66. Benameur, N. *et al.* An improved approach for cardiac MRI segmentation based on a 3D UNet combined with papillary muscle exclusion. *Arab. J. for Sci. Eng.* **50**, 19723–19741, DOI: 10.1007/s13369-025-10310-x (2025).

67. Zhu, F. *et al.* A new method incorporating deep learning with shape priors for left ventricular segmentation in myocardial perfusion SPECT images. *Comput. Biol. Medicine* **160**, 106954, DOI: 10.1016/j.compbiomed.2023.106954 (2023).

Author contributions statement

V.S. and O.B. conceived and conducted the experiments, V.S., O.B., and L.K. validated the results, V.S., O.B., and P.R. made formal analysis, V.S. did investigation, P.R. and L.K. conducted data curation, V.S. and O.B. prepared the draft of the manuscript, P.R., L.K. and I.K. reviewed and edited the manuscript, V.S. and P.R. made visualizations, I.K. and O.B. provided the supervision. All authors reviewed, corrected, and approved the paper.

Additional information

Compliance with ethical standards: We used de-identified, publicly available data from the ACDC and M&Ms challenges; no new institutional review board approval was required for this retrospective study. Our use of the data complied with the respective terms of each challenge. **Competing interests:** The authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SlobodzianManuscriptSupplementaryv4.pdf