Towards Automated Cardiac MRI Assessment: A Multi-Agent CAD Framework for Functional Analysis and Tissue Characterization

Walid Al-Haidri[1], Mukhlis Raza[2], Anatoliy Levchuk[1,3], Kseniya Belousova[1], Maksim Lukin[3], Yeong Hyeon Gu[2(✉)], Mugahed A. Al-antari[2(✉)], Ekaterina Brui[1]

[1] Faculty of physics, ITMO University, St. Petersburg, Russia
{walhaidri, aglevchuk, kmbelousova, ekaterina.brui}@itmo.ru
[2] Department of Artificial Intelligence and Data Science, College of AI Convergence, Sejong University, Seoul, 05006, Republic of Korea
{raza, yhgu, en.mualshz}@sejong.ac.kr
[3] Almazov National Medical Research Centre, St. Petersburg, Russia
lukin_mv@almazovcentre.ru

Abstract. We propose MRI-CARDAIX, a novel multimodal multi-agent framework for integrated analysis of cine and Late Gadolinium Enhancement (LGE) MRI. The proposed computer-aided diagnosis (CAD) framework employs a supervisor agent that coordinates modality-specific cine MRI, LGE MRI, and evidence-retrieval agents for integrated cardiac assessment. All models achieved impressive segmentation accuracy for left ventricular delineation (median DSC ≥ 98.0%), while nnU-Net-v2 and U-Mamba demonstrated the most consistent performance across cardiac structures, with nnU-Net-v2 achieving the highest median DSC for right ventricular segmentation (95.17% [91.69-97.17%]) and comparable myocardial segmentation performance (median DSC 95.70%). For LGE MRI analysis, fibrosis localization achieved an F1-score of 92.98% using Swin-U-Mamba, with precision of 93.01% and recall of 92.76%. Quantitative myocardial fibrosis assessment demonstrated relative volume errors ranging from 5.56% to 11.13% across basal, mid-cavity, and apical regions. Clinical validation of AI-generated cardiac MRI reports using the proposed Total Clinical Reliability Score framework resulted in a mean score of 7.3/10, corresponding to a promising clinical reliability. The supervisor agent achieved 92% routing accuracy (macro-F1 94%) with a mean latency of 1.8 s. Retrieval-augmented answers scored 88% faithfulness and 84% answer relevancy. MRI-CARDAIX provides an interpretable framework for integrated cardiac MRI analysis and reporting. Source code is available on Github https://github.com/MRI-algorithms-and-methods/mri_cardaix

Keywords: Agentic AI; Multi-agent Systems; Cardiac MRI; Multimodal Medical Imaging; Automated Clinical Reporting; 17-segment Segmentation.

1 Introduction

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, accounting for approximately 17.9 million deaths annually [1]. Cardiac Magnetic Resonance (CMR) imaging is the gold standard for non-invasive cardiac assessment. Cine CMR provides high-temporal-resolution evaluation of cardiac function, while late gadolinium enhancement (LGE) MRI enables visualization of myocardial fibrosis and scarring. Clinical interpretation follows the American Heart Association (AHA) 17-segment model [2]. Nonetheless, comprehensive segment-level assessment remains time-consuming and subject to inter-observer variability. Commercial tools such as cvi42 [3] and CAAS MR [4] partially automate analysis but still require separate workflows and manual interpretation. Deep learning has become the dominant approach for automating medical images including cardiac MRI analysis [5, 6]. Most cine MRI methods segment the left ventricle (LV), right ventricle (RV), and myocardium (Myo) for functional assessment [7, 8]. More recent work has extended these techniques to late gadolinium enhancement (LGE) MRI for myocardial scar and fibrosis segmentation [9]. However, accurate localization within the AHA 17-segment model remains challenging because most methods rely on post-hoc geometric partitioning [10, 11]. Cine and LGE MRI are typically analyzed independently, limiting integrated functional structural interpretation within the standardized 17-segment framework and requiring manual cross-modal assessment by clinicians. Multi-agent systems (MAS) coordinate large language models (LLMs) to solve complex medical imaging tasks. Prior agentic frameworks, however, stop short of unified cardiac-specific analysis: M³Builder [12] focuses on general AutoML pipelines rather than integrated cardiac MRI analysis, while MESHAgents [13] lacks supporting end-to-end cardiac MRI analysis and reporting. LLM-based report generation has been improved using RAG methods such as RadAlign [14] remains limited to 2D chest imaging, while cardiac-specific work such as CMR-LLaMA [15] excels at information extraction yet limits generating novel findings directly from raw cine/LGE images. Yet MAS-based joint cine-LGE analysis within the standardized 17-segment model remains underexplored. We therefore propose MRI-CARDAIX and the major contributions are summarized as follows,

1. We propose MRI-CARDAIX, a novel hierarchical multi-agent framework that orchestrates cine functional assessment, LGE tissue characterization, knowledge retrieval, and evidence-grounded reporting for multimodal cardiac MRI analysis.
2. We develop deep learning pipelines for cine and LGE MRIs: the cine agent extracts functional biomarkers via chamber/myocardium segmentation, while the LGE agent performs fibrosis segmentation and 17-segment localization, with both integrated through a shared memory representation.
3. We propose an evidence-grounded LLM reporting framework that combines cine and LGE agent outputs with knowledge base generation to produce clinically coherent reports and support interactive cardiac MRI question answering.
4. We introduce a new multimodal cardiac MRI dataset of paired cine MRI, LGE MRI, and expert clinical reports, enabling end-to-end segmentation-to-reporting development and evaluation.

2 Methodology

As shown in **Fig. 1** MRI-CARDAIX employs a hierarchical multi-agent framework integrating data preprocessing, modality-specific image analysis, knowledge retrieval, and structured report generation. A supervisor agent coordinates task execution and fuses outputs from cine and LGE agents to produce coherent clinical reports.

2.1 Specialized Analysis Agents and Deep Learning Tools

The MRI-CARDAIX comprises three modality-specific analysis agents for cine MRI, LGE MRI, and report generation. **(1) Cine MRI Agent for functional assessment.** This agent processes cine MRI sequences for automated LV, RV, and Myo segmentation at end-diastolic and end-systolic phases. Attention U-Net [16], nnU-Net-v2 [17], U-Mamba [18] and Swin-U-Mamba [19] were evaluated. From these segmentations, the agent computes clinical functional parameters including end-diastolic volume (EDV), end-systolic volume (ESV), stroke volume (SV), ejection fraction (EF), myocardial mass, and regional wall thickness. **(2) LGE Agent for tissue characterization.** This agent deals with LGE pipeline performing scar segmentation and AHA 17-segment localization through a sequence of learned image-analysis steps. First, a ResNet50 predicts the anatomical class of each short-axis LGE slice (basal, mid-cavity, or apical) and estimates the LV center used for standardized CMR cropping. A CNN then segments myocardium and scar tissue. A second CNN uses the cropped MRI data and the predicted slice-class codes to identify the 17 Myo segments. Finally, the segmentation outputs are mapped directly to the 17 AHA segments to localize and quantify fibrosis. **(3) Hybrid Knowledge-Driven: Report Generation and Conversation.** MRI-CARDAIX uses a hybrid RAG framework combining local reports embedded via *mxbai-embed-large* with external biomedical knowledge via Search agent (e.g., PubMed and Google Scholar). Agent outputs are incorporated into modality-aware prompts, while semantically similar reports are retrieved to support report generation and question answering.

2.2 Training and Validation Datasets

DL models (Attention U-Net, nnU-Net-v2, U-Mamba, and Swin-U-Mamba) were fine-tuned using our internal cohort of 124 paired Cine/LGE MRI studies collected at the Almazov National Medical Research Centre, Russia. The data was split into 70% training, 10% validation, and 20% testing. Additionally, we combined the public ACDC (150 cine MRI) [20] and EMIDEC (100 LGE) [21] only for training segmentation models as they lack textual reports.

2.3 Evaluation

The segmentation models, agents, RAG, and LLM components were assessed using a held-out internal test set (n=30). Clinical reliability was assessed in 18 randomly selected cases through blinded review by a board-certified cardiologist with access to the

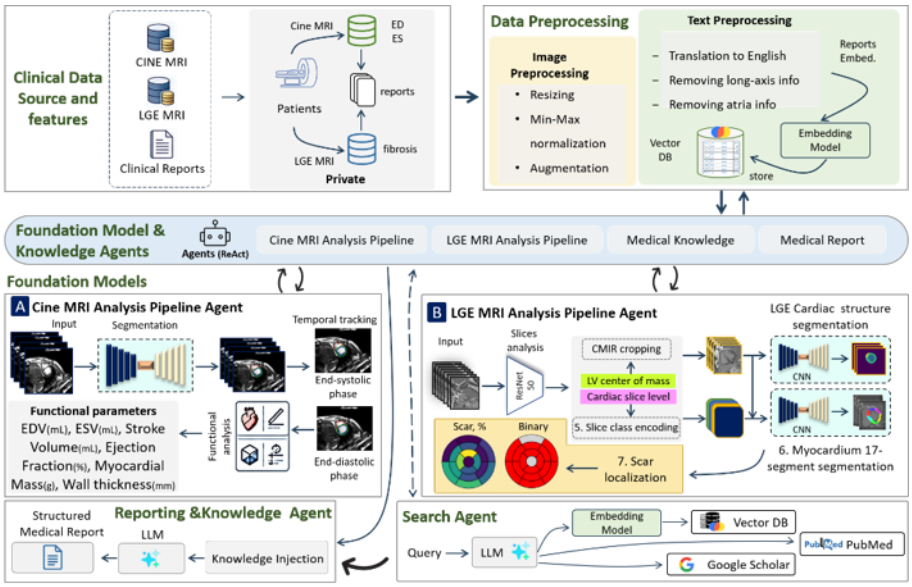

Fig. 1. MRI-CARDAIX: Unified Multi-Agent Architecture for Multimodal Cardiac MRI Analysis, Knowledge Retrieval, and LLM-Driven Clinical Report Generation

original LGE DICOM images and reference reports containing ground-truth (GT) measurements. Segmentation was evaluated using Dice, IoU, F1, precision, and recall, while generated reports were assessed using BLEU-1, ROUGE-L, METEOR, and BERTScore. AI-derived quantitative parameters were compared with GT values, and myocardial tissue characterization was assessed directly from the LGE images. As no standardized metric exists for automated CMR reporting, we developed a 10-point clinical reliability score covering quantitative accuracy (C1: 2), tissue characterization (C2: 2), anatomical localization (C3: 2), clinical consistency (C4: 1), report completeness (C5: 1), radiological terminology (C6: 1), and hallucination safety (C7: 1).

The agentic layer was evaluated on retrieval (36 questions) and routing (47 requests) benchmarks with expert-defined reference labels. Using both *DeepEval* and *RAGAS*, we report four RAG metrics (faithfulness, answer relevancy, context precision, and context recall). Let $q, a, a^*,$ and $\{c_j, \ldots, c_{k=4}\}$ denote the question, generated answer, clinician-written reference answer, and retrieved passages, respectively. Both frameworks use a judge model to decompose text into atomic units $V(\cdot)$ and assess their entailment $C \vDash v$ or relevance. Faithfulness measures how well the generated answer is supported by the retrieved context, while context recall measures whether the evidence required by the reference answer was retrieved. Answer relevancy measures the relevance of the generated answer to the question, whereas context precision evaluates whether relevant evidence is ranked early among the retrieved passages. All metrics range from 0 to 1, with higher values indicating better performance. To assess sensitivity to the evaluation model, all metrics were computed using both a hosted *GPT-4o-mini* judge and a locally hosted *Llama-3.1-8B* judge.

$$D_{RAG} = \{(q_j, a_j^*, C_j)\}_{j=1}^{N_{RAG}}. \quad C_j = \{c_{j,1}, \dots, c_{j,k}\}. \tag{1}$$

$$D_{route} = \{(u_i, F_i, \rho_i^*, \tau_i^*)\}_{i=1}^{N_{route}}. \quad \rho_i^* \in \mathcal{R}. \tag{2}$$

$$Faithfulness = \frac{|\{v \in V(a): C \vDash v\}|}{|V(a)|}, \tag{3}$$

$$Answer\ Relevancy = \frac{1}{|V|} \sum_{v \in V} rel(v, q). \tag{4}$$

$$Context\ Precision = \frac{\sum_{k=1}^{k} p@k.r_k}{\sum_{k=1}^{k} r_k}, \quad p@k = \frac{1}{k}\sum_{j=1}^{k} r_j. \tag{5}$$

$$Context\ Recall = \frac{|\{v \in V(a^*): C \vDash v\}|}{|V(a^*)|}. \tag{6}$$

3 Results

3.1 Segmentation

For cine MRI segmentation, the performance of AI models was evaluated per class: LV, RV, and Myo. As shown in **Fig. 2**, all evaluated methods achieved promising LV segmentation accuracy (median DSC ≥98.32%). Performance differences became more pronounced for the RV and Myo where nnUNet and U-Mamba consistently outperformed Attention U-Net and Swin_Umamba. The nnUNet achieved the highest median DSC for RV DSC (95.17% [Q1:91.69 ~ Q3: 97.17]), while both nnUNet and U-Mamba demonstrated comparable performance for Myo (median DSC ≈ 95.7%). As summarized in **Table 1**, all evaluated models achieved encouraged classification performance (F1: 90.69 ~ 92.98%), with Mamba-based architectures (Swin-U-Mamba, U-Mamba) consistently outperforming CNN-based baselines and showing balanced improvements in both precision and recall. The accuracy of relative Myo fibrosis volume estimation was evaluated separately for the basal, mid-cavity, and apical regions. **Table 2** shows that all models demonstrated progressively larger errors from the basal to the apical segments, indicating that fibrosis quantification is more challenging in apical Myo.

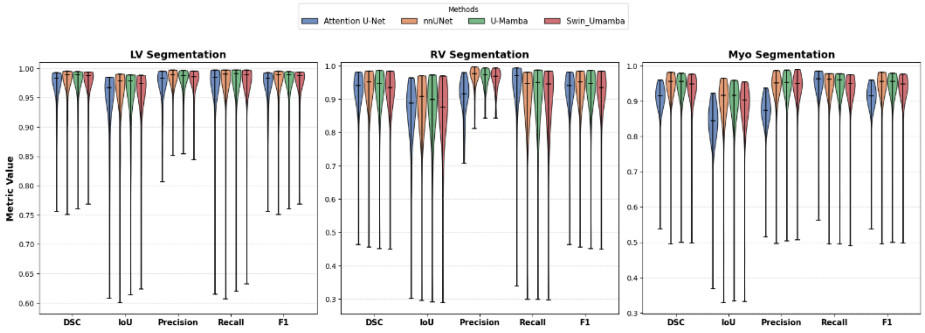

Fig. 2 Cine MRI segmentation performance for LV, RV, and Myo across all AI models.

Table 1. Fibrosis detection evaluation (%)

AI Segmentation Model	F1	Precision	Recall
Attention UNet	90.69	89.97	91.42
nnUNet v2	91.30	92.56	90.08
U-Mamba	92.33	92.70	91.96
Swin-U-Mamba	**92.98**	**93.01**	**92.76**

Table 2. Evaluation (%) of Relative Volume of Myocardium Fibrosis.

	Attention UNet	nnUNet-v2	U-Mamba	Swin-U-Mamba
Basal (1-6)	**5.859**	5.951	6.002	6.045
Mid-cavity (7-12)	**5.563**	6.018	6.154	6.118
Apical (13-16)	**10.038**	10.512	10.669	11.128

Attention U-Net consistently achieved the lowest relative volume error across all three anatomical regions, with errors of 5.859%, 5.563%, and 10.038% for the basal, mid-cavity, and apical segments, respectively.

3.2 LLM Report Generation and Agentic Performance

Tables 3 and 4 summarize the performance of LLMs, agentic workflow, and retrieval framework. Among the evaluated models, Qwen3.5 achieved the best report generation performance, obtaining the highest 85% BLEU-1, 83% METEOR, 87% ROUGE-L, and 97% BERTScore-F1, demonstrating the strongest lexical and semantic agreement with reference reports. MISTRAL and LLaVA also performed competitively, whereas LLaMA, Gemma3, and GPT5.1 showed slightly lower text-generation quality. The agentic framework achieved 91.5% routing accuracy (macro-F1 = 94.2%) with a 100% task success rate. Our system could process requests with a mean latency of 1.78 seconds, a 95th percentile latency of 9.12 seconds, and a throughput of 33.8 queries/min, while maintaining zero execution failures.

Table 3. Comparative evaluation of LLMs on medical report generation

LLM Model	BLEU-1	METEOR	ROUGE-L	BERTScore		
				Precision	Recall	F1
MISTRAL	0.78	0.75	0.85	0.96	0.95	0.95
LLaVA	0.75	0.70	0.82	0.95	0.94	0.94
LLaMA	0.45	0.55	0.53	0.90	0.88	0.89
Gemma3	0.56	0.47	0.59	0.93	0.89	0.91
Qwen3.5	**0.85**	**0.83**	**0.87**	**0.97**	**0.97**	**0.97**
GPT5.1	0.37	0.37	0.52	0.92	0.88	0.90

Table 4. Agentic, system, and RAG evaluation across two LLM judges.

Dimension	Evaluation Metric	Iteration	GPT-4o-mini	Llama-3.1-8B
Agentic	Routing accuracy	47	0.915	0.915
	Routing F1 (macro)	47	0.942	0.942
	Task success rate	42	1.000	1.000
System	Mean latency (ms)	78	1,776	1,771
	P95 latency (ms)	78	9,123	6,602
	Throughput (queries/min)	78	33.8	33.9
	Error rate	78	0.000	0.000
RAG (DeepEval)	Faithfulness	36	0.876	0.738
	Answer relevancy	36	0.838	0.797
	Context precision	36	0.252	0.792
	Context recall	36	0.347	0.533
RAG (RAGAS)	Faithfulness	36	0.367	0.638
	Answer relevancy	36	0.630	0.729
	Context precision	36	0.471	0.518
	Context recall	36	0.176	0.665

Retrieval evaluation showed high faithfulness (0.876) and answer relevancy (0.838) but comparatively lower context precision (0.252) and context recall (0.347), indicating that although generated responses remain clinically coherent, retrieval quality remains the primary bottleneck for further improvement.

3.3 Clinical Validation

The mean Total Clinical Reliability (TCR) score across the 18 evaluated reports was 7.3/10, corresponding to a "Good" level of clinical reliability according to our classification framework. This suggests that the AI-generated reports were generally accurate and captured the key pathological findings, although minor expert verification and corrections were still required before the reports could be considered fully clinically actionable. Due to the limited availability of clinical experts and the time required for comprehensive clinical validation, the current evaluation was conducted by a single investigator. However, evaluation by a single reader may be subject to individual interpretation and evaluator bias. Therefore, for a more robust assessment of the trustworthiness, usefulness, and clinical reliability of MRI-CARDAIX, we recommend future validation using at least three independent cardiologists or relevant clinical experts and reporting inter-rater agreement to quantify the consistency of their assessments. **Fig. 3** illustrates an example of human expert validation, correction, and scoring of AI-generated reports produced by the proposed MRI-CARDAIX framework. The rater evaluates the clinical accuracy and completeness of the generated text based on the outputs of the segmentation and measurement backbones and compares the AI-generated report with the corresponding GT one.

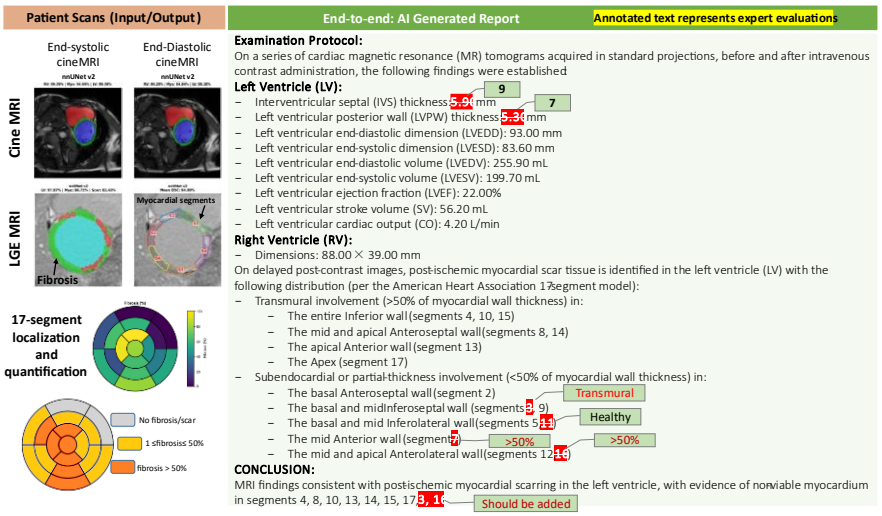

Fig. 3. Multimodal cardiac MRI analysis and automated AI report generation.

4 Discussion

As depicted in **Fig. 3**, MRI-CARDAIX unifies cardiac functional and tissue characterization workflows into a single generated report, achieving excellent segmentation (median DSC ≥ 98.3%), strong fibrosis localization (F1 up to 92.98%), and a mean TCR score of 7.3/10 which required minor expert correction. The RAG system remained faithful and relevant but relied on individual patient reports rather than reference literature, causing population-norm queries to retrieve case-specific measurements. Separately indexing clinical guidelines would therefore be more beneficial than simply expanding the corpus. Routing errors occurred mainly for queries combining viewing verbs with imaging terms, reflecting a keyword-dispatch limitation, while the two LLM judges differed by up to 0.54 in context precision, highlighting the importance of reporting judge identity. Key limitations include evaluation by a single rater, validation of cine MRI against reference reports, and fibrosis localization through expert review of original LGE DICOM images rather than independent re-segmentation. Short-axis LGE views were used by the system while the radiologist had access to long-axis images, potentially affecting localization and parameter extraction. Future work should incorporate multi-rater consensus, multi-planar analysis, and segmentation validation.

5 Conclusion

We presented MRI-CARDAIX, a multimodal multi-agent framework that integrates deep learning-based segmentation with LLM-driven clinical reporting for comprehen-

sive cardiac MRI assessment. By integrating cardiac functional analysis and 17-segment fibrosis localization with evidence-grounded report generation, the system bridges image analysis and clinical decision support, enabling scalable, standardized, and interpretable cardiac diagnosis. In addition to segmentation and reporting accuracy, we evaluated the agentic layer directly, reporting 0.92 routing accuracy, a 1.00 task success rate, and sub-2-second mean latency, alongside retrieval metrics that isolate grounding quality from generation quality. Future work will address the retrieval bottleneck identified here by indexing guideline material separately from case reports and will strengthen intent disambiguation in the supervisor agent.

Acknowledgments. This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Project FSER-2025-0009), which funded the preparation of MRI data, training, and testing of the neural network pipeline for 17-segment segmentation. The development and testing of the agentic system and clinical report generation were supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2024-00460621, Developing BCI-based digital health technologies for mental illness and pain management).

Disclosure of Interests. Not applicable.

References

[1] W. H. Organization and W. H. Organization, "Cardiovascular diseases (CVDS). 2021," ed, 2021.

[2] A. H. A. W. G. o. M. Segmentation *et al.*, "Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: a statement for healthcare professionals from the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association," *Circulation,* vol. 105, no. 4, pp. 539–542, 2002.

[3] I. Circle Cardiovascular Imaging, "cvi42: Cardiovascular MR & CT Analysis Software," ed: Circle Cardiovascular Imaging Inc., 2024.

[4] P. M. Imaging, "CAAS MR: Comprehensive analysis for cardiac MRI," ed: Pie Medical Imaging, 2024.

[5] C. Chen *et al.*, "Deep learning for cardiac image segmentation: a review," *Frontiers in cardiovascular medicine,* vol. 7, p. 508599, 2020.

[6] J. Abbas, D. B. Soomro, S. Huang, and L. Liu, "DualAttendMed: A Coarse-to-Fine Dual-Stage Attention Framework for Interpretable Disease Localization and Classification," *Expert Systems with Applications,* p. 130886, 2025.

[7] G. Simantiris and G. Tziritas, "Cardiac MRI segmentation with a dilated CNN incorporating domain-specific constraints," *IEEE Journal of Selected Topics in Signal Processing,* vol. 14, no. 6, pp. 1235–1243, 2020.

[8] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, "Automatic cardiac disease assessment on cine-MRI via time-series

segmentation and domain specific features," in *International workshop on statistical atlases and computational models of the heart*, 2017: Springer, pp. 120–129.

[9] F. Zabihollahy, M. Rajchl, J. A. White, and E. Ukwatta, "Fully automated segmentation of left ventricular scar from 3D late gadolinium enhancement magnetic resonance imaging using a cascaded multi-planar U-Net (CMPU-Net)," *Medical physics,* vol. 47, no. 4, pp. 1645–1655, 2020.

[10] W. Al-Haidri *et al.*, "Quantitative analysis of myocardial fibrosis using a deep learning-based framework applied to the 17-Segment model," *Biomedical Signal Processing and Control,* vol. 105, p. 107555, 2025.

[11] W. Al-Haidri, A. Levchuk, K. Belousova, V. Fokin, D. Bendahan, and E. Brui, "Deep Learning-Assisted 17-Segment Myocardial Scar Segmentation and Analysis," in *2025 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, 2025: IEEE, pp. 333–336.

[12] J. Feng *et al.*, "M 3 builder: A multi-agent system for automated machine learning in medical imaging," in *International Workshop on Agentic AI for Medicine*, 2025: Springer, pp. 115–124.

[13] W. Zhang *et al.*, "Multi-agent reasoning for cardiovascular imaging phenotype analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2025: Springer, pp. 429–439.

[14] D. Gu, Y. Gao, Y. Zhou, M. Zhou, and D. Metaxas, "Radalign: Advancing radiology report generation with vision-language concept alignment," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2025: Springer, pp. 484–494.

[15] M. Z. Fang *et al.*, "Cardiac magnetic resonance imaging-large language model Meta AI: a finetuned large language model for identifying findings and associated attributes in cardiac magnetic resonance imaging reports," *Journal of Cardiovascular Magnetic Resonance,* vol. 27, no. 2, 2025.

[16] O. Oktay *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999,* 2018.

[17] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods,* vol. 18, no. 2, pp. 203–211, 2021.

[18] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv preprint arXiv:2401.04722,* 2024.

[19] J. Liu *et al.*, "Swin-umamba: Mamba-based unet with imagenet-based pretraining," in *International conference on medical image computing and computer-assisted intervention*, 2024: Springer, pp. 615–625.

[20] O. Bernard *et al.*, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?," *IEEE transactions on medical imaging,* vol. 37, no. 11, pp. 2514–2525, 2018.

[21] A. Lalande *et al.*, "Emidec: a database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac MRI," *Data,* vol. 5, no. 4, p. 89, 2020.