 https://doi.org/10.24191/jeesr.v28i1.012

Two-Stage Deep Learning Framework for Myocardial Infarction Segmentation in Late-Gadolinium Enhancement Magnetic Resonance Imaging Images

Nor Afnan Zharif Nor Kamal, Muhammad Khusairi Osman*, Dayang Suhaida Awang Damit, Siti Noraini Sulaiman, Nur 'Ulya Nasuha Zakaria, Khairul Azman Ahmad, Noor Khairiah A. Karim

***Abstract*— Accurate segmentation of myocardial structures and infarct regions in late-gadolinium enhancement magnetic resonance imaging (LGE-MRI) is essential for diagnosing ischemic heart disease (IHD). However, traditional and single-stage deep learning (DL) methods struggle with small or low-contrast regions such as myocardial scars. This study proposes a two-stage DL framework to address these limitations. Stage 1 segments the LV cavity using DeepLabv3+ (ResNet50), and Stage 2 segments the myocardium and scar using DeepLabv3+ (Xception). The framework was developed through four phases: baseline evaluation, loss and optimizer exploration, two-stage pipeline integration, and final validation with post-processing. Both models were trained using Dice loss and Adam optimizer. Final testing showed high segmentation performance for the LV cavity (Dice = 0.947) and myocardium (Dice = 0.7351). Scar segmentation remained challenging (Dice = 0.0556) due to small size and low contrast. Nonetheless, the modular design enhanced anatomical accuracy and reduced inter-class misclassification, demonstrating its potential for clinical cardiac image analysis.**

***Index Terms*— cardiac imaging, deep learning, LGE-MRI, myocardial infarction, segmentation, two-stage framework.**

I. INTRODUCTION

The human heart is a vital organ responsible for delivering oxygenated blood throughout the body via rhythmic contraction and relaxation of its muscular walls. Among its chamber lies the left ventricle (LV), the most muscular and pressurized chamber, which plays a key role in systemic circulation, therefore its structural and integrity is vital. Any impairment to the structure or function of the LV—such as myocardial thinning, fibrosis, or ischemia—can severely compromise cardiac output and patient survival. Among all cardiovascular disorders, ischemic heart disease (IHD) remains the leading cause of death globally [1]. Recent estimates, IHD affects over 126 million individuals globally and contributing to over 9 million deaths annually [1], [2]. IHD, also referred to as coronary artery disease (CAD), occurs when reduced blood flow to the myocardium, often due to coronary artery narrowing or blockage, leads to oxygen deprivation, tissue damage, and, in severe cases, myocardial infarction (MI) [2]. As the incidence of IHD continues to rise, effective diagnostic and prognostic tools are essential for improving outcomes.

Late Gadolinium Enhancement Magnetic Resonance Imaging (LGE-MRI) is widely recognized as the clinical gold standard imaging modality for visualizing myocardial scarring, offering non-invasive, high-resolution insights into infarction territory [3], [4]. Unlike other imaging modality techniques such as echocardiography, computed tomography (CT), or positron emission tomography (PET), LGE-MRI offers superior spatial resolution and contrast differentiation without ionizing radiation [4]. The procedure involves intravenous injection of gadolinium-based contrast agents, which circulate through the vascular system and gradually accumulate in fibrotic or infarcted myocardial tissue due to increased extracellular space. In the delayed phase, typically 10–20 minutes after administration, gadolinium is washed out from healthy myocardium while persisting in damaged regions, appearing hyperintense on MRI scans [3]. This selective retention enables accurate visualization of scar tissue. However, despite its advantages, manual segmentation of LV structures and infarcted regions in LGE-MRI images remains a challenging due to labour-intensive, time-consuming task that is prone to variability, and subject to inter-observer variability [5]. The segmentation process becomes even more complex, particularly to clinical settings with diverse patient anatomy, low signal-to-noise ratio in LGE-MRI scans (causing low tissue contrast), motion artifacts, and most notably, the relatively small spatial region occupied by scar tissue further complicates accuracy, emphasizing the need for more reliable and automated solutions [5], [6].

This manuscript is submitted on August 4, 2025, revised on October 2, 2025, accepted on October 23, 2025, and published on April 30, 2026.

Nor Afnan Zharif Nor Kamal, Muhammad Khusairi Osman, Dayang Suhaida Awang Damit, Siti Noraini Sulaiman, Nur 'Ulya Nasuha Zakaria and Khairul Azman Ahmad are with Electrical Engineering Studies, Universiti Teknologi MARA Cawangan Pulau Pinang, Permatang Pauh Campus, Seberang Perai, Pulau Pinang, Malaysia (email : dayang671@uitm.edu.my)

Noor Khairiah A. Karim is with Department of Biomedical Imaging, Advanced Medical and Dental Institute, Universiti Sains Malaysia, Kepala Batas, Pulau Pinang, Malaysia (email: drkhairiah@usm.my)

*Corresponding author

Email address: khusairi@uitm.edu.my

II. LITERATURE REVIEW

Conventional image processing and machine learning techniques have long been employed in cardiac segmentation tasks, particularly before the widespread adoption of deep learning. These methods typically rely on domain-specific heuristics such as boundary detection, morphological priors, and hand-crafted texture descriptors. While many approaches have reported moderate success in segmenting the left ventricular (LV) myocardium in LGE-MRI, their performance often declines significantly when applied to infarcted or fibrotic regions, where boundaries are less distinct and anatomical variation is more pronounced.

Albà et al. [7] proposed a graph-cut-based segmentation framework for LGE-MRI using shape constraints and inter-slice consistency to guide myocardial boundary extraction. Although the method achieved a mean Dice coefficient of 0.81, its reliance on anatomical regularity reduced robustness in patients with structural abnormalities. Similarly, Kurzendorfer et al. [8] introduced a polar-space minimal cost path algorithm leveraging texture features, which improved segmentation in mid-ventricular slices but struggled at the apex and base due to reduced contrast and anatomical ambiguity. These methods, while computationally efficient, are highly dependent on consistent image texture and geometry, limiting their generalizability in heterogeneous clinical datasets.

Wei et al. [9] explored the use of cine-MRI as a reference for segmenting myocardium in LGE scans via 3D mesh deformation across registered slices. Despite high segmentation accuracy (Dice ≈ 0.94), the technique required multi-modal image acquisition and precise alignment, which can be impractical in clinical settings and sensitive to motion artifacts. Later, Kurzendorfer et al. [10] advanced a random forest-based pipeline incorporating steerable texture filters and dynamic boundary optimization. While the method achieved solid performance (~0.83 Dice), it showed limitations in discriminating scar tissue, particularly in low-contrast regions or when the scar was small and indistinct from the blood pool (cavity).

To mitigate these limitations, de la Rosa et al. [11] developed a hybrid method combining top-hat morphological transforms for coarse segmentation and a lightweight artificial neural network (ANN) ensemble for refinement. Although more flexible than purely hand-crafted methods, the approach still required extensive pre-processing and was sensitive to image quality and operator-dependent parameters.

Collectively, these studies demonstrate that while conventional methods—such as graph-cut segmentation, mesh deformation, and random forest classifiers—can achieve reasonable accuracy in segmenting large anatomical structures like the myocardium, they consistently struggle with infarct scar detection. This limitation stems from their reliance on hand-crafted features, rigid anatomical priors, and modality-dependent assumptions, which fail to generalize across diverse patient anatomies and imaging conditions. Scar regions in LGE-MRI are typically subtle, irregular, and low in contrast, making them difficult to capture using static geometric or texture-based models. These challenges have motivated a transition toward deep learning-based methods, which offer automatic feature extraction, contextual understanding, and end-to-end optimization—critical capabilities for accurately segmenting small, heterogeneous structures such as myocardial scars.

In response to these limitations, many recent studies have adopted deep learning architectures for cardiac segmentation in LGE-MRI. These models are capable of learning hierarchical and contextual representations directly from data, offering the promise of full automation and improved generalizability. For example, Popescu et al. [12] proposed ACSNet—a CNN that simultaneously segments the LV cavity, myocardium, and scar using anatomy-guided learning. While their network achieved high accuracy for the cavity and myocardium, it reported only moderate performance for scar segmentation (Dice ≈ 0.57), with notable failures in detecting small or low-contrast scar regions. Similarly, Yue et al. [13] employed a shape-regularized spatial constraint network (SRSCN) for myocardium segmentation, achieving Dice scores near inter-observer variation (≈ 0.75), yet scar segmentation was not addressed, underscoring the difficulty of multi-class tasks in single-pass models.

Other efforts have focused specifically on scar segmentation. Zabihollahy et al. [14] utilized a 3D CNN to segment scars in pre-segmented myocardium masks, reporting high Dice values (> 0.93). However, their model relied on manually provided myocardium contours, circumventing the very challenge of fully automated multi-class segmentation. Suwa [15]. attempted to fuse LGE and cine-MRI in a two-channel CNN, achieving good performance in slice-wise scar detection (success rate ≈ 95%) but at the cost of modality dependence and added acquisition complexity. Meanwhile, Righetti et al. [16] developed a CNN for binary classification of scar presence using cine-MRI, achieving high sensitivity but providing no spatial segmentation or localization capability. Despite improvements in representation learning, these single-stage deep learning models still exhibit critical weaknesses. Most notably, they struggle with scar detection, due to several interrelated challenges:

1. ***Small region size:*** Scar regions often occupy a small spatial footprint, making them underrepresented during training.
2. ***Visual ambiguity:*** Scar tissue may have intensity values similar to the blood pool or adjacent myocardium, leading to high misclassification.
3. ***Inter-class dependencies:*** In single-stage models, poor segmentation of one structure—particularly the LV cavity—can propagate errors to surrounding regions like the myocardium or scar. This phenomenon has been consistently observed in segmentation errors clustered around structural boundaries.

These challenges parallel those found in conventional approaches, where dependency on global shape assumptions, handcrafted features, or external modalities often failed to capture the nuanced anatomy present in pathological LGE-MRI

scans. As such, there remains a gap in segmentation quality—particularly in multi-class, structure-aware segmentation pipelines.

To overcome the limitations of conventional and single-stage deep learning methods—particularly in accurately segmenting small, low-contrast scar regions—this study proposes a two-stage deep learning framework. The first stage segments the LV cavity, which typically has clearer anatomical boundaries, while the second stage performs myocardium and scar segmentation guided by the cavity mask. This decomposition allows for modular optimization, where each stage is trained independently and focuses on its respective anatomical target. By introducing the cavity as a spatial prior, the model reduces misclassification at structural boundaries and enhances robustness in anatomically variable or noisy images. As part of this solution, the study also investigates model performance across five deep learning architectures, explores the influence of different loss functions and optimizers, and integrates post-processing techniques to refine final predictions. Collectively, these steps support the overarching objective: to design and evaluate a two-stage deep learning framework that improves automated segmentation of LV structures and myocardial scarring in LGE-MRI through architectural modularity, loss and optimizer tuning, and post-processing enhancement.

III. Methodology

The methodology of this study is structured into a multi-phase pipeline designed to develop, optimize, and evaluate a two-stage deep learning framework for myocardial infarction segmentation in LGE-MRI images. As illustrated in the Fig. 1, the flowchart process begins with data collection and pre-processing to prepare annotated images suitable for training. The segmentation framework is then developed in two separate stages: the first stage segments the LV cavity, and the second stage focuses on myocardium and scar regions.

In Phase 1, all candidate models are trained using default settings to evaluate baseline performance. Based on these results, the best of 2 performing models proceed to Phase 2, where different combinations of loss functions and optimizers are explored. The top configuration for each stage is then integrated in Phase 3 into a unified two-stage deep learning framework, with a cavity-guided post-processing step added between stages. Finally, Phase 4 evaluates the complete system on the segmentation performance of the final model both in term of qualitative and quantitative.

A. Data Collection

Data collection is a critical first step in medical image analysis, involving the acquisition and selection of clinically relevant images that accurately represent the anatomical structures and pathologies under study. In this research, cardiac LGE-MRI data were collected from the Advanced Medical and Dental Institute (AMDI), Universiti Sains Malaysia (USM), Pulau Pinang, Malaysia using a Philips Achieva 3.0 T scanner. The initial dataset comprised 1,497 short-axis (SAX) slices from 41 patients aged 12 to 76, including both normal and pathological cases. Images were provided in DICOM format

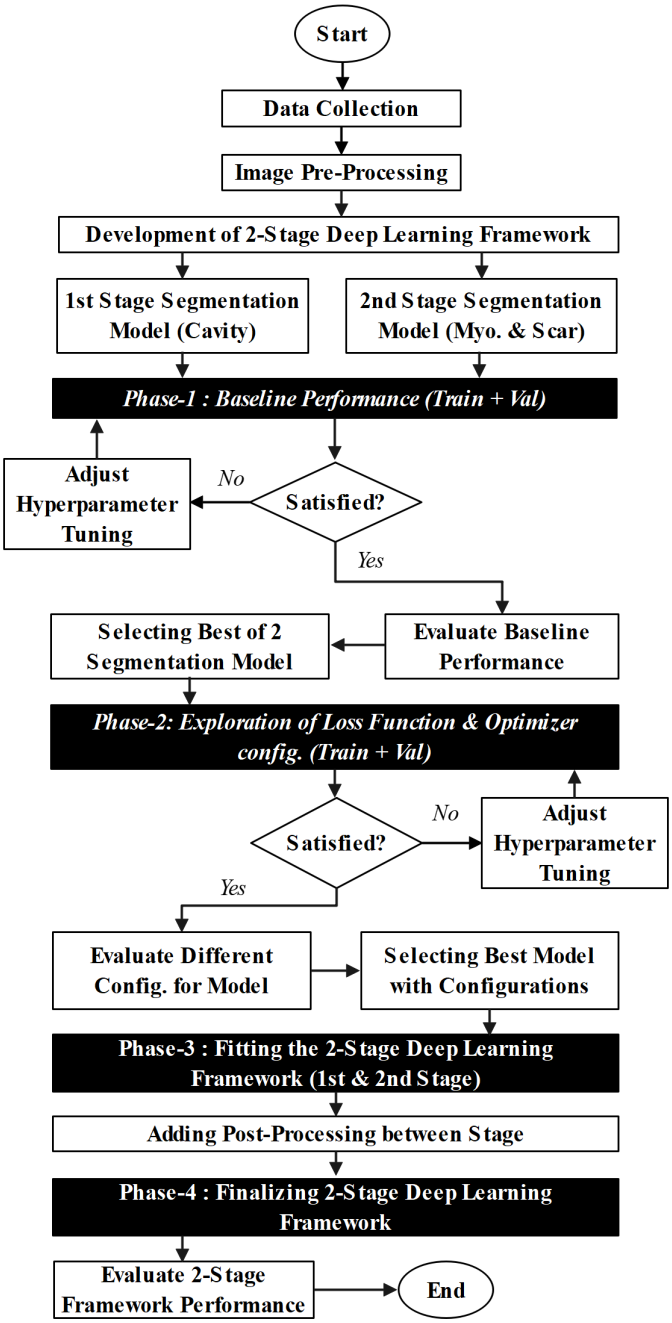


Fig. 1. Development of 2-stage deep learning frameworks for MI segmentation.

and converted to bitmap (.bmp) to preserve intensity fidelity for segmentation tasks. For each patient, 11 mid-ventricular slices were selected because myocardial scars mostly appear in this region, making it the most reliable and clinically useful for analysis. Basal and apical slices were excluded due to minimal scar presence and poor image quality. By focusing on mid-ventricular slices, the most informative regions were captured, thereby enhancing model training consistency and clinical relevance. This yielded 236 high-quality images, representing both healthy and scarred myocardium for model development. Fig. 2 shows the example of curated dataset with its corresponding ground truth (GT).

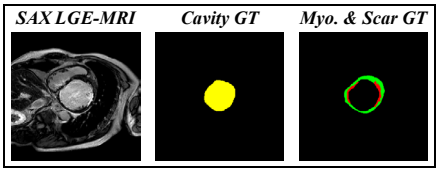


Fig. 2. Multi-class cardiac segmentation dataset specifically on LV region.

B. Image Pre-processing

1) Image Resizing

All LGE-MRI images and their corresponding ground truth masks were resized to a standardized resolution of 320×320 pixels. This ensures compatibility across all network input layers and improves computational efficiency during training while preserving anatomical structure.

2) Data partitioning

The curated dataset of 236 LGE-MRI images was partitioned using a stratified split ratio of 60% training, 15% validation, and 25% testing. Details are provided in Table I.

3) Data Augmentation

TABLE I. DATA PARTITIONING.

Types of Datasets	Data-split Ratio	No of images per classes
Training	60% (140 images)	40 normal, 100 scar
Validation	15% (38 images)	20 normal, 18 scar
Testing	25% (58 images)	20 normal, 38 scar

To improve generalization and reduce overfitting, rotation-based augmentation was applied to the training set. Each image was randomly rotated by 90°, 180°, or 270°, simulating orientation variability during model training. The augmentation was limited to rotation in order to preserve the anatomical structure of the myocardium, cavity, and scar regions.

C. Development of 2-Stage DL Framework

The development of this framework uses and evaluate five deep learning models: UNet, SegNet, and DeepLabv3+ with three different backbones—MobileNetV2, ResNet50, and Xception. Each model was adapted for multi-class segmentation and trained independently for both stages. This unified model setup allowed consistent comparison across architectures while enabling modular optimization of segmentation tasks tailored to anatomical structure complexity.

D. 1st Stage Segmentation Model for Cavity Detection

The first stage of the proposed framework focuses on segmenting the left ventricular (LV) cavity from LGE-MRI images. The cavity typically exhibits clear anatomical boundaries and strong contrast against surrounding myocardium, making it a suitable initial target for guiding subsequent segmentation. All five deep learning models were trained to perform multi-class segmentation including LV cavity. The objective is to produce accurate and stable cavity predictions that serve not only for evaluation but also as spatial priors for guiding the second-stage segmentation. Accurate cavity detection is essential as misclassification at this stage may propagate errors into the downstream myocardium and scar segmentation.

E. 2nd Stage Segmentation Model for Myo. & Scar Detection

The second stage performs segmentation of the LV myocardium and infarcted scar regions, which are more complex due to their irregular shape, small spatial footprint, and low contrast against adjacent tissues. The output from Stage 1—specifically the LV cavity mask—is incorporated into Stage 2 as additional input or guidance to improve boundary definition and inter-class separation. The same five models were retrained for this stage using updated labels that include background, myocardium, and scar. By separating this task from cavity detection, the second-stage models able to specialize in learning finer tissue-level distinctions. This modular approach is particularly beneficial for improving scar detection accuracy, which is typically poor in single-stage segmentation due to its inter-class dependency and limited spatial representation.

F. Phase 1—Baseline Performance

The initial performance of all five deep learning models was assessed in Phase 1 to establish a baseline for each segmentation stage. At this stage, models were trained based on training options (default) specified in Table II, without any modifications to class imbalance, loss functions or optimizers, allowing for a direct comparison under default training conditions. The evaluation was performed separately for Stage 1 (cavity segmentation) and Stage 2 (myocardium and scar segmentation) using the fixed training configuration. Various performance metrics—including Dice coefficient, Intersection-over-Union (IoU), F1-score, Boundary F1-score (BF-score), sensitivity, and specificity—were calculated for each class. These metrics were used to assess segmentation quality and identify models that demonstrated strong baseline performance across classes. This phase provided a reference point for further optimization in subsequent phases.

G. Selection Best of 2 Model Segmentation

Referring from the Phase 1 baseline evaluation results, the top two models for each stage were selected by aggregate or calculate the mean from global performance metrics. This scoring also considered per-class metrics, particularly emphasizing the ability to segment smaller or clinically critical regions such as the scar. Visualizations such as training curves, Dice distribution boxplots and heatmaps were generated to support model comparison and ranking. The two highest-performing models for Stage 1 (cavity) and Stage 2 (myocardium and scar) were then shortlisted for further optimization in Phase 2, where the effects of different loss functions and optimizers would be evaluated. This selection process ensured that only the most promising architectures progressed to the fine-tuning phase (Phase 2).

H. Phase 2—Analysis of Loss Function and Optimizer Parameters

In Phase 2, the top two models selected from each stage were further refined by evaluating the effects of different loss function and optimizer configurations on segmentation performance. This phase aimed to identify the most effective combination of training objectives and optimization strategies for each segmentation task. Five loss functions were investigated: weighted cross-entropy (CE), Dice loss, Focal loss, and two hybrid combinations—Dice + CE and Dice + Focal—each designed to handle class imbalance and enhance boundary sensitivity. Additionally, two optimizers were compared: Stochastic Gradient Descent with Momentum (SGDM) and Adam.

Each model was trained with hyperparameters specified as in Table II for both optimizers, across all ten combinations of loss function (5 losses × 2 optimizers), and their performances were evaluated using the same metrics as in Phase 1. The results were visualized through training curves, boxplots, violin plots and standard deviation bar plot of global metrics for a more comprehensive view of model behaviour across configurations. A weighted scoring system was again used to rank configurations, emphasizing consistency, accuracy, and stability across all classes. This phase ultimately identified the best-performing model–loss–optimizer configuration for each stage, which would then be integrated into the final two-stage segmentation framework.

TABLE II. MODEL'S HYPERPARAMETERS.

Hyperparameter	Types/Value
Optimizer	Adam
Initial learning rate, α	10^{-4}
Maximum epoch	30
Mini-batch size	8
Validation frequency	50
Validation patience	3
Final network generalization	best-validation

I. Selection of Best Models Configuration

Following the evaluation in Phase 2, the best-performing configuration for each stage was determined based on overall consistency and weighted metric scores as in equation (1),

$$Final\ Score = \sum_{k=1}^{M} w_k \cdot \widehat{m_k} \tag{1}$$

where, $\widehat{m_k}$ is the average value of metric k, w_k is the assigned weight for each metric (refer Table III), and M is the number of metrics considered. The highest-scoring configuration per model was selected to advance to the two-stage framework.

This included identifying the most effective model architecture, loss function, and optimizer combination. For Stage 1, the model that demonstrated superior cavity segmentation accuracy and stability across multiple metrics was selected. Similarly, the top configuration for Stage 2 was chosen based on its performance in segmenting both myocardium and scar, with particular emphasis on Dice score for scar regions. The selection was supported by comparative plots mentioned in Section H and the finalized configurations were then carried forward for full integration into the proposed 2-stage DL framework.

TABLE III. WEIGHTED SCORING SYSTEM.

Metrics (Global)	Assign weighted value
Dice	0.25
IoU	0.20
F1-Score	0.20
BF-Score	0.10
Sensitivity	0.15
Specificity	0.10

J. Phase 3— Fitting the 2-Stage DL Framework

In this phase, the best model from each stage was combined into a complete two-stage DL framework. The first-stage model, optimized for LV cavity segmentation, was used to generate a cavity mask for each input image. This mask was then passed into the second-stage model—along with the original LGE-MRI image—as an additional input channel to guide myocardium and scar segmentation. This is the part of the post-processing composite output.

K. Post-Processing

To enhance segmentation precision and reduce anatomical misclassification, a post-processing step was implemented between Stage 1 and Stage 2 of the framework (see Fig. 3 below). The primary goal was to refine the cavity prediction output from the first-stage model before it was used to guide the second-stage segmentation. This process involved two key steps: morphological cleaning and cavity masking.

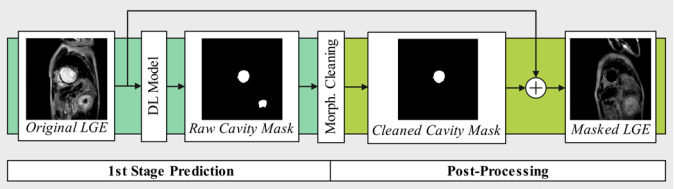


Fig. 3. Post-processing flow, the cleaned cavity mask is created on the original LGE-MRI image after undergoing morphological cleaning process.

1) Morphological Operations

First, these operations were applied to the predicted cavity mask to eliminate noise and remove small false-positive blobs. These operations ensured a cleaner and more anatomically plausible cavity shape. The morphological steps are presented in Fig. 4.

TABLE IV. AGGREGATE SCORE FOR 1ST STAGE BASELINE PERFORMANCE & MODEL CONFIGS. EXPLORATION.

	Model	Dice	IoU	F1-Score	BF-Score	Sensitivity	Specificity	Mean Score	Rank
Phase 1	Deeplabv3+_MobilenetV2	0.9645	0.9391	0.9645	0.9344	0.9632	0.9632	0.9548	1
	Deeplabv3+_ResNet50	0.9423	0.9083	0.9424	0.8643	0.9329	0.9329	0.9205	2
	Deeplav3+_Xception	0.9151	0.8823	0.9151	0.8577	0.8964	0.8964	0.8939	3
	UNet	0.9125	0.8520	0.9125	0.6519	0.94070	0.9407	0.8684	4
	SegNet	0.4864	0.3845	0.4864	0.3677	0.8371	0.8371	0.5665	5
Phase 2	Deeplabv3+_ResNet50 + Dice + Adam	0.4735	0.4507	0.4735	-	0.4849	0.9987	0.5287	*Best*
	Deeplabv3+_ResNet50 + (Dice + CE) + Adam	0.4590	0.4265	0.4590	-	0.4912	0.9978	0.5170	-

TABLE V. AGGREGATE SCORE FOR 2ND STAGE BASELINE PERFORMANCE & MODEL CONFIGS. EXPLORATION.

	Model	Dice	IoU	F1-Score	BF-Score	Sensitivity	Specificity	Mean Score	Rank
Phase 1	UNet	0.3260	0.3190	0.3260	-	0.3288	0.6809	0.3961	1
	Deeplav3+_Xception	0.3261	0.3192	0.3261	-	0.3294	0.6679	0.3938	2
	Deeplabv3+_MobilenetV2	0.3260	0.3191	0.3260	-	0.3293	0.6680	0.3937	3
	Deeplabv3+_ResNet50	0.3258	0.3188	0.3259	-	0.3290	0.6681	0.3935	4
	SegNet	0.2670	0.2241	0.2670	-	0.2254	0.8387	0.3644	5
Phase 2	Deeplabv3+_Xception + Dice + Adam	0.6120	0.5292	0.5953	0.7508	0.6111	0.9011	0.6348	*Best*
	Deeplabv3+_Xception + (Dice + CE) + Adam	0.6056	0.5266	0.5943	0.7143	0.6583	0.9387	0.6396	-

Morphological Refinement & Composite Generation

Input: $\boldsymbol{I_{LGE}}$ (original image), $\boldsymbol{M_{pred}}$ (RGB cavity masks)
Output: $\boldsymbol{C_{out}}$ (3-channel composite images)

For each $\boldsymbol{i} \in [\mathbf{1}, \boldsymbol{N}]$:

$I \leftarrow \text{grayscale}(I_{LGE}(i)),\ M \leftarrow \text{binarize}(M_{pred}(i) = [255,255,0])$
$M_f \leftarrow \text{imfill}(M, \text{'holes'})$
$M_c \leftarrow \text{largestCC}(M_f)$
$M_r \leftarrow \text{imclose}\left(imopen(M_c, SE_{disk}(2)), SE_{disk}(2)\right)$
$M_r \leftarrow \vee \text{dilate}(perim(M_r), SE_{disk}(1))$
$I_m \leftarrow \text{I};\ I_m(M_r) = 0$
$C_{out}(i) = cat(3,\ I_m,\ 255 \times M_r,\ I)$
$save(C_{out}(i))$

Fig. 4. Each predicted cavity mask is binarized, hole-filled, and cleaned by retaining only the largest component. Morphological operations smooth boundaries and reinforce edges. The refine mask zero-masks the LGE image, producing a 3-channel composite $\{I_m, 255M_r, I\}$ for subsequent processing.

2) Cavity Masking

The output from Stage 1 is a cavity mask represented as a single-channel binary image (foreground = LV cavity, background = non-cavity). To ensure compatibility with Stage 2, the refined cavity mask is mapped onto the original LGE-MRI slice, resulting in a three-channel input image in which the LV cavity region is effectively removed. This composite image is then used as the input for the second-stage model. This strategy significantly reduced false positives in scar detection and improved segmentation consistency, particularly around class boundaries affected by inter-class dependencies.

L. Finalizing 2-Stage DL Framework

The final stage of the methodology involved integrating all components optimized models, post-processing, and inter-stage data flow into a complete two-stage deep learning segmentation framework. The best-performing model for Stage 1 was used to segment the LV cavity, and its output was refined through morphological post-processing. Then, the post-processing output is fed into the optimized Stage 2 model for myocardium and scar segmentation. This design enabled a guided segmentation process where Stage 2 benefited from anatomical priors generated in Stage 1, helping resolve ambiguities near class boundaries and enhancing detection of small scar regions. With all components modularized and optimized, the finalized framework was ready for comprehensive evaluation in Phase 4, including both quantitative performance metrics and qualitative visual assessment on same dataset.

This entire framework was implemented using MATLAB's Deep Learning Toolbox and executed on a PC equipped with an NVIDIA GeForce RTX 3060 GPU (12GB) to support efficient training and inference.

M. Framework Evaluation

This section briefly explains on the used performance metrics across Phase 1, 2 and 4. The six-evaluation metrics provide comprehensive evaluation of spatial overlap, boundary precision, and classification accuracy for each anatomical class (cavity, myocardium, and scar).

1) Intersection-Over-Union (IoU)

Also known as the Jaccard Index, IoU measures the overlap between predicted and ground truth regions and is widely used in segmentation challenges and provides a direct measure of region-level agreement. Referring to equation (2), P_c is a predicted binary mask for class c; T_c is the ground truth binary mask. TP_c is true positives, FP_c is false positives and FN_c is false negatives.

$$IoU_c = \frac{|P_c \cap T_c|}{|P_c \cup T_c|} = \frac{TP_c}{TP_c + FP_c + FN_c} \quad (2)$$

2) Dice Similarity Coefficient (DSC)

The Dice score is a spatial overlap index that is particularly useful for small regions, like infarct zones. Dice and IoU are closely related but Dice is more sensitive to small structures.

$$DSC_c = \frac{2 \cdot |P_c \cap T_c|}{|P_c| + |T_c|} = \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c} \quad (3)$$

3) F1-Score

Although mathematically equivalent to the Dice score in binary segmentation, the F1-score is often used in classification contexts to measure harmonic mean of precision (TP) and recall (FP + FN). Thus, for segmentation tasks, F1-score identical to Dice, but may be reported separately when cross-validating classification metrics.

$$F1_c = \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c} \quad (4)$$

4) Boundary F1-Score (BF Score)

The Boundary F1-Score evaluates how well the predicted segmentation matches the ground truth along the object boundary, which is especially important for fine edge delineation. This metric is useful for assessing edge alignment in tasks involving thin or irregular structures. ∂P_c is the predicted boundary for class c and ∂T_c is the ground truth boundary.

$$BF_c = 2 \cdot \frac{|\partial P_c \cap \partial T_c|}{|\partial P_c| + |\partial T_c|} \quad (5)$$

5) Sensitivity

Sensitivity measures the model's ability to correctly identify positive class pixels. A high sensitivity ensures that important pathological regions (e.g., scar) are not missed, which is critical in clinical contexts.

$$Sensitivity_c = \frac{TP_c}{TP_c + FN_c} \quad (6)$$

6) Specificity

Specificity measures the ability to correctly identify negative class pixels. It ensures that background regions are not falsely labelled as anatomical structures, improving the reliability of segmentation.

$$Specificity_c = \frac{TN_c}{TN_c + FP_c} \quad (7)$$

IV. Results and Discussion

This section presents the experimental results and analysis of the proposed two-stage deep learning framework for myocardial infarction segmentation in LGE-MRI images. The results are structured at each stages according to the four development phases outlined in the methodology: baseline model evaluation (Phase 1), optimization of loss function and optimizer configurations (Phase 2), integration into the two-stage segmentation framework (Phase 3), and final system performance with post-processing (Phase 4). Quantitative performance metrics as mentioned previously, are reported alongside qualitative visualization to assess model accuracy, robustness, and clinical relevance. Each phase is discussed in detail to highlight the contributions of architectural design, training strategies, and modular refinement to the overall performance improvement.

A. 1st Stage Segmentation Model Framework Evaluation Results

1) Phase 1— Baseline Performance Results

In Phase 1, five models were benchmarked to assess their baseline segmentation ability for LV cavity under consistent settings. According to Table IV, DeepLabv3+ with MobileNetV2 achieved the highest mean score of 0.9548, while ResNet50 followed at 0.9205. Although the Deeplabv3+ (MobileNetV2) led slightly in average performance, the ResNet50 backbone was selected as a finalist based on its more reliable cavity delineation observed during visual assessment.

The prediction masks generated by Deeplabv3+ (ResNet50) exhibited sharper and more anatomically accurate boundaries, particularly around the LV endocardial contours. While Deeplabv3+ (MobileNetV2) offered higher overlap metrics, its segmentations occasionally lacked structural precision—an important factor in clinical interpretation. Therefore, Deeplabv3+ (ResNet50)'s consistent boundary adherence and cleaner cavity shape motivated its selection for further optimization in Phase 2. Hence, it was selected as one of the top two candidates for further optimization in Phase 2 based on its robust structural precision.

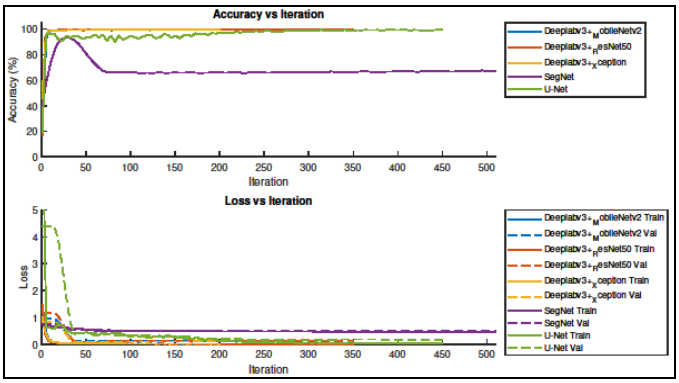


Fig. 5. Training curves of 5 deep learning models with best-validation and early stopping criteria implementation in measuring baseline performance for 1st stage segmentation model.

The training curves in Fig. 5 indicate that both Deeplabv3+ with MobileNetV2 and ResNet50 converged stably within 20–30 iterations, aided by early stopping criteria. However, the MobileNetV2 backbone achieved a slightly faster plateau, reflecting better initial convergence. The class-wise metrics in Fig. 6 show that Deeplabv3+ with both backbones performed well on the cavity and background classes, with MobileNetV2 backbone having a slight edge in Dice and IoU for the cavity. Fig. 7 (a), reveals that, the boxplot of MobileNetV2's Dice

scores had a narrower interquartile range (IQR), indicating greater consistency across test images. Similarly, Fig. 7 (b) shows that, the dice heatmap MobileNetV2's performance was more spatially uniform, whereas ResNet50 exhibited higher variability. Nonetheless, ResNet50's strengths in structural precision and specificity justified its continuation into the next phase.

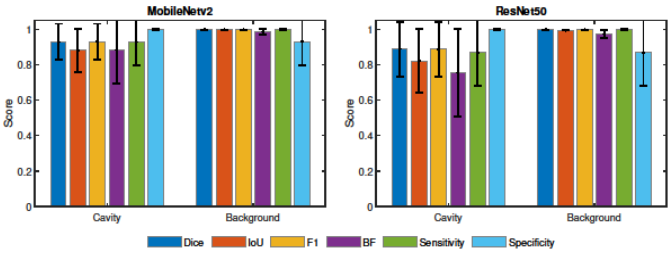


Fig. 6. Metric per-class bar plot shows for best of 2 model (Deeplabv3+) in 1st stage segmentation model with different backbone, MobileNetv2 (left) and ResNet50 (right).

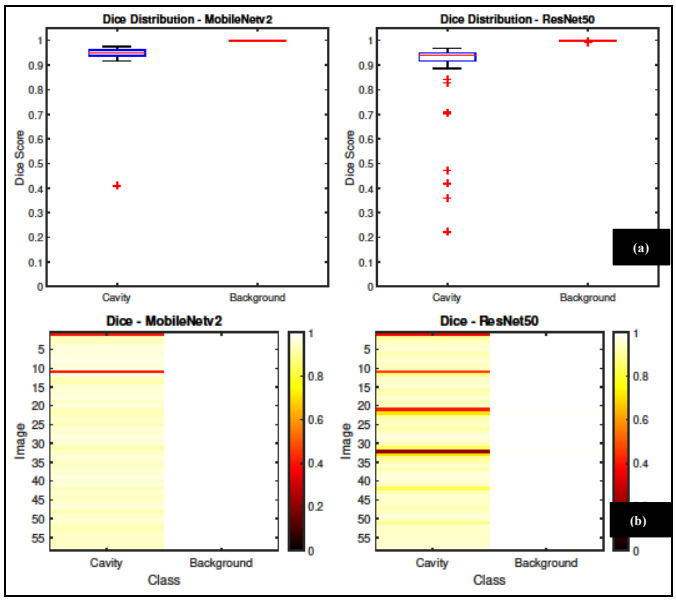


Fig. 7. Dice distribution per-class for MobileNetv2 (left) and ResNet50 (right), for improve qualitative analysis through visualization of, (a) Dice boxplot, and (b) Dice heatmap.

2) Phase 2— Loss & Optimizer Exploration Results

Phase 2 explored various combinations of loss functions and optimizers on the Deeplabv3+ (ResNet50) model. As seen in Fig. 8, the configurations trained with the Adam optimizer consistently demonstrated smoother and faster convergence compared to the SGDM optimizer. The SGDM-based configurations showed more oscillations and slower validation progress, suggesting less stable gradient updates. Notably, both Dice and Dice + CE with Adam optimizer showed the cleanest validation loss curves, validating their suitability for this stage.

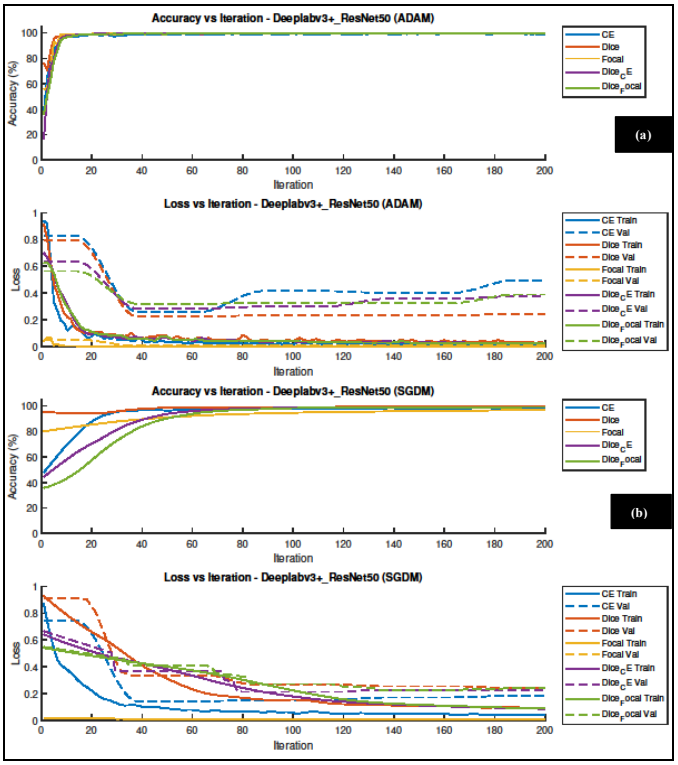


Fig. 8. Training curves of the best model, Deeplabv3+ (ResNet50) with 5 different loss function layers through 2 optimizers, (a) Adam and (b) SGDM.

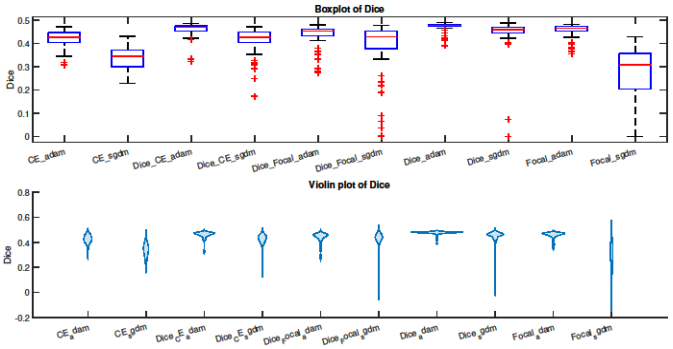


Fig. 9. Dice distribution for various Deeplabv3+ (ResNet50) configurations (loss layer + optimizer), using boxplot and violin plot.

Boxplots and violin plots in Fig. 9 reinforce this observation. Dice with Adam configuration displayed a tight distribution and higher median Dice and IoU scores, while SGDM-based configurations had wider spreads and flatter curves—indicating less reliable performance. Fig. 10, which summarizes global metric standard deviation, clearly highlights that Dice with Adam configuration as the most consistent configuration across metrics. While Dice + CE with Adam showed lower sensitivity, along with higher variability in specificity and boundary performance. Therefore, Deeplabv3+ (ResNet50) with Dice loss and Adam optimizer was selected as the best configuration due to its balanced accuracy, low variance, and reliable convergence behaviour.

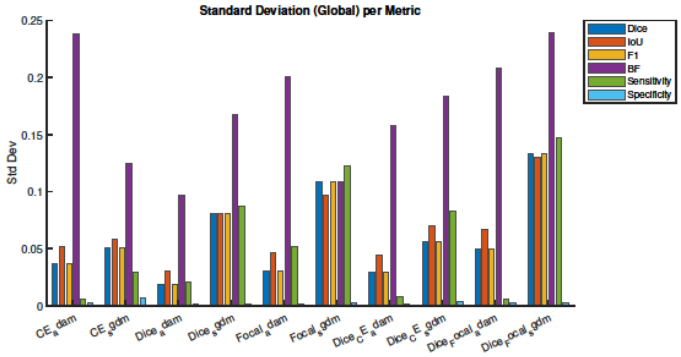


Fig. 10. Bar plot of standard deviation for global metrics per-model configurations of Deeplabv3+ (ResNet50) in presenting the consistency of model's generalization.

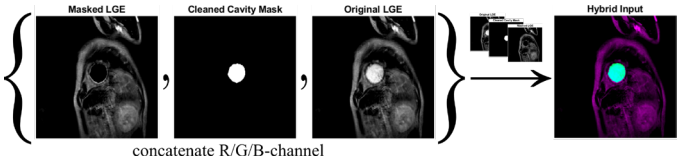


Fig. 11. Post-processing results: the hybrid input is the final process consisting composite of raw cavity, cleaned cavity and the original LGE-MRI arranged in RGB channel.

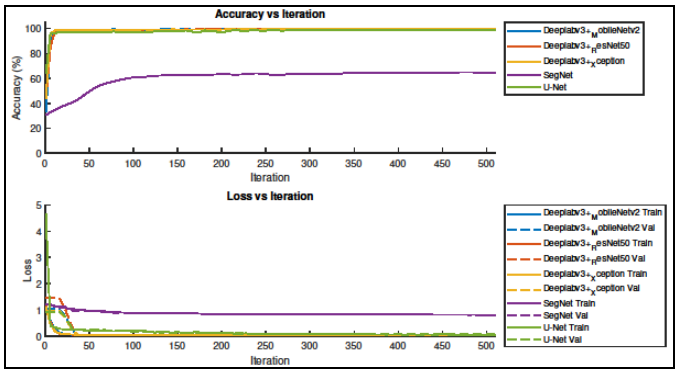


Fig. 12. Training curves of 5 deep learning model with same options in measuring baseline performance for 2nd stage segmentation model.

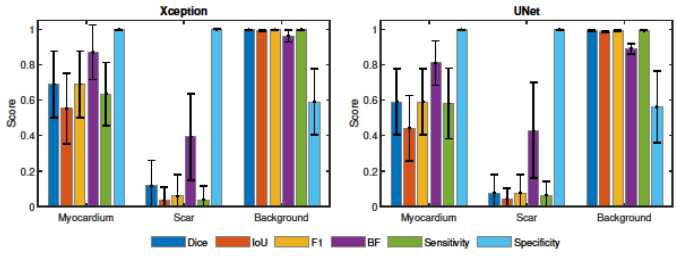


Fig. 13. Metric per-class bar plot shows for best of 2 model in 2nd stage segmentation model, Xception (left) and UNet (right).

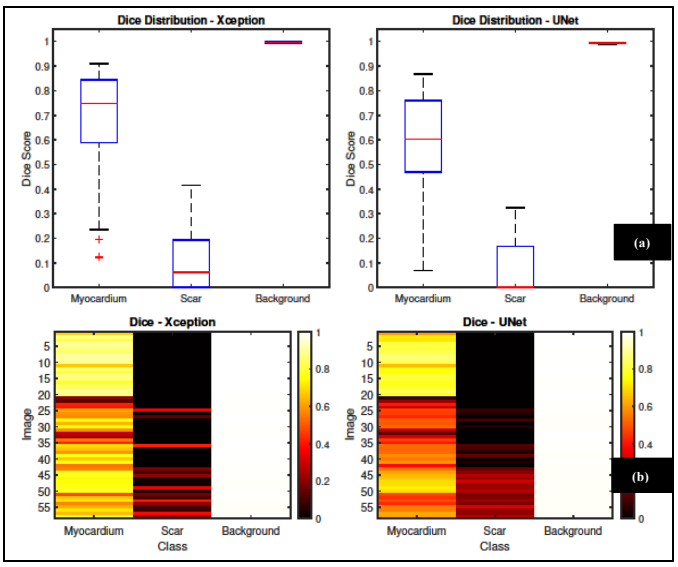


Fig. 14. Dice heatmap plot per-images & per-class for Deeplabv3+ (Xception) (left) and UNet (right), for improve qualitative analysis through, (a) Dice boxplot, and (b) Dice heatmap.

3) Phase 3— Fitting 1st Stage Segmentation Model

In Phase 3, the selected Deeplabv3+ (ResNet50) with Dice loss and Adam optimizer configurations was integrated into Stage 1 of the two-stage framework. The segmentation output of this stage was used to generate cavity masks, which were then morphologically processed and fed into Stage 2 as both anatomical guidance and additional input channels as shown Fig. 11. This modular design allowed Stage 1 to enforce spatial structure and minimize misclassification in later stages. The final performance of Stage 1 (Table VI) confirmed its robustness, with a Dice score (global) of 0.9727, IoU of 0.9490, and high sensitivity and specificity (0.9836)—further validating its role as the structural foundation of the full segmentation pipeline.

B. Analysis of the 2nd Stage Segmentation Model Framework

1) Phase 1— Baseline Performance

In Phase 1 of Stage 2, five models were evaluated to segment myocardium and scar tissue. As shown in Table V, the overall performance was lower than Stage 1, with the top mean scores only reaching 0.3961 (UNet) and DeepLabv3+ (Xception), that is 0.3938. Deeplabv3+ (Xception) led in four out of six metrics, with the highest Dice score (0.3261) along with others, showing promise in handling difficult myocardial structures. Although UNet performed competitively, Deeplabv3+ (Xception)'s superior results in sensitivity (0.3294) and stability across regions made it the preferred model for further exploration. The training loss curves in Fig. 12 show smoother and faster convergence for Deeplabv3+ (Xception) compared to UNet. The per-class metric comparison between these 2 models in Fig. 13 shows that Deeplabv3+ (Xception) outperform UNet in term

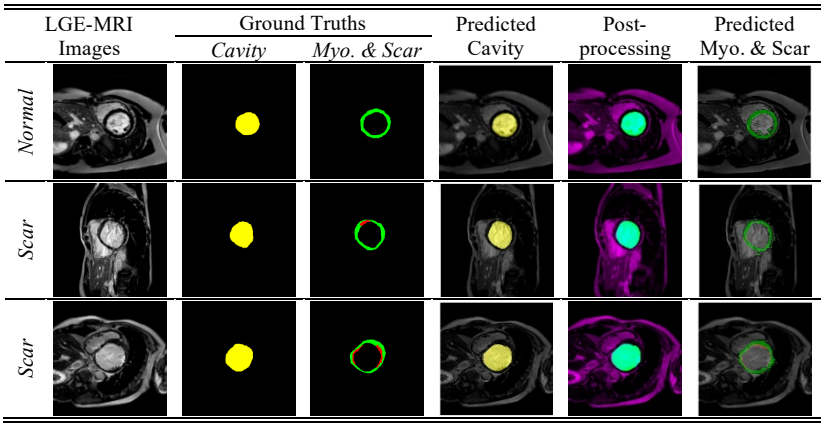
TABLE VI. 1ST STAGE METRICS PERFORMANCE FROM 2-STAGE DEEP LEARNING FRAMEWORK.

Class	Accuracy	Dice	IoU	F1-Score	BF-Score	Sensitivity	Specificity
Cavity	0.9764	0.9470	0.9013	0.9470	0.9107	0.9697	0.9974
Background	0.9974	0.9983	0.9967	0.9983	0.9879	0.9974	0.9697
Global Metrics							
Global Accuracy	Mean Accuracy	Dice	IoU	F1-Score	BF-Score	Sensitivity	Specificity
0.99677	0.9869	0.9727	0.9490	0.9727	0.9493	0.9836	0.9836

TABLE VII. 2ND STAGE METRICS PERFORMANCE FROM 2-STAGE DEEP LEARNING FRAMEWORK.

Class	Accuracy	Dice	IoU	F1-Score	BF-Score	Sensitivity	Specificity
Myocardium	0.7404	0.7351	0.5949	0.7351	0.9157	0.7230	0.9973
Scar	0.6585	0.0556	0.0306	0.0556	0.4804	0.2517	0.9654
Background	0.9637	0.9799	0.9608	0.9799	0.9277	0.9636	0.8101
Global Metrics							
Global Accuracy	Mean Accuracy	Dice	IoU	F1-Score	BF-Score	Sensitivity	Specificity
0.95988	0.7875	0.5902	0.5288	0.5902	0.7746	0.6461	0.9242

TABLE VIII. QUALITATIVE ANALYSIS OF 2-STAGE DEEP LEARNING FRAMEWORK.



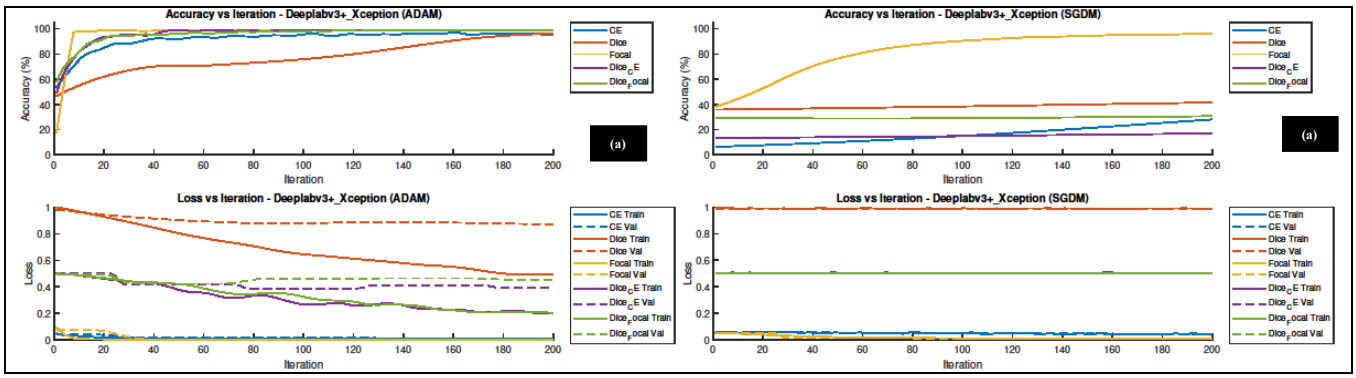


Fig. 15. Training curves of the best model, Deeplabv3+ (Xception) with 5 different loss function layers through 2 optimizers, (a) Adam and (b) SGDM.

of segmenting the myocardium class but almost identical performance in scar class, but qualitative evidence strongly support in Fig. 14(a) which confirms that Dice boxplot of Deeplabv3+ (Xception) achieved more concentrated performance with fewer outliers, while the Dice heatmap in Fig. 14(b) reveals that Deeplabv3+ (Xception) yielded more uniform segmentation across images. These observations confirmed Deeplabv3+ (Xception)'s ability to generalize better on anatomically variable LGE-MRI scans.

2) *Phase 2— Analysis of the Loss and Optimizer*

In Phase 2, DeepLabv3+ (Xception) was trained with multiple loss functions and both SGDM and Adam optimizers. As shown in Fig. 15, models trained with Adam optimizer again demonstrated faster and more stable convergence, while SGDM exhibited more fluctuation, especially in early epochs. Both Dice and Dice + CE with Adam configurations showed early validation and maintained smooth loss progression.

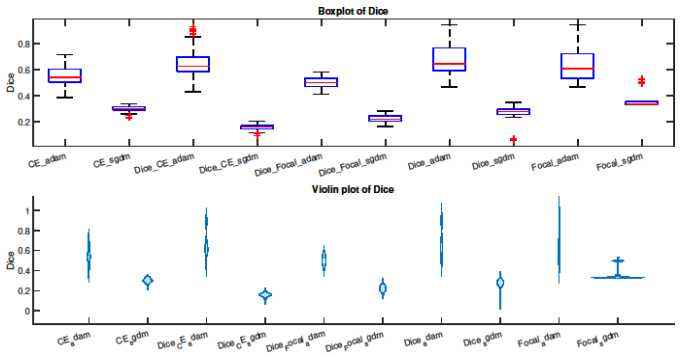


Fig. 16. Dice distribution for various Deeplabv3+ (Xception) configurations (loss layer + optimizer), using boxplot and violin plot.

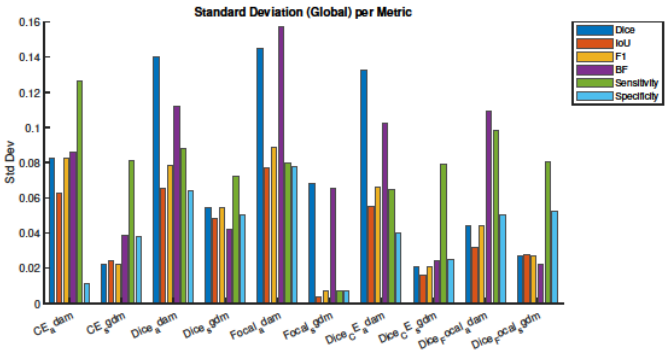


Fig. 17. Bar plot of standard deviation for global metrics per-model configurations of Deeplabv3+ (Xception) in presenting model's consistency.

Fig. 16 further supports this discussion, with Dice and Adam configuration displaying a tight boxplot and peaked violin shape, indicating consistency across test images. While Dice + CE with Adam achieved slightly better sensitivity, Dice with Adam demonstrated lower variability. This is corroborated by Fig. 17, where Dice with Adam showed the lowest standard deviation in global metrics, particularly Dice, IoU, and F1-score. Based on these insights, DeepLabv3+ (Xception) with Dice loss and Adam were selected for the second stage of the final pipeline.

C. *Phase 4—Finalize 2-Stage DL Framework*

The final evaluation of the two-stage framework is summarized in Tables VI, VII, and VIII. Stage 1 achieved strong results, with a Dice (global) of 0.9727 and balanced scores across IoU, F1, sensitivity, and specificity—reflecting high-quality cavity detection. Stage 2 performed well for myocardium (Dice = 0.7351), although scar segmentation remained poor (Dice = 0.0556). The small size, heterogeneous appearance, and low intensity of scar tissue in LGE-MRI scans made it difficult for the model to reliably detect, even with post-processing and multi-channel input.

Qualitative outputs in Table VIII further highlight the effectiveness of the pipeline in cavity and myocardium segmentation but also reveal partial or missed scar regions—an expected limitation in highly imbalanced datasets. Nevertheless, the structural guidance introduced in the two-stage design helped enforce anatomical logic and reduced inter-class confusion between adjacent regions.

V. Conclusion

This study proposed a two-stage deep learning framework for the automated segmentation of LV structures and myocardial infarction in LGE-MRI images. The framework addresses key limitations of traditional and single-stage deep learning methods by separating the segmentation task into two focused stages: cavity detection followed by myocardium and scar segmentation. Development was conducted in four structured phases, including baseline model benchmarking, exploration of loss and optimizer configurations, integration of the best-performing models into a two-stage pipeline, and final evaluation. DeepLabv3+ with ResNet50 (Stage 1) and Xception (Stage 2), both trained with Dice loss and Adam optimizer, emerged as the most effective configurations. Final testing showed excellent segmentation performance (global) for the LV cavity (Dice = 0.9727) and myocardium (Dice = 0.7351), confirming the benefits of guided modular optimization. However, scar segmentation remained challenging, achieving a Dice score of only 0.0556 due to its small size and low contrast in LGE-MRI. Despite this, the two-stage approach improved structural coherence and reduced inter-class confusion. These findings support the value of a clinically informed, staged segmentation strategy for enhancing accuracy and robustness in cardiac image analysis, with future work needed to further refine scar detection through advanced modelling techniques.

This study utilized a dataset collected exclusively from AMDI to maintain consistency in imaging protocols, patient demographics, and annotation standards. Future work will aim to validate the proposed framework on larger, more diverse

datasets, including publicly available dataset, to assess generalizability and robustness of the proposed model across varied clinical settings.

ACKNOWLEDGMENT

The authors would like to thank the Imaging Department of the Advanced Medical and Dental Institute (AMDI), Universiti Sains Malaysia (USM), for providing access to the LGE-CMR datasets, and the radiologists involved in the manual annotation of ground truth segmentations. The authors also thank UiTM Cawangan Pulau Pinang for providing the necessary facilities and support throughout the research process. This research was supported by Ministry of Higher Education through Fundamental Research Grant Scheme (FRGS) (FRGS/1/2023/SKK06/UITM/02/12).

REFERENCE

[1] G. A. Roth *et al.*, "Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update from the GBD 2019 Study," Dec. 22, 2020, *Elsevier Inc.* doi: 10.1016/j.jacc.2020.11.010.

[2] Y. Wang, Q. Li, L. Bi, B. Wang, T. Lv, and P. Zhang, "Global trends in the burden of ischemic heart disease based on the global burden of disease study 2021: the role of metabolic risk factors," *BMC Public Health*, vol. 25, no. 1, Jan. 2025, doi: 10.1186/s12889-025-21588-9.

[3] N. Mewton, C. Y. Liu, P. Croisille, D. Bluemke, and J. A. C. Lima, "Assessment of myocardial fibrosis with cardiovascular magnetic resonance," Feb. 22, 2011, *Elsevier USA*. doi: 10.1016/j.jacc.2010.11.013.

[4] A. S Flett 1, J. Hasleton, Ch. Cook, D. Hausenloy, G. Quarta, C. Ariti, V. Muthurangu, J. C Moon., "Evaluation of techniques for the quantification of myocardial scar of differing etiology using cardiac magnetic resonance," *JACC Cardiovasc Imaging*, vol. 4, no. 2, pp. 150–156, Feb. 2011, doi: 10.1016/j.jcmg.2010.11.015.

[5] Y. Wu, Z. Tang, B. Li, D. Firmin, and G. Yang, "Recent Advances in Fibrosis and Scar Segmentation from Cardiac MRI: A State-of-the-Art Review and Future Perspectives," Aug. 03, 2021, *Frontiers Media S.A.* doi: 10.3389/fphys.2021.709230.

[6] Q. Tao, J. Milles, K. Zeppenfeld, H. J Lamb, J. J Bax, J. H. C. Reiber, R. J. Geest, "Automated segmentation of myocardial scar in late enhancement MRI using combined intensity and spatial information," *Magn Reson Med*, vol. 64, no. 2, pp. 586–594, 2010, doi: 10.1002/mrm.22422.

[7] X. Albà, R. M. Figueras I Ventura, K. Lekadir, C. Tobon-Gomez, C. Hoogendoorn, and A. F. Frangi, "Automatic cardiac LV Segmentation in MRI using modified graph cuts with smoothness and interslice constraints," *Magn Reson Med*, vol. 72, no. 6, pp. 1775–1784, Dec. 2014, doi: 10.1002/mrm.25079.

[8] T. Kurzendorfer, C. Forman, M. Schmidt, C. Tillmanns, A. Maier, and A. Brost, "Fully automatic segmentation of left ventricular anatomy in 3-D LGE-MRI," *Computerized Medical Imaging and Graphics*, vol. 59, pp. 13–27, Jul. 2017, doi: 10.1016/j.compmedimag.2017.05.001.

[9] D. Wei, Y. Sun, S. H. Ong, P. Chai, L. L. Teo, and A. F. Low, "Three-dimensional segmentation of the left ventricle in late gadolinium enhanced MR images of chronic infarction combining long- and short-axis information," *Med Image Anal*, vol. 17, no. 6, pp. 685–697, Aug. 2013, doi: 10.1016/j.media.2013.03.001.

[10] T. Kurzendorfer, A. Brost, and A. Maier, "Random Forest Based Left Ventricle Segmentation in LGE-MRI," Jul. 2017, pp. 152–160. doi: 10.1007/978-3-319-59448-4_15.

[11] E. de la Rosa, D. Sidibé, T. Decourselle, T. Leclercq, A. Cochet, and A. Lalande, "Myocardial infarction quantification from late gadolinium enhancement mri using top-hat transforms and neural networks," *Algorithms*, vol. 14, no. 8, Aug. 2021, doi: 10.3390/a14080249.

[12] D. M. Popescu, H. G. Abramson, R. Yu, C. Lai, J. K. Shade, K. C. Wu, M. Maggioni, N. A. Trayanova, "Anatomically informed deep learning on contrast-enhanced cardiac magnetic resonance imaging for scar segmentation and clinical feature extraction," *Cardiovasc Digit Health J*, vol. 3, no. 1, pp. 2–13, Feb. 2022, doi: 10.1016/j.cvdhj.2021.11.007.

[13] Q. Yue, X. Luo, Q. Ye, L. Xu, and X. Zhuang, "Cardiac Segmentation from LGE MRI Using Deep Neural Network Incorporating Shape and Spatial Priors."

[14] F. Zabihollahy, J. A. White, and E. Ukwatta, "Myocardial scar segmentation from magnetic resonance images using convolutional neural network," in *Medical Imaging 2018: Computer-Aided Diagnosis*, N. Petrick and K. Mori, Eds., SPIE, 2018, p. 105752Z. doi: 10.1117/12.2293518.

[15] K. Suwa, "Editorial for 'Improved Quantification of Myocardium Scar in Late Gadolinium Enhancement Images: Deep Learning Based Image Fusion Approach,'" *Journal of Magnetic Resonance Imaging*, vol. 54, no. 1, pp. 313–314, 2021, doi: https://doi.org/10.1002/jmri.27619.

[16] F. Righetti, G. Rubiu, M. Penso, S. Moccia, M. L. Carerj, M. Pepi, G. Pontone, E. G. Caiani., "Deep learning approaches for the detection of scar presence from cine cardiac magnetic resonance adding derived parametric images," *Med Biol Eng Comput*, vol. 63, no. 1, pp. 59–73, Aug. 2024, doi: 10.1007/s11517-024-03175-z.