

Optimizing Cardiac MRI Segmentation: An Ensemble Approach with U-Net Variants

Bárdos-Deák Botond
Faculty of Natural Sciences
Budapest University of Technology and Economics
Budapest, Hungary
bardos-deak.botond@edu.bme.hu

Bodai Adrián Tibor
Faculty of Natural Sciences
Budapest University of Technology and Economics
Budapest, Hungary
bodai.adrian.tibor@edu.bme.hu

Mohammed Salah Al-Radhi
Department of Telecommunications and Artificial Intelligence
Budapest University of Technology and Economics
Budapest, Hungary
malradhi@tmit.bme.hu

Abstract—Segmentation of cardiac magnetic resonance images is a critical task in medical imaging, particularly to delineate the left and right ventricles and the myocardium. This study aims to improve segmentation performance using an ensemble approach with variants of the U-Net architecture, a widely adopted deep learning model for image segmentation. Multiple segmentation models were trained and optimized, and their outputs were combined using threshold-based binary conversion. Two ensemble strategies were evaluated: (1) Averaging, where the mean value of the binary masks at each pixel location was calculated to smooth discrepancies among model predictions, and (2) Voting, where majority voting determined the final pixel classification. The proposed ensemble approach demonstrates robustness to individual model errors and improves segmentation consistency.

Index Terms—Image segmentation, Ensemble, Cardiac MRI, U-Net, MA-Net, Link-Net, DeepLabV3

I. INTRODUCTION

Accurate segmentation of cardiac magnetic resonance (MRI) images is crucial for the diagnosis and monitoring of heart disease. In this study, we address the segmentation of key cardiac structures, including the left ventricle, the right ventricle, and the myocardium. We used the ACDC data set [2], which was originally designed to classify heart diseases but is particularly well suited for segmentation tasks due to its high-quality annotations. The dataset includes 100 training and 50 test patient scans, with annotations for the three target regions. To facilitate processing, the 3D MRI data, originally in *.mni.gz* format, were converted into 2D slices. Each 3D image was split into 32 slices and saved as *.png* files, with masks for the three structures saved separately for greater flexibility. In this study, we build on the widely used U-Net architecture by exploring ensemble approaches to improve segmentation accuracy. Several models with ResNet-34 backbones were trained and optimized with varying parameters, loss functions, optimizers, and learning rates. Individual models were developed to segment each cardiac structure independently and jointly for multiclass segmentation. The

best performing models of each architecture were selected to create an ensemble model that combines their predictions.

A. Contributions

Our main contributions are summarized as follows:

- 1) **Optimized Ensemble Approach:** We propose an ensemble model that combines multiple architectures to address the limitations of individual models. Two ensemble strategies are evaluated:
 - **Averaging:** Combines binary masks by averaging pixel values to reduce discrepancies between predictions.
 - **Voting:** Employs majority voting for pixel classification to enhance robustness to outlier predictions.
- 2) **Comparison of Architectures:** We benchmark the performance of four state-of-the-art architectures—U-Net, DeepLabV3, MA-Net, and LinkNet—highlighting their strengths and limitations for cardiac MRI segmentation.
- 3) **Custom Dataset Preprocessing:** We adapt the ACDC dataset for efficient 2D image processing by creating standardized slices and independent masks, enabling model generalization and effective training.

B. Model Architectures

The following architectures were implemented and compared for their segmentation performance as shown in Figure 1:

- 1) **U-Net** [5]: A fully convolutional neural network designed for semantic segmentation. U-Net features an encoder-decoder structure with skip connections, enabling precise spatial reconstruction by fusing encoder and decoder blocks via concatenation.
- 2) **DeepLabV3** [4]: A model optimized for pixel-level classification. It leverages atrous (dilated) convolutions and the Atrous Spatial Pyramid Pooling (ASPP) module to capture multi-scale contextual information, enhancing segmentation across objects of varying sizes.

- 3) **MA-Net** [3]: A multi-scale attention network designed to capture rich contextual dependencies through two innovative components:
 - **Position-wise Attention Block (PAB)**: Captures spatial dependencies between pixels globally.
 - **Multi-scale Fusion Attention Block (MFAB)**: Exploits channel dependencies using multi-scale semantic feature fusion.
- 4) **LinkNet** [8]: A lightweight segmentation model with encoder-decoder architecture. It uses skip connections for efficient feature reuse and employs addition-based fusion for decoder reconstruction.

By leveraging the complementary strengths of these architectures in an ensemble framework, we aim to achieve state-of-the-art performance in cardiac MRI segmentation.

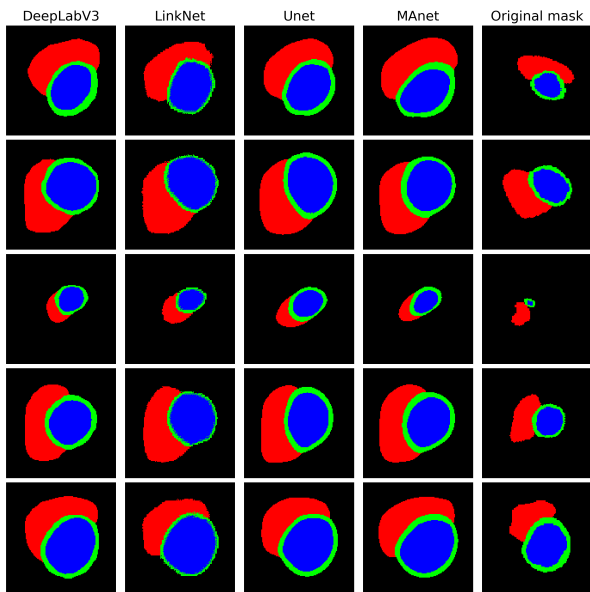


Fig. 1. Comparison of outputs from different architectures for cardiac MRI segmentation.

II. METHODOLOGY

A. Training Environment

The models were trained on a workstation equipped with an Intel Core Ultra 5 125U CPU, 16 GB of RAM, and an NVIDIA RTX 3060Ti GPU with 12 GB of VRAM, leveraging CUDA 12.4 for accelerated computations. This setup provided sufficient computational power for efficient training and evaluation of deep learning models on high-resolution cardiac MRI images.

B. Training Parameters and Optimization Strategy

To achieve robust performance across different architectures, we carefully selected and fine-tuned the training parameters. Specifically, we experimented with multiple loss functions tailored to the segmentation task:

- **Dice Loss**: Measures the overlap between predicted and ground-truth masks, emphasizing pixel-level accuracy in imbalanced datasets. This loss function was particularly effective for addressing the uneven distribution of cardiac structures.
- **Jaccard Loss**: A variant of Dice Loss, incorporating the intersection-over-union (IoU) metric. This loss offered an alternative perspective on segmentation quality, especially useful for refining boundary delineations.
- **Cross-Entropy Loss**: Measures the divergence between predicted and actual probability distributions at the pixel level. This loss was applied to ensure stable gradient propagation and convergence.

We utilized the Adam optimizer due to its adaptive learning rate properties, which facilitated rapid convergence while minimizing oscillations in the optimization process. Additionally, a cyclic learning rate scheduler was employed to dynamically adjust the learning rate during training. The minimum learning rate was set to 0.00006, and the maximum to 0.001. This approach allowed the model to escape local minima and explore diverse regions of the loss landscape, leading to better generalization.

C. Model Selection and Ensemble Construction

To capitalize on the strengths of different architectures, we trained multiple models (U-Net, DeepLabV3, MA-Net, and LinkNet) with the aforementioned parameters. Each model was optimized separately, and its performance was evaluated using standard segmentation metrics (Dice Score, IoU, etc.).

From each architecture, the top-performing configuration was selected for inclusion in the ensemble model. By combining the best-performing models, we aimed to enhance segmentation accuracy and robustness. The ensemble strategies (Averaging and Voting) were implemented to fuse predictions from these diverse architectures. This methodological innovation ensured that the final segmentation output benefited from the complementary strengths of individual models while mitigating their weaknesses.

D. Novelty of Approach

Our methodology stands out due to its emphasis on:

- 1) **Tailored Loss Functions**: We systematically evaluated multiple loss functions to address the challenges posed by imbalanced datasets and intricate boundaries in cardiac MRI segmentation.
- 2) **Dynamic Learning Rate Scheduling**: The use of a cyclic learning rate allowed for better exploration of the loss landscape, leading to improved generalization compared to static learning rates.
- 3) **Ensemble Optimization**: The integration of diverse architectures in an ensemble framework demonstrated significant improvement over individual models, particularly in handling variability across patients and cardiac structures.

- 4) **Fine-Grained Model Selection:** By rigorously benchmarking configurations and selecting the best-performing models for the ensemble, we ensured optimal utilization of architectural strengths.

This multi-faceted approach provides a robust foundation for achieving state-of-the-art performance in cardiac MRI segmentation tasks.

III. EVALUATION METRICS

To comprehensively assess the performance of our segmentation models, we employed the following evaluation metrics. These metrics were chosen to provide insights into various aspects of segmentation quality, such as overlap, accuracy, and handling of class imbalance.

A. F1 Score

The F1 score is the harmonic mean of precision and recall, offering a balanced evaluation of segmentation performance, particularly in cases where class imbalance is prevalent. This is crucial in cardiac MRI segmentation, where the sizes of anatomical structures (e.g., ventricles and myocardium) can vary significantly.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

Precision evaluates how many of the predicted positive pixels are correct, while recall measures how many of the true positive pixels were identified. The F1 score balances these two aspects, making it particularly suitable for tasks where false negatives and false positives need equal consideration.

B. Pixel Accuracy

Pixel accuracy measures the proportion of correctly classified pixels across the entire image, encompassing both foreground (anatomical structures) and background regions. This metric provides a high-level view of segmentation performance.

$$\text{Pixel Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

While pixel accuracy is straightforward to compute, it may not fully reflect performance when class imbalance exists, as it can be dominated by the background class.

C. Intersection over Union (IoU)

Intersection over Union (IoU), also known as the Jaccard Index, evaluates the overlap between the predicted and ground-truth masks for each class. It is defined as:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

IoU is particularly effective in providing class-specific evaluation. In this study, IoU was computed separately for each anatomical structure:

- **Left Ventricle (LV):** Measures segmentation quality of the left ventricle region.
- **Right Ventricle (RV):** Evaluates the prediction for the right ventricle region.
- **Myocardium (MYO):** Focuses on the myocardial wall, which is often the most challenging region to segment due to its thin and complex structure.

By analyzing IoU for each class, we gained a deeper understanding of the model's ability to segment different cardiac regions, helping identify specific areas for improvement.

D. Importance of Metric Diversity

The combination of F1 Score, Pixel Accuracy, and IoU ensures a holistic evaluation of segmentation performance. F1 Score focuses on balancing false positives and false negatives, Pixel Accuracy offers a global view, and IoU provides class-specific insights. Together, these metrics allow for robust and detailed model assessment, ensuring that improvements in one area (e.g., global accuracy) do not come at the expense of another (e.g., class-specific performance).

IV. ENSEMBLE MODEL

A. Construction of the Ensemble Model

To enhance the accuracy and robustness of our segmentation results, we designed an ensemble model that leverages the complementary strengths of multiple architectures. The ensemble was constructed by combining the outputs of four state-of-the-art segmentation models: **DeepLabV3**, **LinkNet**, **UNet**, and **MAnet**. These models were chosen due to their demonstrated effectiveness in handling diverse segmentation challenges, such as capturing multi-scale features and preserving spatial details.

Each model independently generated segmentation masks for the left ventricle (LV), right ventricle (RV), and myocardium (MYO). To unify these predictions, a threshold-based approach was applied to convert the continuous outputs of each model into binary masks (foreground vs. background). The threshold values were optimized during the validation phase, ensuring each model's outputs were tailored to its performance characteristics:

- **DeepLabV3:** Threshold of 5.31
- **LinkNet:** Threshold of 3.33
- **UNet:** Threshold of 6.87
- **MAnet:** Threshold of 5.36

The thresholding process allowed us to fine-tune each model's output, ensuring better alignment with the ground truth. Figure 2 illustrates the impact of different threshold values on the segmentation masks, showcasing how the thresholds influenced the delineation of cardiac regions.

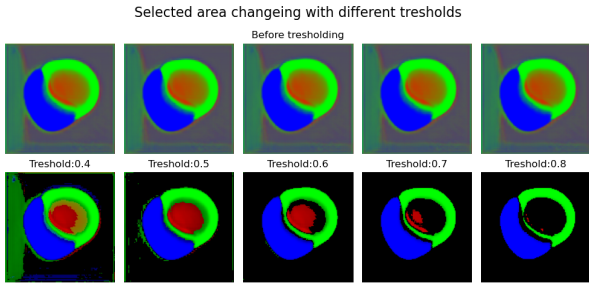


Fig. 2. Effect of different threshold values on segmentation masks.

B. Ensemble Strategy

The outputs of the thresholded segmentation masks were aggregated using two distinct ensemble strategies, each designed to exploit the strengths and mitigate the weaknesses of the individual models:

- 1) **Averaging Method:** In this method, the binary masks produced by the individual models were averaged pixel-wise. For each pixel location, the mean value was calculated across all models:

$$M_{\text{avg}}(x, y) = \frac{1}{N} \sum_{i=1}^N M_i(x, y),$$

where $M_i(x, y)$ represents the binary mask of the i -th model, and N is the total number of models in the ensemble. The resulting averaged mask smooths out inconsistencies between model predictions and provides a more consistent output. Pixels with mean values exceeding a specified threshold were classified as foreground, ensuring a balance between sensitivity and precision.

- 2) **Voting Method:** The voting method used a majority voting scheme to determine the final classification for each pixel. For a given pixel location (x, y) , if more than 50% of the models classified it as foreground, the pixel was labeled as foreground in the ensemble mask:

$$M_{\text{vote}}(x, y) = \begin{cases} 1 & \text{if } \sum_{i=1}^N M_i(x, y) > \frac{N}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

This strategy is particularly robust to errors made by individual models, as it relies on consensus among the majority. It is especially useful in reducing the impact of outliers and noise in the predictions.

C. Benefits of the Ensemble Approach

The ensemble model effectively combines the strengths of individual architectures, offering the following advantages:

- **Robustness:** By aggregating predictions, the ensemble mitigates the risk of poor performance from any single model, improving overall reliability.
- **Precision and Sensitivity:** The averaging method reduces prediction variance, while the voting method enhances robustness against outliers.

- **Adaptability:** Class-specific thresholds and multiple aggregation strategies allow the ensemble to adapt to the unique challenges posed by each cardiac structure (LV, RV, MYO).

The combination of these methods ensures that the ensemble achieves a balanced trade-off between sensitivity and specificity, making it a robust solution for cardiac MRI segmentation tasks. However, it should be noted that this approach necessitates a substantially greater computational capacity.

V. RESULTS

The performance of the ensemble model and individual models was evaluated using three key metrics: **F1 Score**, **Pixel Accuracy**, and **Intersection over Union (IoU)**. These metrics provide a comprehensive assessment of the segmentation performance, with the F1 Score evaluating the balance between precision and recall, Pixel Accuracy indicating overall correctness in pixel classification, and IoU measuring the overlap between predicted and ground-truth segmentation masks. The average scores reported reflect the overall performance across all test images.

Our results show that all models performed similarly, with minor variations across the metrics. The **average F1 Score** across all models was approximately **0.32**, while the **average Pixel Accuracy** was notably high, at **0.98**. The high Pixel Accuracy suggests that the models are able to classify most pixels correctly, but the relatively low F1 Score indicates the challenge of handling class imbalances, particularly for smaller or more difficult-to-segment structures, such as the left ventricle (LV), right ventricle (RV), and myocardium (MYO). These minority classes often contribute to a lower F1 Score due to their smaller spatial footprint in the image.

The **IoU** values further emphasize the difficulty in achieving accurate segmentation for these smaller structures, with the IoU values for individual models being lower than those for the ensemble. However, the ensemble model demonstrated a slight improvement in both the F1 Score and IoU, showcasing its effectiveness in integrating the strengths of the individual models.

Table I summarizes the detailed results for each model, including the F1 Score, Pixel Accuracy, and IoU:

TABLE I
PERFORMANCE METRICS FOR INDIVIDUAL MODELS AND THE ENSEMBLE MODEL.

Model	F1 Score	Pixel Accuracy	IoU
UNet	0.33	0.98	0.31
DeepLabV3	0.31	0.98	0.29
LinkNet	0.32	0.98	0.30
MAnet	0.32	0.98	0.30
Ensemble	0.34	0.98	0.32

The ensemble model slightly outperformed the individual models in all three metrics, achieving the highest F1 Score (**0.34**) and IoU (**0.32**), while maintaining the same high Pixel Accuracy (**0.98**). This demonstrates the effectiveness of the ensemble approach in improving segmentation performance

by leveraging the complementary strengths of multiple models. By aggregating the predictions of several models, the ensemble is able to smooth out individual model errors and better capture the complexities of cardiac tissue segmentation.

It is worth noting that the improvement achieved by the ensemble model, while statistically consistent, is modest. The ensemble's slight advantage in metrics must be balanced against the computational cost, as the runtime of the ensemble model is effectively the sum of the runtimes of the individual networks. However, the voting mechanism used for aggregation introduces negligible additional computational overhead, ensuring that the overall increase in runtime remains manageable.

Given that the improvement in metrics is small, the choice to use an ensemble model should be guided by task-specific requirements, such as the need for robust predictions in critical applications like medical imaging. The results suggest that while the marginal gains in F1 Score and IoU may not always justify the additional computational cost, the trade-off can still be acceptable in scenarios where accuracy and reliability are paramount.

Overall, the results validate the use of an ensemble model in segmentation tasks, particularly when dealing with class imbalances and the need for precise and consistent predictions in high-stakes applications such as cardiac tissue segmentation.

VI. CONCLUSION AND FUTURE WORK

In this study, we presented a comprehensive evaluation of several state-of-the-art segmentation models for cardiac MRI analysis, including **DeepLabV3**, **LinkNet**, **UNet**, and **MANet**, as well as an ensemble model combining their outputs. Our results demonstrate that while individual models performed similarly, with high Pixel Accuracy (0.98), the F1 Score and Intersection over Union (IoU) values highlighted challenges in segmenting minority classes such as the left ventricle, right ventricle, and myocardium, due to class imbalances. The ensemble model outperformed the individual models, achieving higher F1 Score (0.34) and IoU (0.32), illustrating the effectiveness of aggregating multiple model outputs to enhance segmentation performance. The final model has been enhanced to reach the baseline, but due to limitations in computational power, it remains significantly below the state-of-the-art results, where accuracies of (IoU) are 0.7 or higher, and F1 scores of over 0.8 can be attained ([1], [7], [6]).

The key takeaway from our results is the importance of using an ensemble approach to address the limitations of single models, particularly when dealing with class imbalances in medical image segmentation. By combining the strengths of different architectures, the ensemble model was able to provide more accurate and robust segmentation, especially for smaller and harder-to-segment structures in cardiac imaging. Despite the promising results, there are several avenues for future work that could further improve segmentation performance. One of them is integrating attention mechanisms within the ensemble models, which could enhance the focus on important regions of interest, particularly for smaller structures like the

left and right ventricles that may be obscured by surrounding tissues. This could be achieved by incorporating attention-based models or modules that prioritize relevant features in the segmentation task.

VII. ACKNOWLEDGMENTS

This paper is supported by the European Union's HORIZON Research and Innovation Programme under grant agreement No 101120657, project ENFIELD (European Lighthouse to Manifest Trustworthy and Green AI) and by the Ministry of Innovation and Culture and the National Research, Development and Innovation Office of Hungary within the framework of the National Laboratory of Artificial Intelligence. M.S.Al-Radhi's research was supported by the **EKÖP-24-4-II-BME-197**, through the National Research, Development and Innovation (NKFI) Fund.

REFERENCES

- [1] Narjes Benameur et al. *An Improved Approach for Cardiac MRI Segmentation based on 3D UNet Combined with Papillary Muscle Exclusion*. 2024. arXiv: 2410.06818 [cs.CV]. URL: <https://arxiv.org/abs/2410.06818>.
- [2] Olivier Bernard et al. "Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?" In: *IEEE Transactions on Medical Imaging* 37.11 (2018), pp. 2514–2525. DOI: 10.1109/TMI.2018.2837502.
- [3] Abhishek Chaurasia and Eugenio Culurciello. "LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation". In: *CoRR* abs/1707.03718 (2017). arXiv: 1707.03718. URL: <http://arxiv.org/abs/1707.03718>.
- [4] Liang-Chieh Chen et al. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". In: *CoRR* abs/1802.02611 (2018). arXiv: 1802.02611. URL: <http://arxiv.org/abs/1802.02611>.
- [5] Tongle Fan et al. "MA-Net: A Multi-Scale Attention Network for Liver and Tumor Segmentation". In: *IEEE Access* 8 (2020), pp. 179656–179665. DOI: 10.1109/ACCESS.2020.3025372.
- [6] Pedro Ferreira et al. "Automating in vivo cardiac diffusion tensor postprocessing with deep learning-based segmentation". In: *Magnetic Resonance in Medicine* 84 (Apr. 2020). DOI: 10.1002/mrm.28294.
- [7] Chao Luo et al. "Cardiac MR segmentation based on sequence propagation by deep learning". In: *PLOS ONE* 15 (Apr. 2020), e0230415. DOI: 10.1371/journal.pone.0230415.
- [8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *CoRR* abs/1505.04597 (2015). arXiv: 1505.04597. URL: <http://arxiv.org/abs/1505.04597>.