Semi-Supervised Left Atrium Segmentation in 3D Cardiac MRI Using Confidence-Guided Pseudo-Labeling

Peaysararn Rapinrangchang*, Tanasai Sucontphunt
Graduate School of Applied Statistics,
National Institute of Development Administration

Received: 27 July 2025
Revised: 21 October 2025
Accepted: 28 October 2025

Abstract

This study proposes a semi-supervised learning framework for left atrium (LA) segmentation from three-dimensional cardiac Magnetic Resonance Imaging (MRI) using pseudo-labeling. The objective is to improve segmentation performance under limited labeled-data conditions. The proposed method integrates a 3D U-Net architecture with an iterative training pipeline and dynamic confidence-based pseudo-label refinement. Using the Medical Segmentation Decathlon dataset, experiments demonstrate that the semi-supervised model achieves a mean Dice Coefficient (DSC) of 0.9066 ± 0.0043 and a mean Average Hausdorff Distance (AHD) of 2.2409 ± 0.3661, surpassing the fully supervised baseline (DSC: 0.8519 ± 0.0395; AHD: 4.7696 ± 1.3128). Qualitative evaluation further confirms reduced false positives and enhanced anatomical precision. The results indicate that the proposed approach effectively leverages unlabeled data to achieve high segmentation accuracy with minimal manual annotation, providing a practical solution for clinical management of atrial fibrillation (AF).

Keywords: Left Atrium Segmentation, Semi-supervised Learning, Deep Learning

*Corresponding Author; E-mail: 6610422018@stu.nida.ac.th

Introduction

AF is a significant global public health concern. While the age-adjusted prevalence of AF has remained relatively stable between 1990 and 2019, the numbers of affected individuals and Disability-Adjusted Life Years (DALYs) have increased significantly. This rise is primarily attributed to population aging and growth (Roth et al., 2020). Recent studies in the United States have reported an AF prevalence as high as 3.89%, exceeding previous estimations, with variations across demographic groups. For instance, men are more prone to AF than women, while non-Hispanic Whites exhibit a higher prevalence compared with other groups (Oltman et al., 2024). In Thailand, community-based screening among adults aged ≥ 65 years showed an AF prevalence of 2.8% (Suwanwela et al., 2021), and among hypertensive patients, the prevalence was 3.46% (Krittayaphong et al., 2016). These findings highlight a heightened burden in high-risk populations such as the elderly and those with cardiovascular comorbidities, underscoring the need for effective screening and diagnostic methods.

A detailed understanding of the LA structure is crucial for AF management. LA imaging using MRI provides highly accurate structural information, serving as a vital tool for optimal treatment planning. Deep learning—particularly convolutional neural networks (CNNs)—has emerged as a powerful technique for medical image segmentation due to its ability to automatically extract hierarchical and spatial features. This enables more accurate, robust, and scalable segmentation compared with traditional methods that rely on handcrafted features or manual rules. In the context of LA segmentation, deep learning models can better capture complex anatomical variations, which is critical for clinical decision-making. Although deep learning techniques for MRI analysis and segmentation can yield promising results, many rely on fully supervised learning, which requires large amounts of accurately labeled data. This requirement poses a significant challenge, as acquiring expert-annotated medical images is costly and time-consuming. In contrast, semi-supervised learning approaches aim to reduce dependency on labeled data by leveraging both labeled and unlabeled datasets. These methods strike a balance between performance and annotation efficiency, and they have recently gained attention as a promising direction in medical imaging. For instance, a recent study demonstrated that semi-supervised learning techniques achieved significantly improved segmentation performance for brain metastases on multicenter MRI data, especially in low-label scenarios (Ottesen et al., 2024). Their findings underscore the robustness and generalizability of semi-supervised approaches across heterogeneous datasets, highlighting their potential in real-world clinical applications where annotated data is limited. Therefore, integrating deep learning with semi-supervised techniques offers a practical solution to enhance the utilization of available data while reducing the reliance on extensive manual annotation by medical experts (Wang et al., 2022).

In practice, deep learning models for MRI-based segmentation typically require a substantial amount of annotated data to achieve robust and generalizable performance. The effectiveness of these models has been shown to improve with increases in both the number of labeled training samples and the spatial resolution of the input data, reflecting the importance of data volume and structural detail in complex anatomical segmentation tasks (Yang et al., 2024). Similarly, a study evaluating U-Net performance in organ segmentation found that the DSC started at 0.424 with just eight labeled cases and increased substantially to 0.858 with 160 cases, reaching 0.867 when trained with 320 cases. This trend underscores the strong dependency of segmentation accuracy on the availability of sufficient annotated data (Bardis et al., 2020). These observations reinforce the notion that high-quality labeled datasets—often comprising tens to hundreds of annotated cases—remain critical factors in the success of supervised deep learning methods for MRI segmentation.

Despite the clear benefits of using large annotated datasets, assembling such datasets remains a formidable challenge. An analysis of Medical Image Computing and Computer Assisted Intervention (MICCAI) publications between 2011 and 2019 revealed that the geometric mean of dataset sizes grew at annual rates of approximately 21% for MRI, 24% for Computed Tomography (CT), and 31% for functional MRI (fMRI), reflecting increasing community expectations for larger datasets in medical image analysis (Kiryati and Landau, 2021). In a more tangible exemplar, segmentation of multi-organ CT volumes in the AbdomenAtlas project estimated that a trained annotator might spend 30–60 minutes per CT volume under manual annotation, and completing all 8,448 CT volumes by conventional means would require on the order of 1,600 weeks (~30.8 years) of continuous effort (Qu et al., 2023). Such formidable time and labor demands underline the infeasibility of relying solely on fully annotated datasets, motivating approaches such as semi-supervised learning, active learning, pre-annotation, and annotation-correction guidance to alleviate annotation burden.

This research employs semi-supervised learning to address the limited labeled data challenge in LA segmentation from MRI images. In this study, we intentionally adopt an extreme low-data scenario as the labeled set while leveraging additional unlabeled images via pseudo-labeling. This constraint is chosen to simulate real-world clinical settings where expert-annotated data are scarce. The proposed method will be benchmarked against a fully supervised baseline under identical data conditions to evaluate how much semi-supervised learning can mitigate performance degradation due to the low sample size.

Research Objectives

1. To develop a semi-supervised learning framework for LA segmentation in 3D cardiac MRI using pseudo-labeling.

2. To benchmark the proposed method against a fully supervised baseline under limited labeled-data conditions.

Literature Review

Conventional LA segmentation often employs CNNs in conjunction with supervised learning techniques. For instance, Zhang et al. (2021) presented a CNN model integrated with Bayesian filtering for LA segmentation from MRI, achieving a DSC of 94.1% for 2-chamber, 3-chamber, and 4-chamber views. Similarly, Aryan et al. (2022) utilized U-Net with ground-truth MRI data, reporting a DSC of 0.94. However, traditional CNN-based approaches are limited by their reliance on large amounts of labeled data, which can be a significant obstacle in medical applications where such data are scarce. To address this limitation, Swetha et al. (2023) proposed an improved U-Net architecture incorporating skip connections to enhance LA segmentation accuracy while reducing the dependence on extensive labeled data. Furthermore, Uslu et al. (2021) introduced LA-Net, which employs cross-attention modules (CAMs) and enhanced decoder modules (EDMs) for LA segmentation without requiring post-processing, demonstrating advancements in techniques that are less reliant on labeled data.

The DSC has become a widely adopted metric for evaluating segmentation performance in medical imaging, particularly in tasks involving small and irregular structures like the LA. According to Taha and Hanbury (2015), DSC provides a robust and intuitive measure of spatial overlap between predicted and reference segmentations, making it especially suitable for 3D image analysis. However, DSC, being overlap-based, may not fully capture spatial discrepancies, and it therefore should be considered alongside other evaluation metrics in some cases. Similarly, Müller et al. (2022) proposed practical guidelines for selecting evaluation metrics in medical image segmentation, recommending DSC as a primary metric due to its interpretability, while also encouraging the use of complementary measures—such as the AHD and sensitivity—to gain a more comprehensive understanding of model performance, especially in the presence of class imbalance.

To mitigate the limitations associated with extensive labeled data requirements, semi-supervised learning has gained attention for LA segmentation, leveraging unlabeled data in conjunction with a limited amount of labeled data. For example, Wang et al. (2022) presented a Dual-Consistency technique using only 20% labeled data and 80% unlabeled data, employing model-level perturbations and structure-level spatial contextual perturbations to improve LA

segmentation accuracy. Likewise, Shi et al. (2024) employed a Multi-Level Consistency technique emphasizing consistency across multiple levels, including the use of virtual adversarial training to enhance unlabeled data utilization, achieving a DSC of 91.69%. Additionally, Liu et al. (2022) proposed a semi-supervised framework comprising a segmentation model and a classification model with a contrastive consistency loss to facilitate better discrimination between labeled and unlabeled data, achieving a DSC of 89.81%. More recently, Xu et al. (2024) introduced SAMatch, which utilizes a SAM-guided match-based framework to generate pseudo-labels from unlabeled data, demonstrating excellent results in LA segmentation even with limited labeled data, such as achieving a DSC of 89.36% on the Automated Cardiac Diagnosis Challenge (ACDC) cardiac MRI dataset. These techniques demonstrate the potential of semi-supervised learning to improve LA segmentation performance without necessitating large labeled datasets.

While supervised learning can achieve satisfactory results in LA segmentation, it remains dependent on substantial labeled data, which can be a constraint in medical contexts due to the limited availability of such data. Conversely, semi-supervised learning techniques can leverage unlabeled data alongside a small amount of labeled data to enhance LA segmentation performance without excessive reliance on labeled data. Therefore, this research compares the performance of semi-supervised learning against supervised learning to evaluate which technique achieves superior LA segmentation accuracy in limited labeled-data scenarios.

Research Methodology

This study proposes a semi-supervised learning framework for LA segmentation in three-dimensional cardiac MRI, aiming to overcome the challenge of limited annotated data. The proposed method integrates a 3D U-Net architecture with a confidence-based pseudo-labeling strategy and an iterative training pipeline. The methodology comprises four main components: dataset preparation, model architecture, pseudo-labeling strategy, and training procedure.

1) Dataset Preparation

Experiments were conducted using the Medical Segmentation Decathlon (MSD) Task 02: Heart dataset (Antonelli et al., 2022), which consists of 30 cardiac MRI volumes—20 labeled and 10 unlabeled. Two dataset configurations were prepared for different experimental settings:

- Baseline Set: This configuration uses only the labeled data. Ten labeled volumes were used for training and five for validation. The remaining five labeled volumes were excluded from both training and validation, and no unlabeled data were used. This setup serves as the baseline for fully supervised training.

- Semi-Supervised Set: This configuration includes both labeled and unlabeled data. The same ten labeled volumes were used for training and five for validation. The remaining five labeled volumes were reassigned as unlabeled, combining with the ten original unlabeled volumes to form a pool of fifteen unlabeled samples. To enhance the diversity of unlabeled data, augmentation techniques such as random rotation and zoom were applied, resulting in thirty unlabeled samples used for pseudo-label generation.

The configurations of both sets are summarized in Table 1.

Table 1 Dataset Configurations

Configuration	Training (Labeled)	Unlabeled (Original)	Unlabeled (Augmented)	Validation (Labeled)	Total
Baseline Set	10	0	0	5	15
Semi-Supervised Set	10	15	15	5	45

2) Model Architecture

The segmentation model is based on a 3D U-Net architecture, implemented using the MONAI framework due to its optimized support for 3D medical imaging and integration with PyTorch. The architecture comprises an encoder–decoder structure with skip connections and residual units, facilitating the preservation of spatial and contextual information throughout the network. All models were trained using the Adam optimizer with a learning rate of 0.0005, a batch size of 2, and a weight decay of 0.0001. To ensure a fair comparison, these hyperparameters were held constant across both supervised and semi-supervised training regimes. All experiments were executed on an NVIDIA Tesla P100 Graphics Processing Unit (GPU) with 15 GB of Video Random Access Memory (VRAM).

3) Pseudo-Labeling Strategy

To leverage unlabeled data, a confidence-aware pseudo-labeling mechanism was employed. During each training iteration, the current model was used to generate segmentation predictions for the unlabeled volumes. Voxel-wise confidence scores were obtained from softmax probabilities, and only predictions that exceeded a dynamically adjusted confidence threshold were retained as pseudo-labels. The threshold was initialized at 70% and incrementally increased up to 90%, based on the model's validation performance. As the validation DSC improved, the threshold was raised to filter out low-confidence predictions, thereby ensuring the quality of the pseudo-labels. This dynamic adjustment helped reduce the propagation of noisy or unreliable labels into subsequent training stages.

These threshold values (0.70–0.90) were selected based on empirical observations. In our experiments, when the threshold was set below 0.6, the validation DSC often fluctuated significantly, suggesting the inclusion of noisy predictions. Setting the initial threshold (the starting point) at 0.7 helped stabilize training, while increasing the threshold toward 0.9 as the model improved avoided overly strict filtering that would discard useful information.

The total training loss was defined as a weighted sum of two components: a supervised loss from ground-truth annotations and an unsupervised loss from pseudo-labeled data. Initially, the pseudo-label loss was assigned a low weight (0.01) to minimize the influence of potential noise. This weight was gradually increased—up to a maximum of 0.4—according to model performance. This adaptive weighting strategy mitigated confirmation bias and enhanced training stability.

The selected range of 0.01–0.4 for the pseudo-label loss weight was also based on extensive experiments. Fixed weights ranging from 0.1 to 1.0 were initially tested, and the best DSC values were consistently observed when the weight was between 0.2 and 0.4. We then explored adaptive schemes across several ranges (e.g., 0.1–0.3, 0.1–0.5) and found that a dynamic range of 0.01–0.4 yielded the most stable and high-performing results. Starting from 0.01 allowed the model to cautiously incorporate pseudo-labels, while increasing the weight as performance improved enabled better utilization of unlabeled data.

It should be noted that these values were empirically tuned for this specific dataset. When applying the method to other datasets, re-tuning these thresholds and weights is recommended.

4) Training Pipeline

The semi-supervised training follows a four-stage iterative pipeline designed to progressively refine segmentation performance (Figure 1):

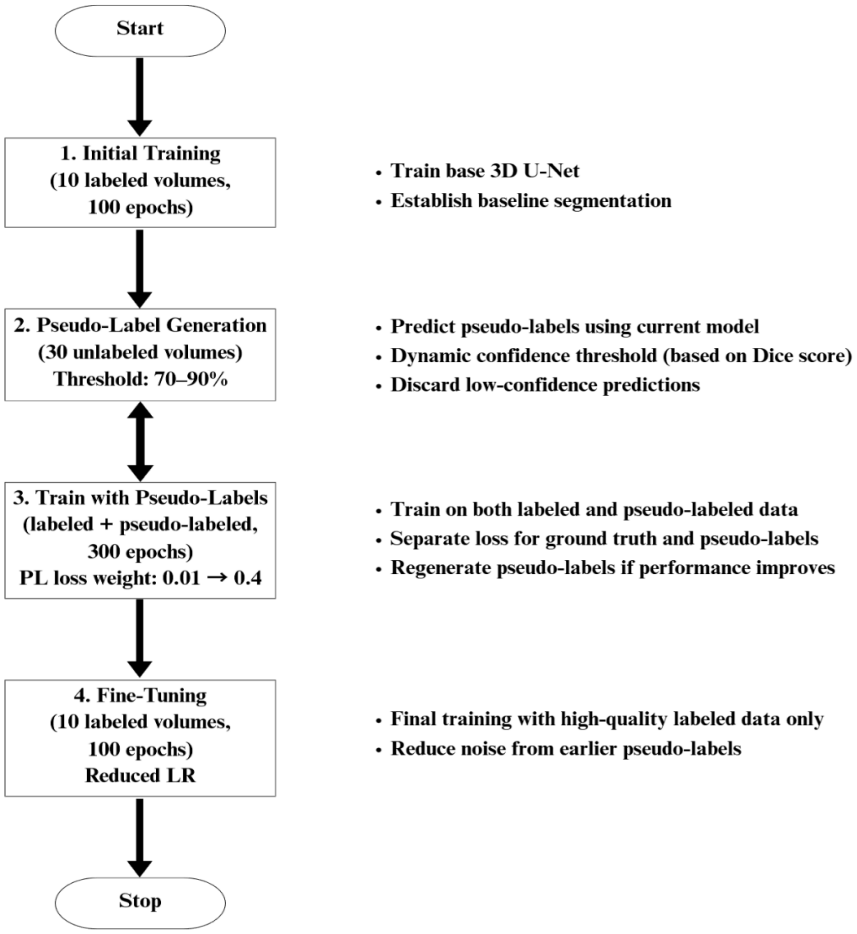
1. Initial Training: The base 3D U-Net is trained for 100 epochs using the 10 labeled training volumes to establish baseline segmentation.

2. Pseudo-Label Generation: The trained model generates pseudo-labels for all 30 unlabeled volumes. A dynamic confidence threshold (70%–90%), determined by the validation DSC, filters predictions to retain only high-confidence labels.

3. Training with Pseudo-Labels: The model is trained for 300 epochs on the combined labeled and pseudo-labeled data. Supervised and unsupervised losses are calculated separately, with the pseudo-label loss weight gradually increasing from 0.01 to 0.4 to mitigate the impact of label noise. Pseudo-labels are regenerated iteratively as model performance improves.

4. Fine-Tuning: After multiple refinement cycles, the model is fine-tuned for 100 epochs using only the original labeled data with a reduced learning rate to stabilize performance and mitigate residual noise effects.

Early stopping is employed throughout training, terminating if the validation loss fails to improve for 20 consecutive epochs. Evaluations are conducted every two epochs to prevent overfitting and optimize computational resources.



All stages use early stopping (20-epoch patience), evaluated every 2 epochs.

Figure 1 Training pipeline

5) Evaluation Metrics

To evaluate segmentation performance, two standard metrics are used: the DSC and the AHD. These metrics assess the overlap quality and boundary accuracy between predicted and ground truth segmentations.

1. Dice Similarity Coefficient (DSC) The DSC quantifies the overlap between the predicted segmentation and the ground truth. It is defined as:

$$DSC = \frac{2|P \cap G|}{|P| + |G|} \quad (1)$$

where:

P = set of predicted voxels

G = set of ground truth voxels

A DSC value ranges from 0 to 1, where 1 indicates perfect overlap and 0 indicates no overlap. This metric is particularly useful for evaluating the segmentation accuracy of anatomical structures.

2. Average Hausdorff Distance (AHD) The AHD measures how closely the boundary of the predicted segmentation aligns with that of the ground truth. It is defined as the symmetric average of the distances from all points on one boundary to the closest point on the other boundary:

$$AHD(A,B) = \max\left(\frac{1}{|A|}\sum_{a\in A}\min_{b\in B}\|a-b\|, \frac{1}{|B|}\sum_{b\in B}\min_{a\in A}\|b-a\|\right) \quad (2)$$

where:

A = set of boundary points from the predicted segmentation

B = set of boundary points from the ground truth segmentation

This metric computes the average of the minimum distances between boundary points of both segmentations. A lower AHD value indicates better boundary alignment and, consequently, higher segmentation accuracy—particularly important in applications involving fine or complex anatomical structures.

Findings

To evaluate the effectiveness of the proposed semi-supervised segmentation framework, a series of experiments were conducted to compare different training strategies: (1) fully supervised learning; (2) a vanilla semi-supervised variant excluding several key components (no confidence thresholding, no pseudo-label weighting, no iterative pseudo-label generation, and no fine-tuning); (3) semi-supervised learning without fine-tuning; and (4) semi-supervised learning with the full configuration. All experiments were performed under identical settings, including consistent dataset splits, hyperparameters, and computational resources.

1. Quantitative Results

Table 2 presents the quantitative results of segmentation accuracy and training time for all models. The full configuration semi-supervised model achieved the best performance, yielding a mean DSC of 0.9066 ± 0.0043 and the lowest mean AHD of 2.2409 ± 0.3661. This significantly outperforms the fully supervised baseline, which attained a DSC of 0.8519 ± 0.0395 and an AHD of 4.7696 ± 1.3128.

Notably, the vanilla semi-supervised variant (DSC 0.8505) showed no improvement in overlap compared to the baseline but did achieve a better boundary alignment (AHD 2.8317). The model without fine-tuning (DSC 0.9045, AHD 2.4762) approached the performance of the full configuration, highlighting the substantial impact of the iterative pseudo-labeling. The markedly lower standard deviations for both DSC and AHD in the full semi-supervised approach indicate more stable and consistent performance across validation samples.

However, this improved accuracy comes at the cost of increased computational time. The full configuration (151 ± 11 min) required substantially more training time than the supervised baseline (53 ± 4 min), reflecting the trade-off between performance and computational resources.

To ensure statistical reliability, each model was trained and evaluated five times independently, and the results are reported as mean ± standard deviation. This approach mitigates the effects of randomness in training and provides a robust measure of consistency across runs.

Table 2 Comparison of segmentation performance

Model	Dice Score (Validation Set)	Average Hausdorff Distance	Training Time (min)
Fully Supervised	0.8519 ± 0.0395	4.7696 ± 1.3128	53 ± 4
Semi-Supervised (Vanilla)	0.8505 ± 0.0298	2.8317 ± 0.5691	92 ± 17
Semi-Supervised (No Fine-Tuning)	0.9045 ± 0.0071	2.4762 ± 0.4735	126 ± 10
Semi-Supervised (Full Configuration)	0.9066 ± 0.0043	2.2409 ± 0.3661	151 ± 11

Figure 2 illustrates the training dynamics of both models. In the supervised setting, training loss decreased steadily, while the validation DSC plateaued around 0.85. In contrast, the semi-supervised model exhibited fluctuations in training loss during epochs 100–250, coinciding with the integration of pseudo-labeled data. Despite these fluctuations, the validation DSC continued to improve, ultimately surpassing the supervised baseline following the fine-tuning stage. These observations underscore the importance of iterative pseudo-label refinement in enhancing segmentation accuracy, even under increased training complexity.

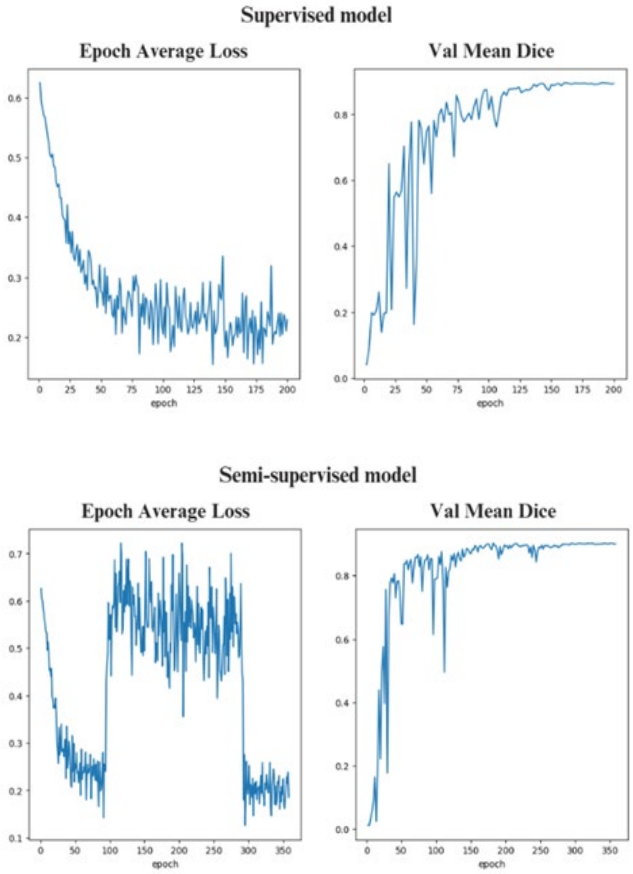


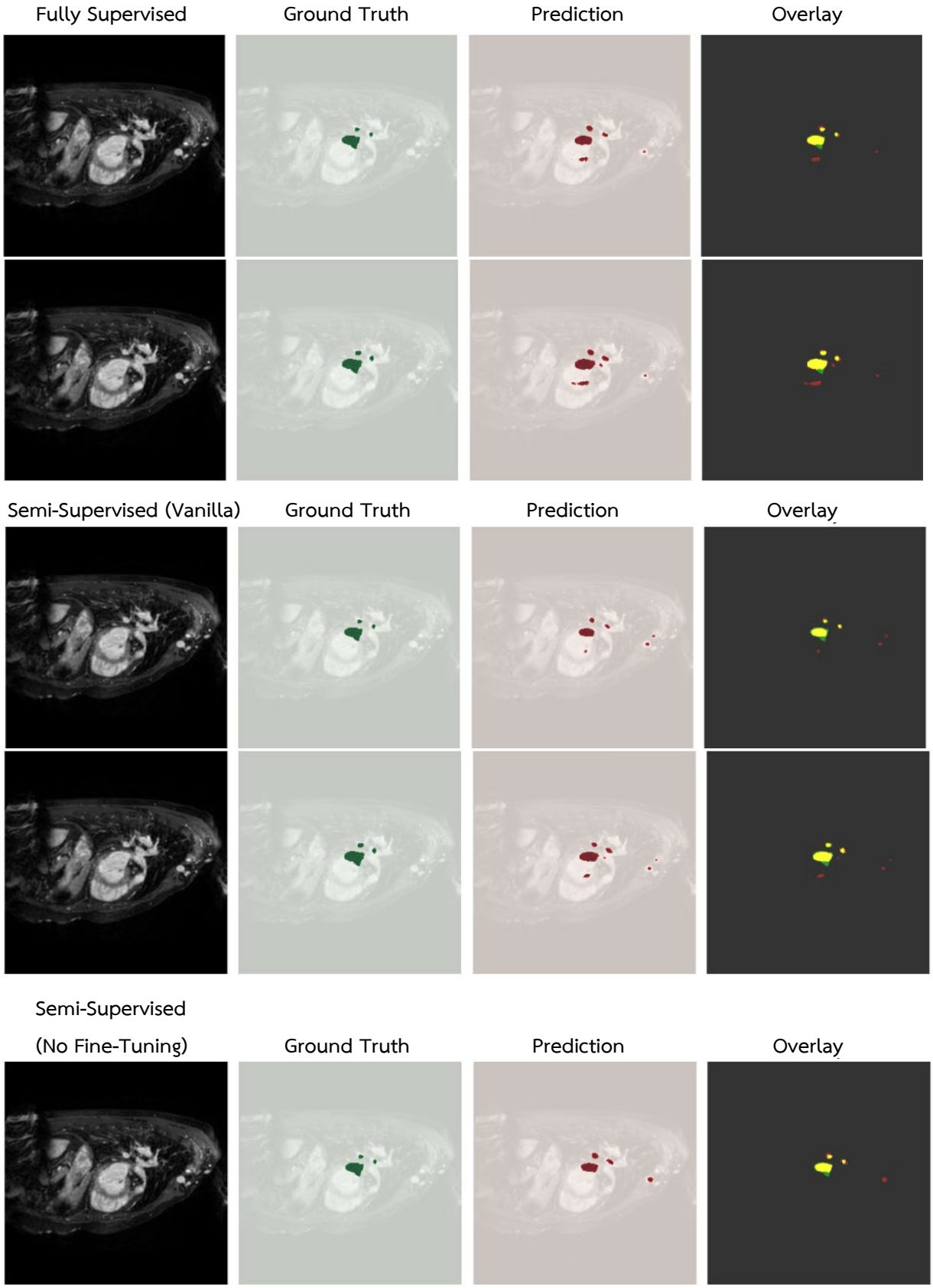
Figure 2 Training loss and validation Dice score for supervised vs. semi-supervised models

2. Qualitative Evaluation

In addition to quantitative metrics, a qualitative assessment was performed on withheld validation volumes to evaluate the anatomical plausibility and visual consistency of the segmentation results. As shown in Figure 3, all semi-supervised variants produced fewer false positives in non-cardiac regions and demonstrated more anatomically accurate delineations of the LA boundary compared to the fully supervised model.

Among the semi-supervised methods, both the configuration without fine-tuning and the full configuration yielded noticeably cleaner segmentations than the vanilla semi-supervised variant, which exhibited residual artifacts and over-segmentation in peripheral regions. The full-configuration model achieved the lowest number of false positives overall, while the no-fine-tuning variant produced slightly higher false positives but still performed substantially better than the vanilla baseline.

These visual observations are consistent with the quantitative results reported in Table 2, reinforcing that iterative pseudo-label refinement and fine-tuning contribute to improved anatomical precision and reduced segmentation noise.


Fully Supervised
Ground Truth
Prediction
Overlay
Semi-Supervised (Vanilla)
Ground Truth
Prediction
Overlay
Semi-Supervised
(No Fine-Tuning)
Ground Truth
Prediction
Overlay

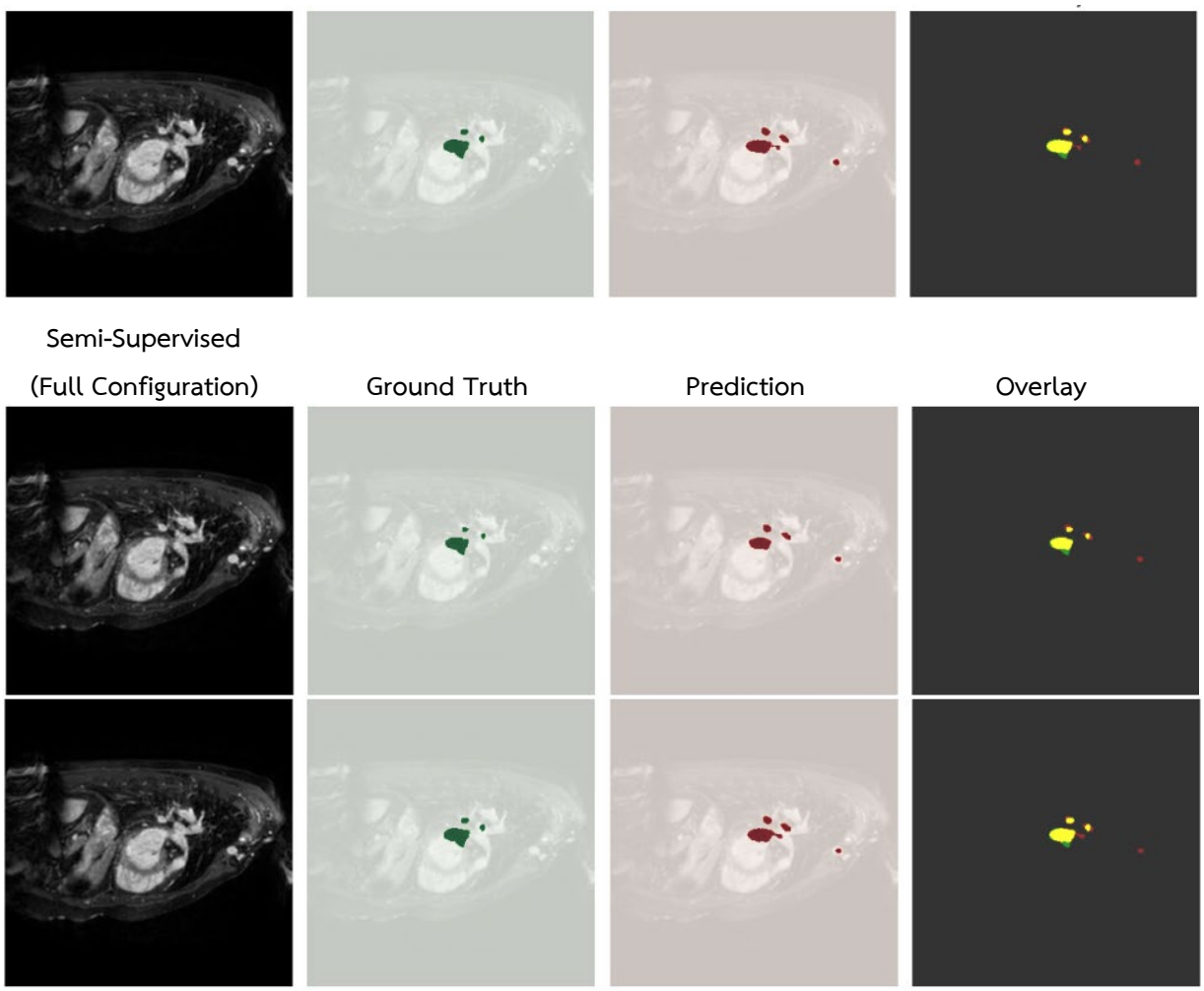


Figure 3 Qualitative comparison of segmentation results

Discussion

The experimental results clearly demonstrate the effectiveness of the proposed semi-supervised framework, particularly in data-scarce environments. The substantial improvement in mean DSC, combined with reduced variance, suggests that the model benefits significantly from leveraging unlabeled data through pseudo-labeling. This finding aligns with existing literature emphasizing the role of unlabeled data in enhancing model generalization.

A key observation is that the semi-supervised model continues to improve even during stages of fluctuating training loss—especially when pseudo-labels are incorporated. This indicates that the model successfully learns meaningful representations from pseudo-labels despite the inherent noise. Qualitative results further confirm more anatomically plausible segmentations, supporting the hypothesis that unlabeled data can guide the model toward better anatomical priors.

To mitigate the impact of noisy pseudo-labels, several strategies were implemented. First, a confidence-aware filtering mechanism was applied to exclude low-probability predictions, with the

confidence threshold dynamically increased from 0.70 to 0.90 as the model improved. Second, the loss weight for pseudo-labeled data was initially set to a very low value (0.01) and gradually increased to 0.4 based on the validation DSC, limiting the early influence of noisy labels. Third, data augmentation was applied to the unlabeled data pool to enhance diversity and reduce overfitting to spurious structures.

Despite these strategies, some noise remains inevitable—primarily due to the limited training data, which constrains the model's ability to generalize during early training. Increasing the size and diversity of the dataset represents a promising direction for further reducing pseudo-label noise. Incorporating additional labeled examples or employing self-supervised pretraining could further enhance the model's robustness to label noise and anatomical variability.

Suggestion

1. Suggestions for Research Utilization

This study presents a semi-supervised segmentation framework based on a 3D U-Net architecture for LA segmentation from cardiac MRI. The framework was intentionally designed and evaluated under extremely limited labeled-data conditions to reflect practical constraints in clinical and research environments. By incorporating pseudo-labeling with dynamic confidence filtering, the method effectively leverages unlabeled data to improve segmentation performance while reducing dependence on manual annotation.

Experimental results demonstrate that the semi-supervised model consistently outperforms its fully supervised counterpart in both accuracy and stability, despite being trained on only 10 labeled volumes. This highlights the framework's capability to operate effectively in data-scarce scenarios, addressing one of the key challenges in medical image analysis.

The iterative refinement of pseudo-labels and the adaptive loss-weighting strategy further enhance model generalization without imposing significant annotation overhead. Nevertheless, the relatively small and homogeneous dataset may limit generalizability to more diverse clinical settings, and the training process remains computationally demanding.

2. Suggestions for Further Research

Future research will explore the scalability and adaptability of the proposed framework through the following directions: 1) Validation on large-scale datasets: Evaluating generalization across multi-center or population-level cardiac MRI datasets, 2) Transfer learning from big data: Pretraining on large datasets and fine-tuning on small labeled subsets for data-scarce domains, 3) Multi-scale pseudo-labeling: Incorporating anatomical context at multiple resolutions to improve pseudo-label robustness, 4) Adaptive uncertainty-aware weighting: Integrating uncertainty estimation

into the loss function to better handle noisy pseudo-labels, and 5) Efficiency-oriented implementation: Developing lightweight models suitable for clinical environments with limited computational resources. These directions aim to enhance scalability, robustness, and clinical applicability in future cardiac image-analysis systems.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the author(s) utilized ChatGPT as a language support tool to check grammar and explore alternative phrasings that could enhance clarity. Following the use of this tool, the author(s) reviewed and revised the content as necessary and take(s) full responsibility for the final content of the publication.

References

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B. A., et al. (2022). The medical segmentation decathlon. *Nature communications*, *13*(1), 4128.

Aryan, R., Kejriwal, V., Patel, V., Aggarwal, A., Khanna, V., Thomas, S. B., et al. (2022, November). Left Atrium Segmentation Using Deep Learning Model. *The Proceedings of the 19th 2022 IEEE India Council International Conference (INDICON)*, 1-5. IEEE.

Bardis, M., Houshyar, R., Chantaduly, C., Ushinsky, A., Glavis-Bloom, J., Shaver, M., et al. (2020). Deep learning with limited data: organ segmentation performance by U-Net. *Electronics*, *9*(8), 1199.

Kiryati, N., and Landau, Y. (2021). Dataset growth in medical image analysis research. *Journal of imaging*, *7*(8), 155.

Krittayaphong, R., Rangsin, R., Thinkhamrop, B., Hurst, C., Rattanamongkolgul, S., Sripaiboonkij, N., et al. (2016). Prevalence and associating factors of atrial fibrillation in patients with hypertension: a nation-wide study. *BMC Cardiovascular Disorders*, *16*(1), 57.

Liu, Y., Wang, W., Luo, G., Wang, K., and Li, S. (2022). A contrastive consistency semi-supervised left atrium segmentation model. *Computerized Medical Imaging and Graphics*, *99*, 102092.

Müller, D., Soto-Rey, I., and Kramer, F. (2022). Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, *15*(1), 210.

Oltman, C. G., Kim, T. P., Lee, J. W., Lupu, J. D., Zhu, R., and Moussa, I. D. (2024). Prevalence, Management, and Comorbidities of Adults With Atrial Fibrillation in the United States, 2019 to 2023. *JACC: Advances*, *3*(11), 101330.

Ottesen, J. A., Tong, E., Emblem, K. E., Latysheva, A., Zaharchuk, G., Bjørnerud, A., et al. (2025). Semi-Supervised Learning Allows for Improved Segmentation with Reduced Annotations of Brain Metastases Using Multicenter MRI Data. *Journal of Magnetic Resonance Imaging*, *61*(6), 2469-2479.

Qu, C., Zhang, T., Qiao, H., Tang, Y., Yuille, A. L., and Zhou, Z. (2023). Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. *Advances in Neural Information Processing Systems*, *36*, 36620-36636.

Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., et al. (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. *Journal of the American college of cardiology*, *76*(25), 2982-3021.

Shi, Z., Jiang, M., Li, Y., Wei, B., Wang, Z., Wu, Y., et al. (2024). MLC: Multi-level consistency learning for semi-supervised left atrium segmentation. *Expert Systems with Applications*, *244*, 122903.

Swetha, S., Rafee, A., Manjula, S. H., and Venugopal, K. R. (2023, December). Optimizing Left Atrium Segmentation: A Modified U-NET Architecture with MRI Image Slicing. In *2023 IEEE 2nd International Conference on Data, Decision and Systems (ICDDS)* (pp. 1-6). IEEE.

Suwanwela, N. C., Chutinet, A., Autjimanon, H., Ounahachok, T., Decha-Umphai, C., Chockchai, S., et al. (2021). Atrial fibrillation prevalence and risk profile from novel community-based screening in Thailand: A prospective multi-centre study. *IJC Heart and Vasculature*, *32*, 100709.

Taha, A. A., and Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, *15*(1), 29.

Uslu, F., Varela, M., Boniface, G., Mahenthran, T., Chubb, H., and Bharath, A. A. (2021). LA-Net: A multi-task deep network for the segmentation of the left atrium. *IEEE transactions on medical imaging*, *41*(2), 456-464.

Wang, J., Liu, X., Yin, J., and Ding, P. (2022). DC-net: Dual-Consistency semi-supervised learning for 3D left atrium segmentation from MRI. *Biomedical Signal Processing and Control*, *78*, 103870.

Xu, G., Qian, X., Shao, H. C., Luo, J., Lu, W., and Zhang, Y. (2024). A SAM-guided and Match-based Semi-Supervised Segmentation Framework for Medical Imaging. *arXiv preprint arXiv:2411.16949*.

Yang, Q., Wang, C., Pan, K., Xia, B., Xie, R., and Shi, J. (2024). An improved 3D-UNet-based brain hippocampus segmentation model based on MR images. *BMC Medical Imaging, 24*(1), 166.

Zhang, X., Noga, M., Martin, D. G., and Punithakumar, K. (2021). Fully automated left atrium segmentation from anatomical cine long-axis MRI sequences using deep convolutional neural network with unscented Kalman filter. *Medical image analysis*, *68*, 101916.