Ensemble of Deep Convolutional Neural Networks with Monte Carlo Dropout Sampling for Automated Image Segmentation Quality Control and Robust Deep Learning Using Small Datasets

Evan Hann[1], Ricardo A. Gonzales[1], Iulia A. Popescu[1], Qiang Zhang[1], Vanessa M. Ferreira[1], and Stefan K. Piechnik[1]

Oxford Centre for Clinical Magnetic Resonance Research (OCMR), Division of Cardiovascular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

Abstract. Recent progress on deep learning (DL)-based medical image segmentation can enable fast extraction of clinical parameters for efficient clinical workflows. However, current DL methods can still fail and require manual visual inspection of outputs, which is time-consuming and diminishes the advantages of automation. For clinical applications, it is essential to develop DL approaches that can not only perform accurate segmentation, but also predict the segmentation quality and flag poor-quality results to avoid errors in diagnosis. To achieve robust performance, DL-based methods often require large datasets, which are not always readily available. It would be highly desirable to be able to train DL models using only small datasets, but this requires a quality prediction method to ensure reliability. We present a novel segmentation framework utilizing an ensemble of deep convolutional neural networks with Monte Carlo sampling. The proposed framework merges the advantages of both state-of-the-art deep ensembles and Bayesian approaches, to provide robust segmentation with inherent quality control. We successfully developed and tested this framework using just a small MRI dataset of 45 subjects. The framework obtained high mean Dice similarity coefficients (DSC) for segmentation of the endocardium (0.922) and the epicardium (0.942); importantly, segmentation DSC can be accurately predicted with low mean absolute errors ($\leq$0.035), in the absence of the manual ground truth. Furthermore, binary classification of segmentation quality achieved a near-perfect accuracy of 99%. The proposed framework can enable fast and reliable medical image analysis with accurate quality control, and training of DL-based methods using even small datasets.

Keywords: Automated quality assessment · Segmentation · Ensemble learning · Monte Carlo sampling.

1 Introduction

Cardiovascular diseases (CVD) are a leading cause of mortality worldwide [1]. Cardiac magnetic resonance (CMR) imaging is a powerful tool in the diagnosis and treatment of CVD, providing comprehensive analysis of cardiac structure and function, especially the left ventricle (LV). Accurate segmentation of the LV is an essential step for the quantification of clinically important parameters, such as volumes, ejection fraction and mass. Despite advances of automated segmentation methods, manual delineations and quality assurance are still the current clinical standard for performing and validating automated segmentation.

Automated LV segmentation has been extensively studied over the past decade, with progress ranging from classical machine learning to advanced deep learning (DL) approaches. The latter was recently enabled by data availability and hardware development. There have been a number of international challenges and collective efforts to benchmark state-of-the-art segmentation accuracy, providing valuable CMR cine SSFP images of the LV, such as the Sunnybrook Cardiac Dataset [2], the Automatic Cardiac Diagnosis Challenge [3], and the UK Biobank [4].

Given the time-consuming task of manual annotation of CMR images in typical clinical workflow, there is significant interest in fully automatic segmentation. Initial efforts required manual extraction of relevant image features with prior knowledge to achieve satisfactory accuracy. A series of LV segmentation methods have been proposed using the publicly available Sunnybrook Cardiac Dataset of 45 subjects [2]. Among others, the proposed approaches use deformable models [5, 6], image-based [7–9] and model-based [10, 11] methods. However, the hand-crafted approaches can fall short in generalizability when dealing with unfamiliar new data. Furthermore, they often require manual adjustments, which limit implementation of fully-automatic tools in modern clinical practice.

With recent advancements of DL, data-driven neural networks can learn end-to-end for image segmentation, reducing the need for hand-crafted approaches. Nevertheless, even state-of-the-art DL methods can still fail on unfamiliar testing data [3]. Case-by-case visual inspection of segmentation quality is still necessary, which is laborious, time-consuming, and defies the benefits of fully-automated methods. Moreover, to achieve robust performance, end-to-end deep learning-based methods require larger and more representative datasets [3,12], which can be time-consuming to curate and not always readily available. Training of DL models requiring only small datasets would be desirable, but demands a quality prediction method in real-world applications, to flag poor-quality results. We therefore present a DL approach, with automated quality prediction, which holds the DL models accountable, even when trained on small datasets. We validated this novel framework on the Sunnybrook Cardiac Dataset for LV segmentation.

1.1 Related Work

There is increasing interest in developing accountable DL-based segmentation methods with inherent quality control. Bayesian approaches have been proposed

to provide means of uncertainty estimation for prediction. In particular, Monte Carlo sampling-based neural networks have been used to perform medical image segmentation, as well as quality control [13, 14]. To implement the Monte Carlo sampling approach, a deep convolutional neural network can be modified by adding dropout units, which randomly "turn off" some internal connections within the neural network [13, 14]. While dropout units are activated only for training in standard DL, they can be activated for testing or deployment to generate many different segmentation samples. The agreement among the samples can be exploited to predict segmentation evaluation metrics, such as Dice similarity coefficient (DSC), without the need of a reference manual segmentation. [13] has successfully demonstrated the capability of the Monte Carlo dropout (MCD) approach for whole brain segmentation.

Alternatively, deep ensembles have also been used to estimate uncertainty and predict segmentation quality [15–17]. Successful applications include segmentation of the brain, prostate, and cardiovascular structures [15, 17, 18]. Similar to Monte Carlo sampling, deep ensembles also generate multiple candidates, then exploit the agreement among candidates to predict output quality or uncertainty. The difference is that a single trained neural network with Monte Carlo dropout can theoretically generate unlimited number of segmentation candidates, whereas the number of candidates generated by deep ensembles is limited by the number of independent neural networks trained. For example, an ensemble of 50 independently-trained neural networks can generate up to 50 different segmentation candidates [18]. This makes deep ensembles more computationally expensive to train and deploy than the Monte Carlo dropout approach. Despite this disadvantage, deep ensembles tend to generate more diverse prediction samples, offering higher accuracy and robustness in uncertainty estimation compared to Bayesian approaches [18,19]. In addition, selecting the segmentation candidate with the best predicted quality as the final output for deep ensembles can improve the overall accuracy and robustness [15, 17]. The same mechanism has not been applied to Bayesian approaches. Therefore, deep ensembles and Bayesian approaches have their own merits and pitfalls.

It has been shown that using an ensemble of multiple MC-dropout models can improve classification accuracy for handwritten digit and character recognition tasks [20]. In this work, we further explore the idea of combining novel deep ensemble frameworks such as [15, 17] and Bayesian approaches for reliable medical image segmentation and quality control.

1.2 Contributions

The contributions of this work are as follows: (1) we propose a novel ensemble of deep convolutional neural networks with Monte-Carlo dropout to merge the advantages of both deep ensembles and Bayesian approaches for reliable medical image segmentation and quality control; (2) we show that deep ensembles can generate diverse segmentation candidates for reliable quality prediction; (3) we add Monte Carlo dropout in the individual neural networks to efficiently generate a large number of segmentation samples; (4) the proposed framework adopts

a novel automatic selection of the final optimal segmentation from multiple candidates [15, 17], and we demonstrate that the proposed framework can produce more accurate segmentation; (5) the proposed approach predicts the quality of segmentation accurately even when trained with a highly-limited (small) dataset.

2 Methods

2.1 Data

The Sunnybrook Cardiac Dataset [2] comprises 45 subjects divided into normal controls and 3 different pathological groups: heart failure with ischemia, heart failure without ischemia, and hypertrophic cardiomyopathy. The dataset was randomly split into 38 training subjects (355 images) and 7 testing subjects (65 images). The testing data comprised of two subjects from each of the 3 pathological groups, and one subject from the normal control. For each subject, the short-axis cine SSFP CMR images were provided with manually drawn contours on both endocardial and epicardial borders at end-diastole, which were

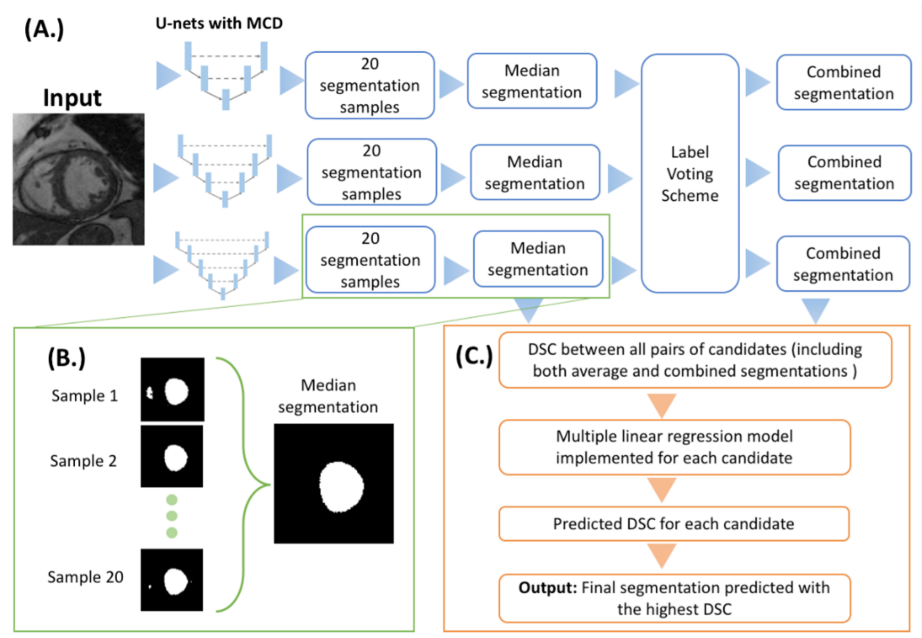


Fig. 1. (A) Overview of the ensemble framework of multiple independently-trained U-nets and combined segmentation models; (B) illustration of generating the median segmentation from 20 samples by each U-net, and (C) the segmentation quality control pipeline.

considered the ground truth for training and testing in this work. Images at end-systole were not used for the development of this work, as only the endocardial contours were provided, without the epicardial contours. The training data were augmented by randomly rotating within $\pm 10^\circ$ to prevent overfitting. In total, 85200 augmented training images were generated.

2.2 Overview of the Ensemble Framework

The proposed ensemble framework (Fig. 1A) involved multiple independently-trained U-nets [21] implemented with MCD and their combined segmentation models generated via a label voting scheme [22], with a quality control pipeline to predict the segmentation accuracy and to select the optimal result [15, 17]. The MCD approach (Fig. 1B) used the median of 20 generated segmentation samples for each MCD U-net. The quality control pipeline (Fig. 1C) calculated inter-candidate DSC for quality prediction via multiple linear regression, and selected the final optimal segmentation.

2.3 U-Nets with Monte Carlo Dropout

In the proposed ensemble framework, 6 U-nets [21] with different numbers of convolutional layers (7, 11, 15, 19, 23, 27) were implemented based on [15, 17] to perform segmentation of the LV endocardium and epicardium. By varying the number of convolutional layers across individual U-nets, it was expected to increase prediction diversity of the ensemble for robust quality control. The U-nets were modified by adding MCD units similar to [14]. The dropout units were activated during both training and testing with a dropout rate of 0.5. In this work, each U-net was set to generate 20 different segmentation samples for each anatomical structure (the endocardium or the epicardium) in a given input, as shown in Fig. 1B. The median segmentation candidate was calculated as the mean of the 20 Monte Carlo samples, with thresholding at 0.5, to obtain a binary mask. In other words, 6 median segmentations were produced from a total of 120 samples by the 6 U-nets.

2.4 Combined Segmentation Models

In addition to the 6 U-nets, 6 combined segmentation models (Fig. 2) were also implemented via a pixelwise label voting scheme [15,17,22] to provide additional segmentation candidates for the ensemble. Figure 2 exemplifies the process of generating combined segmentations from 4 models. The input (Fig. 2A) is the median segmentations independently generated by the multiple MCD U-nets. The input segmentations are added up pixel-by-pixel (Fig. 2B) to produce multiple combined segmentations with different thresholds (Fig. 2C). In this work, 6 combined segmentations were generated for each input medical image.

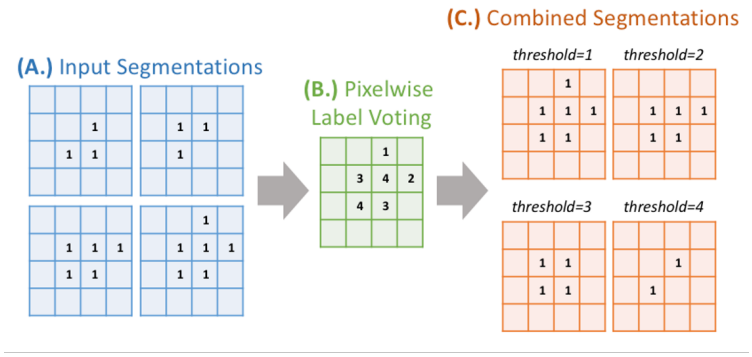


Fig. 2. Illustration of a combined segmentation of 4 models using a label voting scheme: (A) input segmentations are added up to generate (B) a pixelwise vote map, used to calculate (C) combined segmentations with different thresholds. In this work, 6 median segmentations generated by the 6 U-net models with Monte Carlo dropout approach were considered as the input segmentations.

2.5 Prediction of Segmentation Quality

For the quality control component (Fig. 1C), a multiple linear regression model was implemented for each of the 12 candidate segmentation models (including both U-nets with MCD and combined models) based on [15, 17] to predict the ground truth DSC, calculated between the candidate segmentation and the manual ground truth segmentation. The independent variables of the regression model were inter-candidate DSCs calculated between all possible pairs of the 12 candidate segmentations. Via the regression model, the inter-candidate DSCs can associate to the ground truth DSC. The regression parameters have been established using the same ground truth data used for training each individual neural network. Once trained, the regression models can predict DSC of the test segmentation on a per-case basis and in the absence of a manual ground truth segmentation. In this work, the proposed framework adopted a novel mechanism to choose the best final output, with the highest predicted DSC, from multiple candidate segmentations [15, 17].

We also implemented another segmentation quality prediction method based on [13] for comparison. This DSC prediction was calculated by averaging over the DSCs of all possible pairs of Monte Carlo segmentation samples, available only in the MCD models, excluding the combined models in the evaluation.

2.6 Evaluation

Each of the 12 candidate models implemented in the ensemble framework was evaluated for its segmentation performance, measured in terms of mean DSC (and standard deviation), independently for the endocardium and the epicardium. For the U-nets with MCD, only the median segmentations, not the Monte Carlo segmentation samples, were evaluated.

For the quality control component, the regression-based DSC prediction was evaluated independently for each candidate model for both the endocardium and the epicardium. The mean absolute error (MAE) and the Pearson correlation coefficient (r) were calculated between the predicted DSC and the observed ground-truth DSC derived from the manual segmentation. In addition, evaluation of the Monte Carlo-based DSC prediction was also reported for comparison.

3 Experiments Results

The methods were implemented in Python using TensorFlow, Keras and Scipy modules. The neural networks were trained for 240 epochs each, taking 6 h and 48 min in total, with an additional 6 min for the DSC regression models, on a desktop computer equipped with a NVIDIA Titan X GPU. The testing on 7 subjects (65 images) took 12 min and 7 s (i.e. 11 s per image).

3.1 Segmentation Performance

The mean DSC results for all the candidate segmentation models and the proposed ensemble framework are shown for both the endocardium and the epicardium (Table 1). The best mean DSC obtained by a single U-net model was 0.916 (U-net 15) in segmenting the endocardium, and 0.939 (U-net 23) in segmenting the epicardium. The best combined model (Combined Model 3) achieved a mean DSC of 0.920 and 0.941 for the endocardium and the epicardium, respectively. In comparison, the proposed framework outperformed all single and combined models, with a mean DSC of 0.922 and 0.942 for the endocardium and the epicardium, respectively. Furthermore, the framework, comprising of U-nets with Monte Carlo dropout (MCD) in this work, also achieved better performance than the reported results in [17], which implemented U-nets without MCD for the ensemble using the same training and testing datasets. This demonstrates the potential improvement on robustness and accuracy brought forth by integrating the deep ensemble framework with the Bayesian approach, subject to further cross-validation to mitigate the limitation of having a small testing dataset.

Figure 3 shows an example of an apical slice image in the testing dataset (Fig. 3A), with the corresponding manual segmentation of the epicardium (Fig. 3B), and the segmentations by U-net 23 (Fig. 3C), Combined Model 3 (Fig. 3D), and Combined Model 2 chosen by the ensemble framework (Fig. 3E). Despite U-net 23 and Combined Model 3 respectively being the best among the U-nets and the Combined Models, they were outperformed by the proposed ensemble framework when compared to the ground truth manual segmentation. The framework chose the segmentation generated by Combined Model 2, as its predicted DSC (0.82) was higher than the predicted DSCs for U-net 23 (0.77) and Combined Model 3 (0.81). This demonstrates that the on-the-fly selection of segmentation can improve overall segmentation quality by choosing the most-optimal candidate.

Table 1. Mean Dice similarity coefficients (DSCs) for U-nets with Monte-Carlo Dropout (MCD), Combined Models, and the proposed ensemble framework. Standard deviations shown in brackets.

Model	Endocardium DSC	Epicardium DSC
U-net 7 with MCD	0.486 (0.270)	0.569 (0.266)
U-net 11 with MCD	0.878 (0.172)	0.895 (0.166)
U-net 15 with MCD	0.916 (0.127)	0.938 (0.107)
U-net 19 with MCD	0.913 (0.128)	0.936 (0.124)
U-net 23 with MCD	0.915 (0.130)	0.939 (0.124)
U-net 27 with MCD	0.913 (0.128)	0.934 (0.127)
Combined Model 1	0.810 (0.161)	0.856 (0.131)
Combined Model 2	0.913 (0.127)	0.935 (0.123)
Combined Model 3	0.920 (0.126)	0.941 (0.122)
Combined Model 4	0.916 (0.127)	0.936 (0.126)
Combined Model 5	0.887 (0.177)	0.904 (0.175)
Combined Model 6	0.550 (0.297)	0.617 (0.286)
Proposed Framework	**0.922 (0.125)**	**0.942 (0.122)**

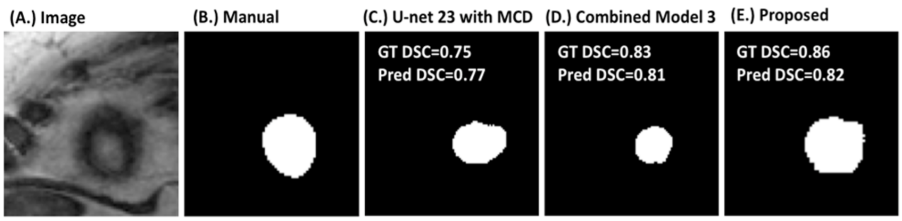


Fig. 3. Example of (A) an input image with (B) its corresponding manual segmentation of the epicardium, (C) segmentation generated by the best single neural network – U-net 23 with Monte Carlos dropout (MCD), (D) segmentation generated by the best Combined Model – Combined Model 3, and (E) final optimal segmentation chosen by the proposed ensemble framework – Combined Model 2, for the epicardium. The corresponding ground truth (GT) Dice similarity coefficients (DSCs) and the predicted (Pred) DSCs are shown.

3.2 Regression-Based DSC Prediction Accuracy

For the evaluation of the DSC prediction via multiple linear regression, the mean absolute errors (MAE) and Pearson correlation coefficients (r) are reported in Table 2 for both the endocardium and the epicardium. All the regression models achieved excellent performance in predicting the ground truth DSC, with very low MAE (from 0.011 to 0.035) and very high Pearson r (0.90 to 1.00).

Table 2. Mean absolute error (MAE) and Pearson coefficient (r) for DSC prediction using regression described in [15, 17]. All r had p<0.0005.

Model	Endocardium		Epicardium	
	MAE	**r**	**MAE**	**r**
U-net 7 with MCD	0.016	1.00	0.011	1.00
U-net 11 with MCD	0.026	0.97	0.018	0.97
U-net 15 with MCD	0.030	0.92	0.021	0.97
U-net 19 with MCD	0.030	0.93	0.023	0.97
U-net 23 with MCD	0.032	0.92	0.020	0.97
U-net 27 with MCD	0.028	0.93	0.024	0.96
Combined Model 1	0.035	0.94	0.021	0.97
Combined Model 2	0.032	0.90	0.023	0.97
Combined Model 3	0.032	0.91	0.022	0.97
Combined Model 4	0.030	0.94	0.021	0.96
Combined Model 5	0.027	0.97	0.023	0.97
Combined Model 6	0.019	1.00	0.014	1.00
Proposed Framework	0.034	0.90	0.023	0.97

The scatter plots (Fig. 4) also reflect the high agreement between the DSC prediction (x-axis) and the ground truth (y-axis) for both the endocardium (Fig. 4A) and the epicardium (Fig. 4B). Most cases clustered closely along the identity line, indicating very accurate DSC predictions. Using a binary threshold at 0.7, the segmentations were classified into good ($\geq$0.7) or poor quality ($<$0.7) with an excellent accuracy of 98% and 99% for the endocardium and the epicardium, respectively, consistent with the performance reported in [15]. This demonstrates the accuracy and practicality of the proposed quality predictions to flag potentially problematic segmentations to human attention for clinical applications.

3.3 Comparison with Monte Carlo-Based DSC Prediction

The Monte Carlo (MC)-based DSC prediction [13] was also evaluated for comparison. The MC-based prediction achieved generally good performance (Table 3), but with higher MAE (from 0.52 to 0.177) and lower Pearson r (0.54 to 0.98) when compared to the regression-based prediction (Table 2). Moreover, the scatter plots (Fig. 5) show that the data points deviate farther from the identity line, with a lower classification accuracy (95%), compared to the regression-based prediction (Fig. 4). Thus, regression-based DSC prediction demonstrated the expected advantages over the intrinsic MC-based agreement measures.

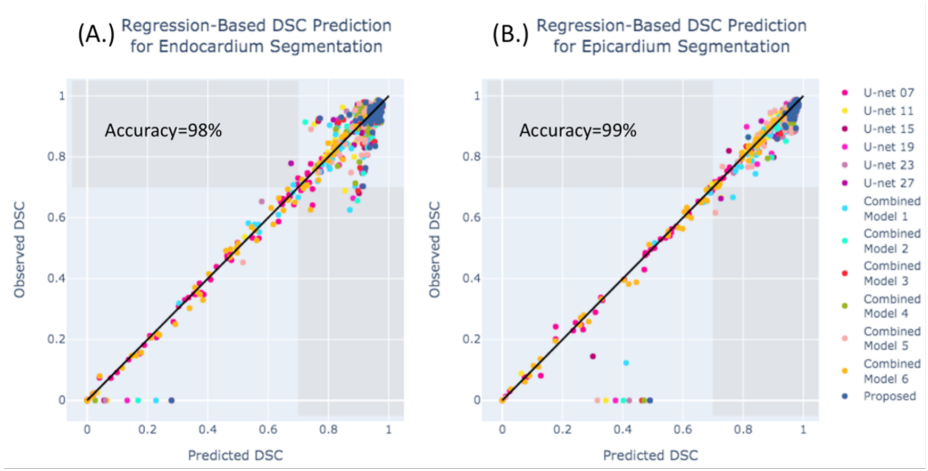


Fig. 4. Scatter plots of the regression-based predicted Dice similarity coefficient (DSC) (x-axis) versus the observed ground-truth DSC (y-axis) for (A) the endocardium and (B) the epicardium. With the quality prediction dichotomized by a binary threshold of 0.7, the DSC prediction achieved a very high classification accuracy of 98% and 99% for the endocardium and the epicardium, respectively. The black diagonal line is the identity line.

Table 3. Mean absolute error (MAE) and Pearson coefficient (r) for DSC prediction using average DSC over all possible pairs of Monte Carlo samples based on [13], available to U-nets with Monte-Carlo dropout (MCD) only. All r had $p<0.0005$.

Model	**Endocardium**		**Epicardium**	
	MAE	**r**	**MAE**	**r**
U-net 7 with MCD	0.177	0.86	0.150	0.92
U-net 11 with MCD	0.052	0.88	0.033	0.94
U-net 15 with MCD	0.062	0.54	0.045	0.74
U-net 19 with MCD	0.068	0.89	0.048	0.96
U-net 23 with MCD	0.064	0.93	0.047	0.98
U-net 27 with MCD	0.066	0.91	0.059	0.95

An example is shown in Fig. 6 showing an input image (Fig. 6A), the corresponding manual segmentation (Fig. 6B), and the automatic epicardium segmentation (Fig. 6C), with a table detailing the DSC prediction results (Fig. 6D). The automatic segmentation was derived from the median of the 20 segmentation samples generated by the U-net 15 with MCD. The MC-based quality control method falsely predicted a high DSC of 0.917 (Fig. 6D top row) with

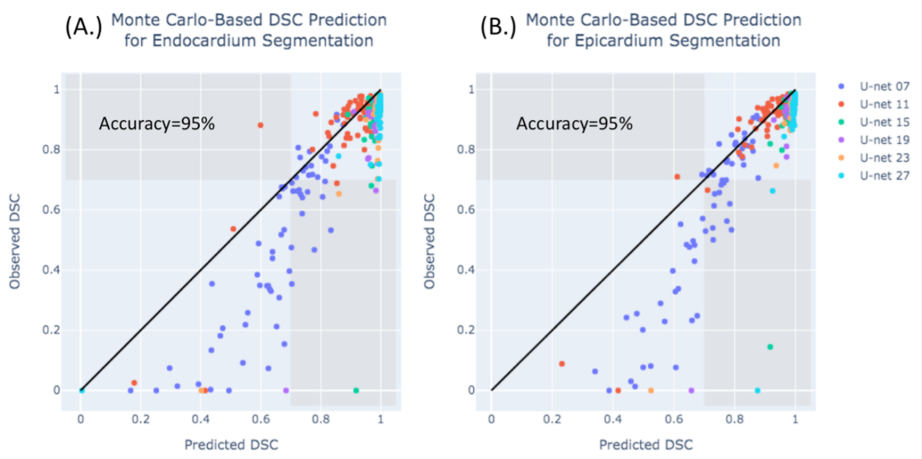
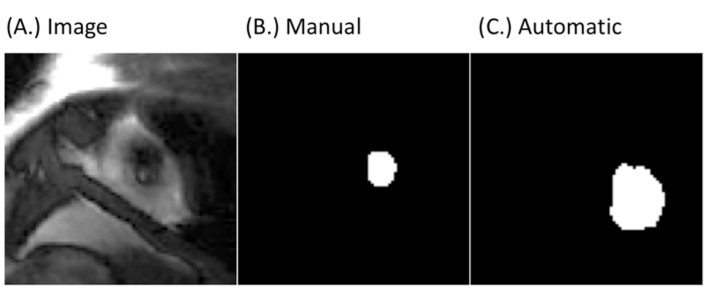


Fig. 5. Scatter plots of the Monte Carlo-based predicted Dice similarity coefficient (DSC) (x-axis) versus the observed ground-truth DSC (y-axis) for (A) the endocardium and (B) the epicardium for U-nets 7 to 27. With a binary threshold of 0.7, the DSC prediction achieved a segmentation quality classification accuracy of 95% for both the endocardium and the epicardium. The black diagonal line is the identity line.



(D.) Quality prediction for the automatic segmentation

Quality Control Method	Predicted DSC	Predicted Classification
Monte Carlo-Based	0.917	Good
Regression-Based	0.301	Poor
Ground Truth	**0.145**	**Poor**

Fig. 6. Example of an (A) input image with (B) its manual segmentation and (C) a poor-quality automatic segmentation, obtained by averaging 20 samples generated by U-net 15 with Monte Carlo dropout (MCD). Table (D) shows quality prediction of the automatic segmentation by the Monte Carlo-based method (top row), the regression-based method (middle row), with the ground truth (bottom row).

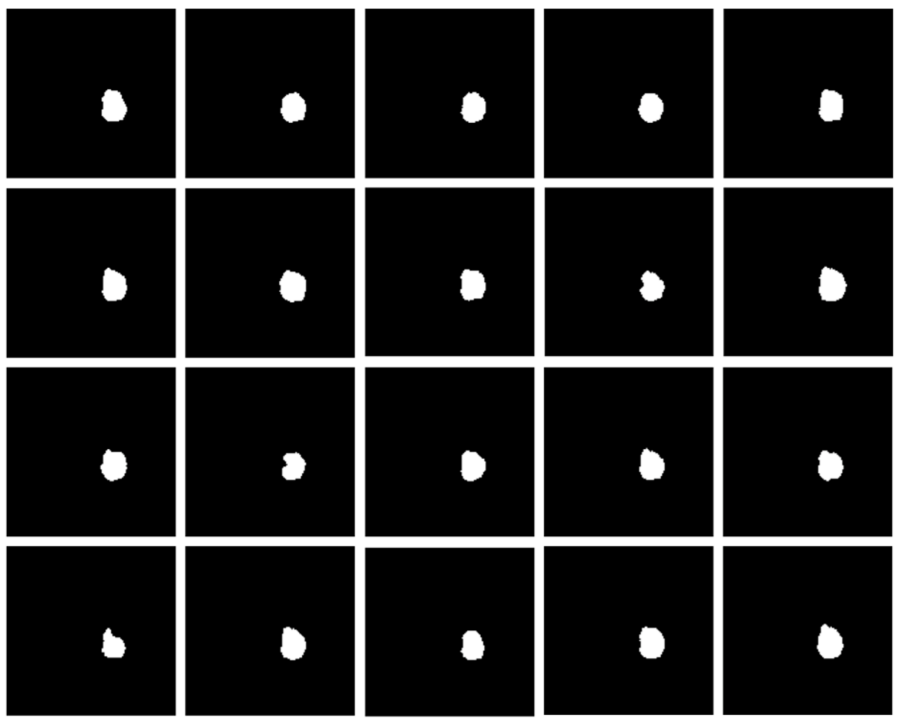


Fig. 7. 20 Monte Carlo segmentation samples generated for the median segmentation in Fig. 6C are shown. The samples lacked diversity in prediction as they highly resemble each other, causing a high Monte Carlo-based predicted Dice similarity coefficient (0.917) despite low agreement with the ground-truth Dice similarity coefficient (0.145).

an incorrectly predicted label of "good quality" for the automatic segmentation, while the regression-based method predicted a low DSC of 0.301 (Fig. 6D middle row). The regression-based method achieved a result closer to the ground truth DSC of 0.145 (Fig. 6D bottom row), and also correctly flagged the poor-quality segmentation.

Figure 7 and 8 are shown for further insights into the differences in prediction performance by the two quality control methods. The MC segmentation samples for the automatic segmentation (Fig. 6C) are shown in Fig. 7. Despite having 20 segmentation samples, the MC samples lacked diversity in prediction and were prone to making the same segmentation mistake – falsely locating the epicardium. This led to an undesirable consequence of predicting a high DSC while the actual ground truth DSC was low. Figure 8 shows the 12 candidate segmentations, which were utilized for the DSC prediction via multiple linear regression in the proposed ensemble framework. Compared with MC samples, the segmentations show more prediction diversity, consistent with the observed advantage of deep ensembles reported in [19].

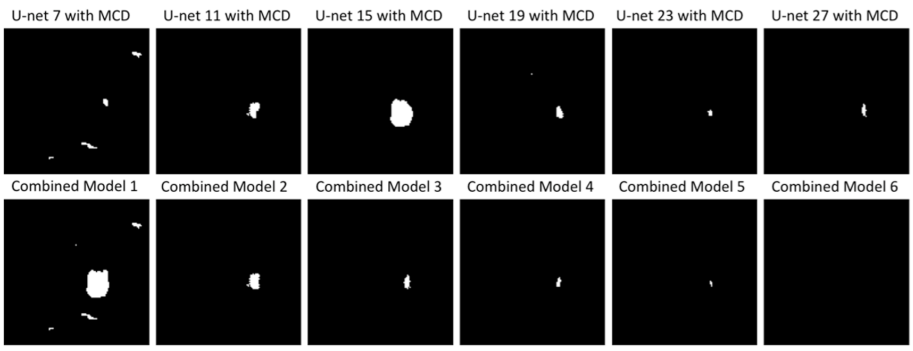


Fig. 8. The proposed framework provided 12 candidate segmentations, with high prediction diversity, are shown for the same input image in Fig. 6A. The segmentation generated by U-net 15 was compared with other candidates to predict a Dice similarity coefficient (0.301), correctly classifying the segmentation as bad quality.

4 Conclusion

In this work, we validated a novel deep ensemble segmentation framework integrated with Bayesian Monte Carlo sampling. The proposed framework can delineate the left ventricular endocardium and epicardium with a high mean DSC of 0.922 and 0.942, respectively. It has inherent quality control, which can predict the segmentation quality in terms of expected DSC with excellent accuracy. We have shown that the regression-based DSC prediction integrated in the framework outperformed the conventional Monte Carlo-based approach, which lacked prediction diversity. This framework successfully merged the advantages of deep neural network ensembles and Bayesian approximation, enabling reliable automatic image segmentation, even for deep learning models trained on small datasets. This can potentially accelerate the advancement of deep learning approaches for diagnostic imaging by reducing requirements of large training datasets.

References

1. Roth, G.A., et al.: Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet 392(10159), 1736–1788 (2018)
2. Radau, P., et al.: Evaluation framework for algorithms segmenting short axis cardiac MRI. The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge (2009)
3. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Transactions on Medical Imaging 37(11), 2514–2525 (2018)

4. Petersen, S.E., et al.: Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank - rationale, challenges and approaches. Journal of Cardiovascular Magnetic Resonance 15(1), 46 (2013)
5. Constantinides, C., et al.: Semi-automated cardiac segmentation on cine magnetic resonance images using GVF-Snake deformable models. The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge (2009)
6. Casta, C., et al.: Evaluation of the dynamic deformable elastic template model for the segmentation of the heart in MRI sequences. The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge (2009)
7. Huang, S., et al.: Segmentation of the left ventricle from cine MR images using a comprehensive approach. The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge (2009)
8. Lu, Y., et al.: Automatic image-driven segmentation of left ventricle in cardiac cine MRI. The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge (2009)
9. Jolly, M.P.: Fully automatic left ventricle segmentation in cardiac cine MR images using registration and minimum surfaces. The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge (2009)
10. O'Brien, S., Ghita, O., Whelan, P.F.: Segmenting the left ventricle in 3D using a coupled ASM and a learned non-rigid spatial model. The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge (2009)
11. Wijnhout, J., et al.: LV challenge LKEB contribution: Fully automated myocardial contour detection. The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge (2009)
12. Chen, C., et al.: Deep learning for cardiac image segmentation: A review. Frontiers in Cardiovascular Medicine 7, 25 (2020)
13. Roy, A.G., et al.: Inherent brain segmentation quality control from fully ConvNet Monte Carlo sampling. In: Frangi, A.F., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. pp. 664–672. Springer International Publishing, Cham (2018)
14. DeVries, T., Taylor, G.W.: Leveraging uncertainty estimates for predicting segmentation quality. arXiv (2018)
15. Hann, E., et al.: Quality control-driven image segmentation towards reliable automatic image analysis in large-scale cardiovascular magnetic resonance aortic cine imaging. In: Shen, D., et al. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. pp. 750–758. Springer International Publishing, Cham (2019)
16. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. arXiv (2016)
17. Hann, E., Popescu, I.A., Zhang, Q., Gonzales, R.A., Barutçu, A., Neubauer, S., Ferreira, V.M., Piechnik, S.K.: Deep neural network ensemble for on-the-fly quality control-driven segmentation of cardiac MRI T1 mapping. Medical Image Analysis 71, 102029 (2021)
18. Mehrtash, A., et al.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Transactions on Medical Imaging 39(12), 3868–3878 (2020)
19. Fort, S., Hu, H., Lakshminarayanan, B.: Deep ensembles: A loss landscape perspective. arXiv (2020)
20. Pop, R., Fulop, P.: Deep ensemble bayesian active learning : Addressing the mode collapse issue in Monte Carlo dropout via ensembles. arXiv (2018)

21. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., et al. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
22. Li, X., et al.: Estimating the ground truth from multiple individual segmentations incorporating prior pattern analysis with application to skin lesion segmentation. In: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. pp. 1438–1441 (2011)