Automated Semantic Segmentation of Cardiac Magnetic Resonance Images with Deep Learning

Sujata Sinha, Thomas Denney, Yang Zhou , Jingyi Zheng
Auburn University, AL 36849, USA
{sujata.sinha, dennets, yangzhou, jingyi.zheng}@auburn.edu

***Abstract*—Machine learning algorithms, especially deep learning architectures, have demonstrated immense potential for biomedical segmentation, often surpassing expert-level performance. For cardiac magnetic resonance (CMR) imaging, semantic segmentation is critical to deriving clinical measures such as myocardial mass and volume. However, challenges still exist. Manual delineation by domain experts is time-consuming and subject to human errors. With semi-automated segmentation techniques, it is challenging to analyze images of the same subject twice, end-diastole and end-systole of the cardiac cycle. To address these challenges, we propose a deep learning-based end-to-end analytical pipeline for automated segmentation of short-axis CMR imaging. The automated pipeline successfully avoids the problem of human subjectivity and achieves expert-level segmentation accuracy. With a large heterogeneous data inclusive of subjects with varying conditions, our model overcomes the data-homogeneity and achieves 99.9% dice similarity score, which outperforms the current state-of-art work.**

***Index Terms*—cardiac magnetic resonance (CMR) imaging, semantic segmentation, deep learning, U-Net, Fully Convolutional Network**

I. Introduction

The World Health Organization ranks cardiovascular diseases (CVDs) as the number 1 cause of global deaths, and estimates 17.9 million deaths because of CVDs [1] each year. Researchers have been investigating CVDs in order to improve the treatment and reduce mortality for years. To better study CVDs, various medical imaging techniques have been employed such as echocardiography, computed tomography (CT), and cardiovascular magnetic resonance imaging (CMR, also known as cardiac MRI), etc. As a non-invasive investigation technique, CMR can produce detailed images of the cardiovascular system with good quality. Unlike single-photon emission computed tomography (SPECT), CMR does not pose the radiation burden, which makes it a safer choice.

Quantitative evaluation of cardiac anatomical structures and their segmentation by clinicians yield meaningful information such as myocardial mass, left ventricle (LV) volume, right ventricle (RV) volume, ejection fraction (EF), etc. The anatomical structure of our interest in this paper is the LV. Traditionally, CMR is manually segmented by domain experts. Clinicians typically take 20 minutes to analyze the CMRs of one subject at the end-diastole and the end-systole. Manual segmentation is time-consuming and suffers from subjective errors. Therefore, researchers started to leverage statistical models and machine learning models [2], [3] to train semi-automated or automated segmentation tools. These models have demonstrated great potential in achieving expert-level performance, but require feature engineering or domain knowledge to improve their performance.

Advancements in hardware, including faster GPUs and TPUs, and the availability of large datasets, greatly benefit the artificial intelligence research especially deep learning (DL). DL-based algorithms avoid human intervention by automatically learning the features from the data, and often surpass human performance in various tasks such as cancer detection [4] and image recognition [5]. With more cardiac MRI images available, DL-based segmentation models gradually show expert-level performance which outperforms traditional methods and become popular in CMR research [6]–[8]. Although extensive research has been done in the field of automatic semantic segmentation of cardiac MRI, several barriers still exist, such as data homogeneity, heterogeneous brightness and intensity values of the LV cavity due to blood flow, lack of large public data, and training performed on shallow networks. Due to several technical constraints, the automatic semantic segmentation of CMR images remains challenging.

To overcome the challenges, we propose a DL-based segmentation pipeline to automatically draw the endocardial and epicardial borders of the LV for a large and heterogeneous CMR data. The data contains 100,199 short-axis cine CMR images recorded from 1344 subjects including volunteers without CVDs and patients with various CVDs such as mitral regurgitation and myocardial infarction. Our data is more heterogeneous and much larger compared with small datasets collected mostly from healthy subjects in most previous work [6]–[10]. After pre-processing the CMR images, we modify and compare three DL architectures: U-net, residual U-net, and Fully convolutional network (FCN). The testing results show that the modified models can achieve expert-level performance and the best one reaches 0.9990 dice score. The automated segmentation models require no human intervention, which saves lots of time and effort for clinicians. Also, the pre-trained network architectures can facilitate transfer learning and automated segmentation for other homogeneous/non-homogeneous datasets.

The remainder of the paper is organized as follows: we present previous work in Section II, introduce our methodology and present the results in Section III and IV, respectively. We discuss the future work and conclude in Section V.

II. Previous Work

Biomedical images have been studied extensively in the past few decades [11], [12]. For cardiac MRI, lots of studies have been conducted to study semi-automated semantic segmentation [11]. Most of the existing work depend considerably on dynamic programming, active contours, graph cut, or some atlas fitting strategies [2], [3]. While semi-automated segmentation tools relieve the workload of clinicians to a certain extent, they still require human intervention, which introduces inter- and intra-observer variability.

With the resurgence of artificial intelligence and machine learning algorithms, various attempts have been made to automate the task of cardiac MRI semantic segmentation. Tran (2016) [13] was one of the first attempts to use fully convolutional networks (FCNs) to segment LV myocardium and RV on short-axis CMR images, and FCNs outperformed several traditional techniques. In the following years, different variations of FCNs have been proposed [9], [10] and numerous works [14], [15] have been done to improve the segmentation performance of CMR by investigating various loss functions. Emad et.al [8] trained a CNN model on 33 patients and achieved 0.9866 accuracy. Besides FCNs, researchers also exploit the U-Net [16] architecture which works well for pixel-wise prediction and shows promising performance even with small amount of training images. For instance, Abdelmaguid et.al [6] trained a U-Net architecture on 2000 images and gave the dice score in the range of $0.9450 - 0.9650$. However, one major limitation of these works is the unavailability of large data sets while they attempt to overcome this by employing various data augmentation techniques.

To address the shortcomings, Bai et.al [7] trained a FCN inspired architecture on a larger dataset with $93,500$ images of $4,875$ subjects from the UK Biobank, and yielded a remarkable performance with 0.94 dice score. However, the UK Biobank dataset is relatively homogeneous, as pointed out by the authors [7]. With the majority of subjects being healthy, only a small proportion of the data corresponds to self-reported cases of CVDs [17].

Compared to the existing works that suffer from the lack of large datasets, shallow networks for training, and the homogeneous nature of test subjects, the automated segmentation pipeline we propose overcomes these restrictions. Our dataset is more extensive and heterogeneous, including healthy volunteers and patients with hypertension, mild to moderate and severe mitral regurgitation, and myocardial infarction with and without diabetes. These patient groups are representative of the most common situations that change the shape of the heart due to CVDs. Also, four different variations of DL models were trained and evaluated to achieve expert-level segmentation capabilities.

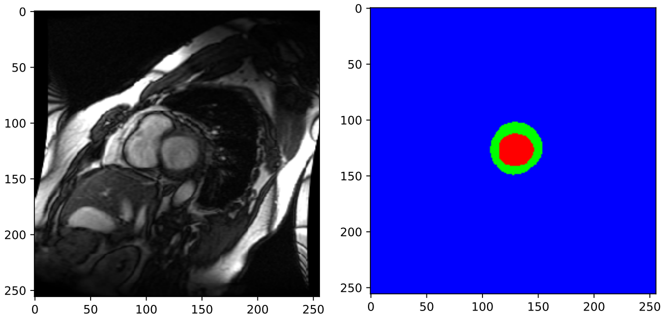
III. Methodology

In this section, we first introduce our data and the techniques used for data augmentation and pre-processing. We then discuss the DL models in detail and further compare their performance.

A. Dataset Description

The CMR images were acquired as part of a separate clinical study at the University of Alabama at Birmingham (UAB). The SCCOR dataset consists of $100,199$ short-axis cine CMR images recorded from $1,344$ subjects. These subjects include normal volunteers and patients with hypertension, mild to moderate and severe mitral regurgitation, and myocardial infarction with and without diabetes. The resolution of images is 256×256, and the mask is $256\times256\times3$, where 3 represents the respective channels.

HELIX dataset containing $8,502$ short-axis cine CMR images collected from 38 patients was also obtained from UAB. A major characteristic of this dataset is that all patients have diastolic dysfunction. HELIX data is in the same format as SCCOR, but there is only one pathology and one visit. HELIX dataset is not involved in the training of deep learning models and are used only during the testing phase.



(a) CMR Image

(b) Ground Truth Mask

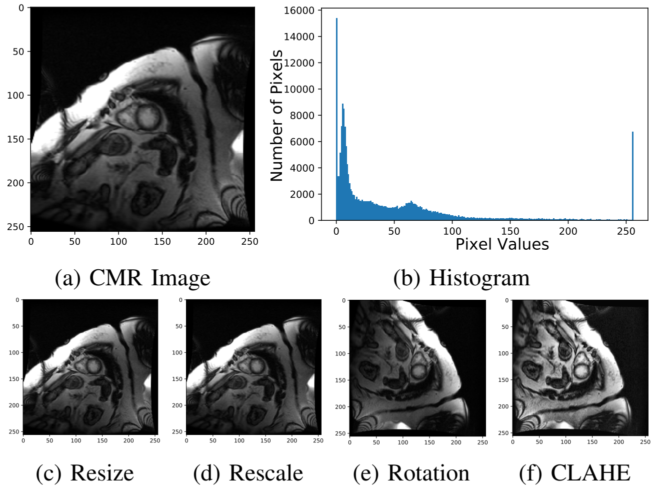
Fig. 1: Illustrative example of the dataset. (a) CMR image of one subject (b) the corresponding ground truth mask obtained using the custom written software [18]. The ground truth masks are further verified by experts to ensure minimal errors.

Manual annotation was performed by a team of experts at Auburn University MRI Research Center using custom-written software [18]. Papillary muscles were considered part of the blood pool.

B. Data Preparation and Augmentation

The custom-written software [18] yields 256×256 dimensional grayscale images and the corresponding expert-annotated RGB ground truth masks of the same dimensions. The CMR images and masks are in Digital Imaging and Communications in Medicine standard (DICOM) and tagged image file format (TIFF) respectively. Since the DL architectures have specific dimensional requirements, we standardize the cardiac images and masks to $256\times256\times1$ and $256\times256\times3$ dimensional numpy arrays respectively. Also, the pixel size ranges from $1.5mm \times 1.5mm$ to $1.8mm \times 1.8mm$. To ensure all images have identical height, width, and channel, we rescale the images and set the in-plan resolution to be $1.5mm \times 1.5mm$. The primary pre-processing operations performed on the training dataset include:

1) Random Rotation: The training images and the corresponding masks are both rotated at a random angle ranging from $-90°$ to $90°$.
2) CLAHE: Contrast Limited AHE (CLAHE) [19] is randomly applied in order to limit contrast amplification, thereby reducing the problem of noise amplification.



(a) CMR Image (b) Histogram

(c) Resize (d) Rescale (e) Rotation (f) CLAHE

Fig. 2: Illustration of Data Pre-processing pipeline. (a) a sample cardiac MR image from the training set, (b) the corresponding histogram, a graphical representation of the pixel values of the image, (c) resized sample image to 256×256 dimensional representation, (d) Rescaling to $1.5mm \times 1.5mm$, (e) random rotation $\in (-90°, 90°)$, (f) CLAHE to improve contrast.

The pre-processing and augmentation pipeline involves image resizing, rescaling, random rotation, and CLAHE, as illustrated in Figure 2. The histogram in panel (b) gives an intuition of the contrast, brightness, and intensity distribution which implies the image in panel (a) is dominated by darker pixel values. After conducting intensive experiments, we decide to have two datasets ready for training, augmented data which goes through four steps (c)-(f) and non-augmented data which only goes through two steps, i.e. resizing (c) and rescaling (d).

C. Deep Learning Architectures

In this section, we describe four deep learning models inspired by the U-Net [16], ResUNet [20], and fully convolutional network (FCN) [21]. They are modified to suit our task and perform well on the evaluation metrics.

1) U-Net with Dropout Layers: The U-Net [16] encoder-decoder network inspires the first deep learning model (Table I). The contracting path or encoder, constituted by 2D convolutional layers, and the expansive path or the decoder, composed of transposed 2D convolutions, form the two major parts of the architecture. The encoder uses repeated 3×3 padded convolutional filters with the rectified linear unit(ReLU) as activation. It is followed by 50% dropout and 2×2 max pooling layer for downsampling. The convolutions have 32, 64, 128, 256, and 512 filters in the contracting path. Each downsampling step in the encoder path doubles the feature map.

TABLE I: U-Net Architecture with Dropout Layers

	Layer Name	No. of Stacks
Input Layer: 256 x 256 x 1 dimensional input		
Encoder	2D Convolutional Layer 50% Dropout 2D Convolutional Layer Max Pooling	x 4
	2D Convolutional Layer 50% Dropout	x 2
Decoder	2D Transposed Convolutional Layer Concatenate from Downsampling Path 2D Convolutional Layer 50% Dropout 2D Convolutional Layer	x 5
Output Layer: 256 x 256 x 3 dimensional Predicted Mask		

Expanding path or decoder upsamples the channel and concatenates feature maps from the encoder path at each stage via skip connections [16]. The expanding path performs localization, segmentation, and increases the output resolution. The model uses truncated Gaussian distribution centered around zero as the kernel initializer. Strides of 2 for max pooling operations is another parameter that is kept constant throughout the experiments (Table V). There are 18 convolution layers, nine on the contracting path, and nine on the expanding path.

2) U-Net without Dropout Layers: The second architecture trained to automate segmentation of CMR images is summarized in Table II. The model is a variation to the first architecture and uses no dropout layers. The hyperparameters used in the experiments are listed in Table V. Our experiment results demonstrate that the U-Net models with and without dropout layers, both achieve expert-level performance, implying that our U-Net models do not suffer from severe overfitting.

TABLE II: U-Net Architecture without Dropout Layers

	Layer Name	No. of Stacks
Input Layer: $256 \times 256 \times 1$ dimensional input		
Encoder	2D Convolutional Layer 2D Convolutional Layer Max Pooling	x 4
	2D Convolutional Layer	x 2
Decoder	2D Transposed Convolutional Layer Concatenate from Downsampling Path 2D Convolutional Layer 2D Convolutional Layer	x 5
Output Layer: 256x256x3 dimensional Predicted Mask		

3) Residual U-Net: Inspired by the ResUNet [20], the third DL model uses a series of residual blocks with skip connections, where each residual unit employs identity mapping [22]. The network architecture is demonstrated in Table III. It follows the U-Net [16] encoder-decoder paradigm as an encoder,

TABLE III: Residual U-Net (ResUNet) Architecture

	Layer Name	No. of Stacks
Input Layer: 256 x 256 x 1 dimensional input		
Encoder	2D Convolutional Layer	x 1
	Batch Normalization Activation Layer 2D Convolutional Layer Addition - Input from previous stack to the next stack	x 5
Bridge	Batch Normalization Activation Layer 2D Convolutional Layer Addition - Input from previous stack to next stack	x 1
Decoder	2D Transposed Convolutional Layer Concatenate from Encoder Path Batch Normalization Activation Layer 2D Convolutional Layer Batch Normalization Activation Layer 2D Convolutional Layer Addition - Input from previous stack to next stack	x 3
2D Convolutional Layer Activation Layer Output Layer: 256 x 256 x 3 dimensional Predicted Mask		

TABLE IV: Fully Convolutional Network (FCN) Architecture

	Layer Name	No. of Stacks
Input Layer: 256 x 256 x 1 dimensional input		
Convolutions and Pooling Layers	2D Convolutional Layer 50%Dropout 2D Convolutional Layer Max Pooling	x 5
	Upsample and Concatenate	
Output Layer: 256 x 256 x 3 dimensional Predicted Mask		

a bridge, and a decoder forming the major components of the network.

The model employs padded 3×3 convolutional filters to increase image depth and batch normalization to smooth the objective function (Equation 1). Since batch normalization may induce a severe gradient explosion during parameter initialization, skip connections in residual networks [23] are employed to alleviate this problem. Instead of using dropout layers to avoid overfitting like the previous models, this model has taken care of the overfitting problem with batch normalization due to its regularization property. The size of the initial filter is 16, and the number of filters are doubled in the encoder at each subsequent convolutional layers. In the decoder network, transposed 2D convolutions perform image upsampling operations and a 1×1 convolution as the output layer generates the desired $256 \times 256 \times 3$ dimensional segmentation masks [22].

4) Fully Convolutional Network: The VGG-16 [24] network forms the backbone architecture of the fourth model (Table IV), with modifications to suit our task of semantic segmentation. Each block in the model uses a 50% dropout layer sandwiched between padded 2D convolutions. With the initial number of filters chosen to be 32, we double the kernels in every subsequent blocks. Each convolutional layer extracts image features and uses strides of 2 instead of 1. Sigmoid as the output activation function is computed instead of softmax probabilities. In our model, the absence of dense layers reduces the number of trainable parameters and the training time. Feature maps learnt at different scales are upsampled to the original resolution using transposed 2D convolutional layer, with strides of 2 and padding “same” (Table V).

TABLE V: Hyperparameters and their values that are kept constant during training of the four deep learning models

Model Hyperparameters	Value
Kernel Size (height and width of 2D Convolutional window)	3 x 3
Strides (Maxpooling)	2 x 2
Padding	”same”
Dropout Level	50%
Kernel Initializer	Truncated Gaussian Distribution
Activation Function	ReLU
Activation Function (Output Layer)	Sigmoid
Learning Rate	0.00001

IV. Results

In this section, we describe the training of deep learning models, and quantitatively assess and compare their performance. All source code for the proposed pipeline and DL models can be obtained from the author’s repository at https://github.com/szs0210/CMR-Segmentation.

A. Training

The model training typically follows augmenting the data, optimizing the objective function, and tuning hyperparameters. To validate and compare four DL models, we randomly split $1,344$ subjects into training, validation, and testing sets with 8:1:1 holdout technique. Quality control measures are taken to ensure the randomization of subjects with various health status (e.g. hypertension, mitral regurgitation, and myocardial infarction) in each of the three sets. Then the augmentation pipeline, as describe in section III-B, is performed on-the-fly before images being fed as inputs into DL models.

Various loss functions have been explored to achieve impressive semantic segmentation performance. In a multi-class context, Novikov et al. (2017) [25] showed that the segmentation performance can be improved with binary cross-entropy (BCE) loss, which can be mathematically formulated as:

$$\text{BCE}\,(p, \hat{p}) = -\left(p \log\left(\hat{p}\right) + \left(1-p\right)\log\left(1-\hat{p}\right)\right) \tag{1}$$

For a CMR image X, let Y and $\hat{Y}$ be the ground truth contours and predicted masks respectively. Then $P(Y=0) = p$ and $P(Y=1) = 1-p$ define the ground truth masks. Pixel values of the prediction $\hat{Y}$ are computed using the sigmoid activation function:

$$\mathbf{P}(\hat{Y}=0) = \frac{1}{1+e^{-x}} = \hat{p} \tag{2}$$

$$\mathbf{P}(\hat{Y}=1) = 1 - \frac{1}{1+e^{-x}} = 1-\hat{p} \tag{3}$$

We optimize the binary cross-entropy loss using Adam optimizer [26] at a momentum of 0.00001. To have a good generalization performance, we use early stopping if the validation loss does not improve for 20 epochs. Table V reports the hyperparameter values used to train, evaluate, and test the network architectures to achieve expert-level segmentation.

TABLE VI: Performance evaluation of the network architectures trained on randomly sampled cardiac MRI images of the SCCOR dataset. A comparative study is conducted among various models on varying size of training sets.

	U-Net with Dropout Layers			**U-Net without Dropout Layers**			**ResUNet**			**FCN**		
Dataset size	**Dice**	**IoU**	**Accuracy**	**Dice**	**IoU**	**Accuracy**	**Dice**	**IoU**	**Accuracy**	**Dice**	**IoU**	**Accuracy**
2000	0.9946	0.9892	0.9964	0.9964	0.9928	0.9972	0.9967	0.9934	0.9973	0.9654	0.9333	0.9820
Augmented 2000	**0.9954**	**0.9908**	**0.9969**	**0.9971**	**0.9942**	**0.9975**	**0.9970**	**0.9940**	**0.9974**	**0.9867**	**0.9738**	**0.9916**
5000	0.9963	0.9926	0.9973	0.9976	0.9952	0.9977	0.9979	0.9958	0.9979	0.9936	0.9874	0.9959
Augmented 5000	**0.9972**	**0.9944**	**0.9978**	**0.9979**	**0.9959**	**0.9979**	0.9979	0.9958	0.9979	**0.9969**	**0.9938**	**0.9977**
10000	0.9963	0.9926	0.9973	0.9981	**0.9963**	0.9963	**0.9986**	**0.9971**	**0.9982**	0.9950	0.9902	0.9967
Augmented 10000	**0.9975**	**0.9950**	**0.9978**	0.9981	0.9962	**0.9980**	0.9982	0.9965	0.9981	**0.9969**	**0.9937**	**0.9977**

TABLE VII: The number of trainable parameters and respective time taken to train the models are reported

Model	**No. of Parameters**	**Training Time**
U-Net with Dropout Layers	8,480,739	1310 minutes ~21.83 hours
U-Net without Dropout Layers	8,480,739	986 minutes ~16.43 hours
ResUNet	4,715,507	670 minutes ~11.17 hours
FCN	7,564,259	843 minutes ~14.05 hours

Since the background of the short-axis image is dark, maximum pixel values are required to be selected. Max pooling operation offers abstract representation and generalization of input, ensuring selection of brighter pixels. The input matrices are padded by zeros which facilitate accurate image analysis. With 50% dropout level, the models can learn more robust features. Considering the instability of gradients and slow learning caused by poor initialization, we choose the truncated Gaussian distribution as the kernel initializer to circumvent this problem.

The U-Net inspired architectures train 8,480,739 trainable filter weights and take the longest training time. The Residual U-Net learns robust features with a smaller network and takes the shortest time to train. Table VII summarizes the number of trainable filter weights and the training times for the respective DL models. For the batch size, extensive experiments have been conducted with batch sizes to be 2, 4, 8, 16, 32, and 64, and it turns out 4 is the optimal choice. Therefore, a batch size of 4 and 200 iterations are kept constant for all the training and experiments conducted in this work. The DL models are trained in Tensorflow 2.1.0 and take approximately 11 to 22 hours on Nvidia GeForce RTX 2080 Ti GPU.

B. Quantitative Assessment

To quantitatively assess the segmentation performance of different DL architectures, we compute and report five commonly used metrics including the Dice Coefficient, Intersection over Union (IoU) or Jaccard Index, Pixel Accuracy, Precision, and Recall. All metrics, as in equations (4)-(8), range from 0 to 1, the higher the better. While pixel accuracy can serve as an evaluation metric, a higher accuracy does not necessarily imply a better segmentation ability of the model. The same for precision and recall. Also, our data suffers from class imbalance. Therefore, we prefer to use the Dice Coefficient and IoU as the evaluation metrics to assess the segmentation performance.

$$\text{Dice Coefficient} = \frac{2TP}{2TP + FP + FN} \tag{4}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{5}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

where TP, TN, FP, and FN stand for the number of true positives, true negatives, false positives, and false negatives, respectively.

We randomly draw CMR image samples of 2000, 5000, and $10,000$ from the SCCOR training set. Each of the four DL models is trained on the non-augmented and augmented samples, respectively. Table VI summarizes the Dice Score, IoU and Pixel Accuracy for each case. In general, data augmentation and a large training dataset improve the performance of deep learning models. When trained on a larger augmented dataset, FCN shows a significant improvement with the IoU

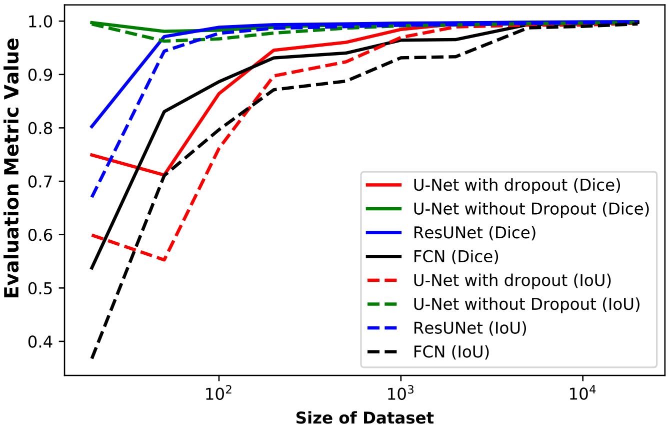
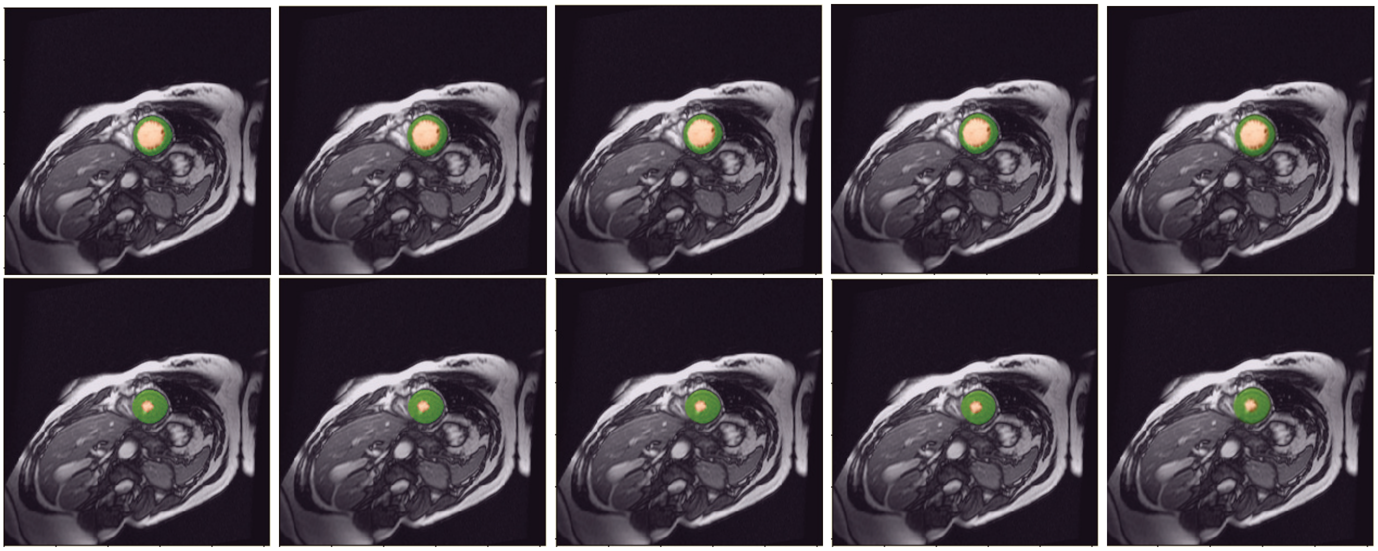


Fig. 3: Performance evaluation of the models : U-Net with 50% dropout, U-Net without dropout layers, Residual U-Net (ResUNet) and Fully Convolutional Network (FCN); over evaluation metrics Dice Similarity Score and Intersection over Union (IoU) for different dataset size



(a) Ground Truth

(b) U-Net with Dropout Layers

(c) U-Net without Dropout Layers

(d) ResUnet

(e) FCN

Fig. 4: Illustrative example of predicted segmentation results for short-axis CMR images against the expert-annotated ground truth image. The first column is the contoured cardiac image in its end-diastole and end-systole volumes respectively. The subsequent columns show the prediction made using the deep learning models: U-Net with 50% dropout, U-Net with no dropout layers, Residual U-Net and Fully Convolutional Network, trained and tested on 1344 subjects.

TABLE VIII: The Dice Similarity Score (Dice), Intersection over Union (IoU), Accuracy, Precision and Recall between automated and manually segmention results of short-axis CMR images. We report the model performance on 10% holdout test set of SCCOR dataset and the HELIX dataset with $8,502$ images

SCCOR: 10,019 test images					
Model	**Dice**	**IoU**	**Accuracy**	**Precision**	**Recall**
U-Net with Dropout Layers	0.9981	0.9962	0.9981	0.9985	0.9985
U-Net without Dropout Layers	0.9986	0.9972	0.9982	0.9988	0.9988
ResUNet	**0.9990**	**0.9976**	**0.9983**	**0.9989**	**0.9989**
FCN	0.9974	0.9947	0.9979	0.9982	0.9981
HELIX: 8,502 test images					
Model	**Dice**	**IoU**	**Accuracy**	**Precision**	**Recall**
U-Net with Dropout Layers	**0.9963**	**0.9926**	**0.9969**	**0.9971**	**0.9970**
U-Net without Dropout Layers	0.9957	0.9914	0.9964	0.9957	0.9957
ResUNet	0.9958	0.9917	0.9964	0.9959	0.9959
FCN	0.9955	0.9910	0.9966	0.9961	0.9961

increasing by 6.48%. However, this observation might not always hold true. For instance, the ResUNet performs better without augmentation when the size of the training set is increased from 2000 to $10,000$.

An additional observation about the behavior of deep learning models, when trained on varying size datasets, can be visualized in Figure 3. It represents the quantitative assessment metric scores - Dice and IoU. We observe that the models perform better with more training data. However, the U-Net architecture demonstrates significantly better performance regardless of the dataset's size, especially for small training datasets containing approximately 20 to 50 CMR images. ResUNet also achieves good performance with a dice score of 0.8 and above, even for small datasets.

The DL models are then trained and tested on $100,199$ cardiac MR images of the heterogeneous SCCOR dataset. As the gold standard for model evaluation, we further test our DL models on HELIX dataset as well as on a randomly sampled test set containing 134 subjects from SCCOR. Table VIII summarizes their segmentation performance. The dice scores are $\in [0.9974, 0.9990]$ for the testing set of SCCOR while they are $\in [0.9955, 0.9963]$ for HELIX. The IoUs are in the range of $0.9947 - 0.9976$ for SCCOR while they are $0.9910 - 0.9926$ for HELIX. Overall, our models can achieve expert-level segmentation performance on both SCCOR and HELIX data with dice scores and IoUs over 99%, significantly higher than the current state of the art performance scores.

Figure 4 shows the automated segmentation results, generated by the four DL architectures, of the same subject at end-diastole (the first row) and end-systole (the second row) of the cardiac cycle. The first column is the expert-annotated CMR images serving as our ground truth. Unlike semi-automated segmentation which finds it challenging to analyze images from the same subjects twice during the cardiac cycle, the DL-based automated segmentation models can easily overcome the problem.

The experiments demonstrate that larger training data can yield an impressive performance for DL models. Backed up by a huge dataset and on-the-fly augmentation, our models are able to achieve expert-level segmentation.

V. Conclusion

In this paper, we propose an end-to-end analytical pipeline with multiple stages for automated short-axis CMR segmentation. Compared with the manual delineation and semi-automated segmentation, our automated segmentation includes zero human intervention, which greatly reliefs the workload of clinicians and avoids the subject error. The automated process is driven by DL models, and we analyze four models: two variations of the U-Net, a ResUNet, and FCN architectures. Our models are then tested on two separate data SCCOR and HELIX and shows expert-level performance with the dice score reaching 0.9990, which is one of the highest scores, as documented in various previous works. We also overcome the limitation of the unavailability of a heterogeneous dataset. Our models are trained on a dataset with subjects falling into categories of normal, hypertension, mild to moderate and severe mitral regurgitation, and myocardial infarction with and without diabetes. We expect our models to serve as good source models and lay a good foundation for problems similar to automated semantic segmentation of CMR images. This would benefit the biomedical research, and significantly facilitate the use of DL techniques in biomedical image analysis.

Acknowledgment

We would like to thank Louis J. Dell'Italia, MD for his support in collecting the cardiac MRI data.

References

[1] "World health organisation. cardiovascular diseases (cvds) fact sheet."

[2] I. B. Ayed, H. mei Chen, K. Punithakumar, I. G. Ross, and S. Li, "Max-flow segmentation of the left ventricle by recovering subject-specific distributions via a bound of the bhattacharyya measure." *Medical Image Analysis*, vol. 16, no. 1, pp. 87–100, 2012.

[3] D. Grosgeorge, C. Petitjean, J. N. Dacher, and S. Ruan, "Graph cut segmentation with a statistical shape model in cardiac mri," *Computer Vision and Image Understanding*, vol. 117, no. 9, pp. 1027–1035, 2013.

[4] B. E. Bejnordi, M. Veta, P. J. van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, O. Geessink, N. Stathonikos, M. C. van Dijk, P. Bult, F. Beca, A. H. Beck, D. Wang, A. Khosla, R. Gargeya, H. Irshad, A. Zhong, Q. Dou, Q. Li, H. Chen, H.-J. Lin, P.-A. Heng, C. Haß, E. Bruni, Q. Wong, U. Halici, M. Ümit Öner, R. Cetin-Atalay, M. Berseth, V. Khvatkov, A. Vylegzhanin, O. Kraus, M. Shaban, N. M. Rajpoot, R. Awan, K. Sirinukunwattana, T. Qaiser, Y.-W. Tsang, D. Tellez, J. Annuscheit, P. Hufnagl, M. Valkonen, K. Kartasalo, L. Latonen, P. Ruusuvuori, K. Liimatainen, S. Albarqouni, B. Mungal, A. George, S. Demirci, N. Navab, S. Watanabe, S. Seno, Y. Takenaka, H. Matsuda, H. A. Phoulady, V. Kovalev, A. Kalinovsky, V. Liauchuk, G. Bueno, M. M. Fernandez-Carrobles, I. Serrano, O. Deniz, D. Racoceanu, and R. Venâncio, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer." *JAMA*, vol. 318, no. 22, pp. 2199–2210, 2017.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[6] E. Abdelmaguid, J. Huang, S. Kenchareddy, D. Singla, L. Wilke, M. H. Nguyen, and I. Altintas, "Left ventricle segmentation and volume estimation on cardiac mri using deep learning." *arXiv preprint arXiv:1809.06247*, 2018.

[7] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, F. Zemrak, K. Fung, J. M. Paiva, V. Carapella, Y. J. Kim, H. Suzuki, B. Kainz, P. M. Matthews, S. E. Petersen, S. K. Piechnik, S. Neubauer, B. Glocker, and D. Rueckert, "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *Journal of Cardiovascular Magnetic Resonance*, vol. 20, no. 1, p. 65, 2018.

[8] O. Emad, I. A. Yassine, and A. S. Fahmy, "Automatic localization of the left ventricle in cardiac mri images using deep learning." in *Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, vol. 2015, 2015, pp. 683–686.

[9] G. Kedenburg, C. A. Cocosco, U. Köthe, W. J. Niessen, E. jan P. A. Vonken, and M. A. Viergever, "Automatic cardiac mri myocardium segmentation using graphcut," in *Progress in biomedical optics and imaging*, vol. 6144, no. 30, 2006.

[10] M. Khened, V. A. Kollerathu, and G. Krishnamurthi, "Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers," *Medical Image Analysis*, vol. 51, pp. 21–45, 2019.

[11] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep learning for cardiac image segmentation: A review." *Frontiers in Cardiovascular Medicine*, vol. 7, p. 25, 2020.

[12] J. Zheng, M. Liang, A. D. Ekstrom, L. Ge, W. Yu, and F. Hsieh, "On association study of scalp eeg data channels under different circumstances," in *International Conference on Wireless Algorithms, Systems, and Applications*, 2018, pp. 683–695.

[13] P. V. Tran, "A fully convolutional neural network for cardiac segmentation in short-axis mri." *arXiv preprint arXiv:1604.00494*, 2016.

[14] M. Chen, L. Fang, and H. Liu, "Fr-net: Focal loss constrained deep residual networks for segmentation of cardiac mri," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, pp. 764–767.

[15] J. Sander, B. D. de Vos, J. M. Wolterink, and I. Išgum, "Towards increased trustworthiness of deep learning segmentation methods on cardiac mri," in *SPIE Medical Imaging 2019: Image Processing*, 2019, p. 1094919.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[17] A. Fry, T. J. Littlejohns, C. Sudlow, N. Doherty, L. Adamska, T. Sprosen, R. Collins, and N. E. Allen, "Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population." *American Journal of Epidemiology*, vol. 186, no. 9, pp. 1026–1034, 2017.

[18] W. Feng, H. Nagaraj, H. Gupta, S. G. Lloyd, I. Aban, G. J. Perry, D. A. Calhoun, L. J. Dell'Italia, and T. S. Denney, "A dual propagation contours technique for semi-automated assessment of systolic and diastolic cardiac function by cmr." *Journal of Cardiovascular Magnetic Resonance*, vol. 11, no. 1, pp. 30–30, 2009.

[19] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. T. H. Romeny, and J. B. Zimmerman, "Adaptive histogram equalization and its variations," *Graphical Models graphical Models and Image Processing computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355–368, 1987.

[20] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[22] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[23] G. Yang, J. Pennington, V. Rao, J. Sohl-Dickstein, and S. S. Schoenholz, "A mean field theory of batch normalization," in *ICLR 2019 : 7th International Conference on Learning Representations*, 2019.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR 2015 : International Conference on Learning Representations 2015*, 2015.

[25] A. A. Novikov, D. Major, D. Lenis, J. Hladuvka, M. Wimmer, and K. Bühler, "Fully convolutional architectures for multi-class segmentation in chest radiographs," *CoRR*, vol. abs/1701.08816, 2017. [Online]. Available: http://arxiv.org/abs/1701.08816

[26] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *ICLR 2015 : International Conference on Learning Representations 2015*, 2015.